NSF Workshop on PIM: Problems and Opportunities

"*Searching, Finding, Filtering and Auto-Classification*" Breakout Group Report

Participants:

Facilitators: Nick Belkin, Susan Dumais (report author), Diane Kelly
Scribe: Luna Dong
Participants:

Nick Belkin (Rutgers University), http://www.scils.rutgers.edu/~belkin/belkin.html
Rick Boardman (Google), http://www.iis.ee.ic.ac.uk/~rick/
Susan Dumais (Microsoft Research), http://research.microsoft.com/~sdumais
Luna Dong (University of Washington), http://www.cs.washington.edu/homes/lunadong/
Jaime Teevan (MIT), http://people.csail.mit.edu/teevan/
Diane Kelly (University of North Carolina), http://www.ils.unc.edu/~dianek/
Brian Ross (University of Illinois), http://www.psych.uiuc.edu/people/showprofile.php?id=7

Overview:

We spent most of our time discussing how to help users re-find information in rich personal stores. Personal information comes from many different sources (email, files, web pages, calendar appointments, instant messages, rss feeds, newspapers, notes, music, images, videos, etc.), in many different formats, and in the context of many different primary activities (writing a paper, creating a presentation, organizing a meeting, reviewing a technical paper, planning a trip, catching up on email, reading the latest news headlines, etc.). People may organize the information into directory or folder structures, they may add annotations, or they may do nothing and rely on full-text search to find it again. The ability to handle the diversity of information types and metadata quality is critical in accessing personal information.

Looking for information in a personal store is different in many ways from a search in an unknown collection like the Web. Perhaps the most important difference is that people are familiar with many different characteristics of information as well as the contexts in which they have previously encountered it. Because people remember different characteristics of the information they are looking for at different times (e.g., who sent an email, when you created a document, the topic of a memo), it is important to support a wide variety of access routes. The idea of fast and flexible access to personal digital memories was popularized by Vannevar Bush in his seminal paper in 1945. Although the technologies are quite different than those envisioned by Bush, the latest operating systems (e.g., Apple's Tiger OSX and Microsoft's Vista OS) and new desktop search tools (e.g., Copernic, Google, HotBot, Lookout, MSN, X1, Yahoo) provide the infrastructure to support some key of his vision. Key challenges

remain in combining automatic and human organization, and in providing interfaces to help users specify their information needs and understand the results returned.

Description:
We first considered characteristics of human memory, including what cues people remember about information and what kinds of initial processing will be most useful at retrieval time. We then turned to the topic of how to design systems that can better support people in organizing and harvesting personal memories.

Cues for search.
What do people remember about their personal information?
- Content. What is it about? This is a primary search cue on the Web and will be important in personal information as well. Because people have previously interacted with (created, read, modified) information in their personal stores, they will also remember a wide range of other cues about these items. Techniques for supporting access along many different dimensions, as well as tools for capturing metadata about objects (either automatically by recording attributes like time, or manually by allowing people to file or annotate items) are important to develop for PIM applications.
- Context. What was I doing when I encountered the item? What happened just before or after? How similar is the retrieval context to the context at time of previous encounters?
- Time. When did I initially encounter it? When did I subsequently use it?
- People. Who was involved?
- Storage location. Where did I file the item?
- Physical characteristics. What does the item 'look' like? This includes physical characteristics of the item (e.g. size, position, type, font) as well as the context in which I encountered it (e.g., other items that were around, ambient music, time of day, location).
- Distinctiveness. How distinct is this item? Slameka and other psychologist have shown that items which are distinct are easier to retrieve. The distinctiveness of duration, location and attendees are important in predicting which electronic calendar appointments people will find memorable (Horvitz, Dumais and Koch).
- Encoding effects. What did I do with the item? How items are processed when they are initially encountered has a large effect on how easily they can later be retrieved. For example, items that are processed more deeply are retrieved more easily (Craik and Lockhart). This might have interesting implications for automatic vs. manual filing of email. Manual filing (and other types of explicit organization) should improve the memorability of the item, but at the cost of additional processing time.
- Recency and frequency. Two important factors in retrieval from human memory are how often an item has previously been encountered (frequency and the spacing of practice) and when it was encountered (recency). Anderson and

Schooler have argued that these characteristics of human memory can be considered as a rational adaptation to statistics of the item encountered in the world. From an information systems perspective, this suggests that temporal and usage factors should be incorporated into access schemes.
- Recall vs. recognition. It is much easier for people to recognize items from among a set of alternatives than to recall or generate the items. Retrieval of personal information is an interesting case, which may lie between these extremes. When people search for information they have seen before they remember some attributes, with varying degrees of accuracy. The variety of attributes that are stored when an item is encountered and how the information is presented will both improve this kind of cued recall.

