Don’t Drop the Ball: Re-finding Personal Information

Personal information management has been described as a game of catch, where a person tosses their personal information into the future, in hopes of being able to catch the information later when it is needed [2]. This report focuses on the catching aspect of personal information management, discussing current approaches to and problems with how people return to previously encountered information when it has become useful.

As an example, imagine Alex, who organized his company’s football team several years ago, and was recently asked by the current captain where he purchased the team jerseys. Alex must use his memory and the organizational structure he created when managing the football team to re-find the company’s name. He could use the structure he created to help him remember the company name by searching, for example, for email communications with the jersey producing company in an old email folder or for the invoice in a file directory. He could also search the Web, using an old bookmark to return to the company’s Web page or issuing a search to an Internet search engine for something like, “football jerseys” and browsing the result list for a familiar looking Web page.

However, the organizational structure Alex created several years ago is probably difficult for him to operate in effectively now. Further, the bookmarks Alex made are likely to have changed, and he is unlikely to be able to recognize the company’s Web page should it be presented to him, let alone be able to issue an Internet search that finds it. His hunt for the name of the company where he purchased the team jerseys will probably require significant effort, and if he is unable to find the name, the current captain will be required to also expend significant effort repeating research that Alex has already performed.

As can be seen from this example, re-finding personal information is an important problem that is difficult to solve. The amount and types of information that people routinely encounter, create, use and/or save in digital form are expanding dramatically. We can assume that this increase will continue, as computing becomes ever more ubiquitous and part of our daily lives, creating a great need for effective re-finding solutions. Current tools for re-finding even textual personal information are only in their infancy, and are based on rather traditional information retrieval models, without taking into account the particular characteristics of the personal information situation [4].

Below we discuss several important controversial statements on the topic of re-finding, highlighting key arguments for and against each statement. Short term and long term goals that arise from the statements are highlighted, as are any resources needed to pursue resolution of the controversy.

Note that in this report, terms such as *search* or *finding* do not refer exclusively to keyword search (*e.g.*, Alex’s Internet search for “football jerseys”), or even directed search (*e.g.*, Alex’s search for the company name), but can also refer to the entire information seeking process (*e.g.*, the new captain’s effort to learn about a good company from which to purchase new football jerseys) [1].

Finding = Re-finding

2 participants agree, 4 disagree

The first controversial statement is that re-finding is essentially the same behavior as finding. In this section we discuss whether we believe the two behaviors are the same, and, if they differ, what the important aspects of that difference are. A significant feature of re-finding is that people tend to know a lot of meta-information about the item they are seeking. For example, if Brooke wanted to purchase a CD she saw earlier on Amazon.com, she would probably return to Amazon.com, and use information about how she originally encountered the CD to follow a similar path to return to it.

Nonetheless, the strategies that people tend to employ when searching for new information versus returning to previously viewed information appear to be similar. Teevan, et al. [6], found that regardless of whether people were looking for information on the Web (usually a finding behavior) or in their files and email (usually a re-finding behavior), they tended to navigate to their information target via a series of small steps, using the various meta-information they knew about that target to inform the steps. For example, even if Brooke were searching for a new CD to purchase, she might know the basic genre of music she likes, what sort of CD cover art tends to appeal to her, and that Amazon.com is a store where music CDs can be purchased. She could use this information to find the CD by visiting Amazon.com, navigating to her preferred music genre, and then browsing for appealing cover art.

This finding behavior is very similar to the previous example where Brooke was re-finding, except that when finding for the first time she does not have personal experience with the actual target. Instead, the meta-data she uses is based on prior experience with similar items. Those who argue there is a difference between finding and re-finding claim that there is a qualitative difference between the meta-data a user knows about their information target based on experience with the actual item, such as exactly what the item looked like or when it was last seen, and other types of meta-data a user might have about an unknown target.

It has also been argued that when searching, a person experiences considerably more frustration when unable to locate the target if the target has been seen before than if it has not. For example, Brooke is likely to find it more frustrating to not be able to return to a CD she's already seen on Amazon than to not be able to find a new CD she likes. The amount of frustration a user experiences with a search is probably related to the searcher's expectation that the item exists, but whether a person can or cannot have a similar level of expectation for an item that has not been seen before as for an item that has is a matter of debate.

Another difference cited between finding and re-finding is that users can easily recognize their target when it has been seen before, rather than having to think about and determine that a particular item is indeed what is being sought. Those who believe re-finding and finding are the same again believe that sometimes items that have been seen before can take effort to recognize, while new items might be immediately recognizable as relevant.

Those participants that believe finding and re-finding are the same, believe that the same tools should support both behaviors. However, if re-finding is indeed found to be

qualitatively different, it remains an open question as to whether the two behaviors should be supported differently, and, if so, how.