There was some discussion of the extent to which searching and browsing are qualitatively different activities or are extremes of a continuum, but we did not resolve the issue. There was agreement that people do both and that they need to be able to go back and forth between them easily. We also discussed whether queries could be thought of as data, again an issue which we did not resolve. A common theme of our discussion was the need to support access using a wide range of cues that people might remember about items of interest.

Harvesting personal memories.
What kinds of aids can we provide to people to make it easier for them to access their personal memories? We broke this problem down into two main areas: communicating information needs and understanding the results that are presented.

C*ommunicating information needs.* Today the most common way for users to specify their information need is to type keywords into a small "search box", or browse a hierarchy of folders organized along a single attribute (typically folder name). We discussed a number of alternative techniques. Relevance feedback, in which people mark some items as relevant, is a well-know technique for improving the relevance of items in batch mode evaluations (Salton and Buckley). Interestingly there are few examples of its use in operational systems. Encouraging people to say more about their information needs has also been shown to improve retrieval accuracy in laboratory studies (Kelly). Spelling correction is a simple technique that works well when people misspell what they are looking for, although even here the details of how and when alternatives are presented has a large influence on the success of the approach (Mayer). Tabbed completion is another alternative that has been successful in some settings, although it is difficult for novices to discover. Recommendations are another technique for suggesting related items or query terms that might also be of interest. Capabilities that support richer interaction are also possible, including specification by selecting regions of the current document, or richer representation and use of facets. Implicit queries can also be generated using current document contexts.

Understanding the results presentation. Today most retrieval systems return a long list of results which typically includes a title, url and short contextual description. Several experimental systems have explored alternatives, but few are widely used. Some systems have presented richer summaries such as thumbnails, query-relevant thumbnails, or additional details on demand. Others have grouped the results in some way, e.g., by site or by content using text clustering or text classification techniques. The use of richer faceted metadata has been explored by several groups, and appears to be especially important for personal information retrieval since people remember many attributes of items that they have previously encountered (e.g., Dumais et al.). Some metadata can be automatically captured (e.g., the time an item is receive, author, recipient, subject, interaction history, etc.), but we need to support users in specifying additional metadata as well. Popular folder structures are one type of metadata, but others could include things like a "keeper button" that allows users to mark the current item as important or to save the current context for subsequent presentation. There is a tight coupling between storage and retrieval and we need to consider both in designing systems. There was an interesting discussion of the extent to which the same cues are useful for finding and re-finding.

Key research challenges:
There are tremendous opportunities to go beyond the popular search box and a long list of results to help people to specify their information needs and to understand the relations among results that are returned.

There are large individual differences (both across individuals and within an individual for different task) in the strategies for organizing and retrieving information. Understanding the costs and benefits of investing in saving and organizing information is an important first step – i.e., to what extent are the costs invested in organizing information worth it in terms of retrieval accuracy or speed; can we develop tools to mitigate costs and improve benefits? The evolution of content, strategies and access patterns over time is an important dimension that is just beginning to be explored.

The ability to handle diverse types of information and metadata is critical for accessing personal information. People create and encounter many different kinds of information and it all needs to be accessible. People remember many cues about information they have previously encountered and systems need to provide multiple access routes and allow users to switch easily between them. Iteration and interaction, rather than one-shot querying, need to be supported. Richer visualizations showing the relationships among retrieved items might be appropriate for some information analysis applications.

It is also important to develop new techniques for specifying information needs that go beyond a simple search box. People bring much more to retrieval situations than 2.5 words, including rich search histories. Researchers should

work to develop techniques to elicit and capture this information and incorporate it into retrieval. Understanding what and when to elicit (either explicitly or automatically) are important topics for future research.

Information retrieval is driven by information needs, so a richer understanding of users' tasks and contexts is central to developing systems that support the management and retrieval of personal information.

Final thoughts and a parting image:
The image below is a screen shot of the final poster our group used to summarize our discussion at the workshop. It might not be memorable to readers, but to those of us who were involved in its creation, it will serve as an important memory cue for years to come!