Re-finding in Personal Information = Re-finding on the Web

6 participants agree, 0 disagree

Regardless of whether finding and re-finding are the same behavior, we also discussed whether there was a qualitative difference in re-finding behavior based on the corpus. Although there is consensus among the participants in our breakout group that re-finding in one's personal information space is the same as re-finding on the Web, this statement is not taken to be true among the general PIM community. It is our belief that once information has been seen, it enters a person's personal information space, regardless of whether that information continues to reside on the Web or under direct control of the individual.

This is an open question about PIM in general, and not necessarily unique to the problem of finding and re-finding information. Many believe that there is a fundamental difference between information one believes they have control over and information that others have control over, and while this is likely to be true, the degree to which this makes *re-finding* qualitatively different in the two situations is unclear to us.

People shouldn't have to do any work in advance to make re-finding easier.

5 participants agree, 1 disagrees.

Earlier, personal information management was described as a game of catch. But should it really be necessary to "toss" information into the future in order to be able to catch it when needed? Or should relevant information be provided to a user regardless of their previous interaction with that information?

As an example, Alex, mentioned above, might have created considerable organizational structure when organizing his company's football team. This structure would serve useful to him when later asked to re-find the name of the company he ordered jerseys from. If he does not have a rich organizational structure, he might have a more difficult time re-finding that information. Organizational structure allows the user to use recognition, rather than recall, in their search process.

While organizational structure likely serves an important purpose, we believe that it need not necessarily be the created by the user, but could also be automatically generated by the system. Further, the organizational structure need not be static. Alex could issue a query for "football jersey", be reminded of any similar searches he ran earlier, and then use one of those similar searches to find the company, essentially using the search results like dynamic folders. Similarly, Yee, et al. [7] create dynamic organizational structure by allowing users to browse faceted meta-data.

Short term goal: Make advance work unnecessary for re-finding.

Note that while there is disagreement as to whether advance organization should be required of the user, none of the participants believe effectively being able to find information will make the process of organizing obsolete. Information organization furthers the user's understanding of the information space and helps the user remember the information being organized. In fact, people who file their information rather than pile it are more likely to use keyword search when looking for something [6], perhaps

because of the role organization plays in helping them memorize and understand the information.

People should not have to do any work at all to re-find.

3 participants agree, 3 disagree

Just as catching a ball is only a part of a greater game such as football, so is re-finding, and, indeed, all of information management, just an activity that is part of a greater task. While in our original example, Alex was asked to re-find the name of the company he originally purchased team jerseys from, that name was necessary only because the team needed new football jerseys. While in this report we primarily discuss re-finding in isolation, it is important to consider the activity's greater context.

Ideally, a user would not have to do any work to re-find information at all. The previous section talked about doing away with the "tossing" in personal information management. The participants who believe that users should not have to do any work to re-find believe the "catching" of personal information should also be done away with. Instead, relevant information should just appear when and where the user needs it.

Examples of this exist already in many small and task specific ways. For example, many email clients support filling in the recipient's email address as the user starts to type his name so that the user doesn't need to actually find the address. One could imagine even more clairvoyant systems that know, based on the user's context, the likely email recipient and automatically fill in the recipient field. Other examples of task-embedded re-finding include the Remembrance Agent [3], which re-finds documents relate to a document being currently composed, and Aria [5], which re-finds images related to an email being currently composed.

Another argument in support of having relevant information automatically appear in the context of the task is that we unanimously agree that information that the user doesn't remember having encountered can still have value. Such information is difficult to re-find, since the user does not even remember it exists. Pushing relevant information on the user could serve as an important reminding function.

Those who disagree with the statement, "People should not have to do any work at all to re-find," think that it is fine as an ideal, but entirely impractical as a solution. If a computer will never be able to perfectly guess the user's information needs, there will always be a need for information seeking tools. Thus, it is best put as a long term goal.

> *Long term goal: Make it so people don't have to re-find.*

Re-finding is always part of another task. It's reuse that matters, not re-finding.

6 participants agree, 0 disagree

The participants were unanimous in their belief that the only reason to re-find information is to use the information target to accomplish some task of which re-finding is only a step. This assertion supports the argument above that users should not have to do any work in order to re-find appropriate information. That is, the ideal system is one that, in the process of accomplishing some task, is able to suggest information that the person has already encountered when it is needed, without forcing the person to leave the task of interest in order to engage in re-finding behavior. In order to achieve the ideal, it

is important to understand the relationships between task types and potential information support for those tasks.

Medium term goal: Classification of tasks for which information support is important.

Pruning is good for re-finding. Support for losing is as important as support for re-finding.

3 participants agree, 3 disagree

There was considerable disagreement as to whether pruning information from an individual's personal information store would aid re-finding. The argument in favor of pruning is that people are currently subject to information overload, and do not want to have to interact with as much information as they do. By removing information from the user's information space, the user can feel more in control of that space and be better able to find important information nuggets.

Those that disagree with the statement, "pruning is good," believe that a good information management system can provide relief from information overload by *virtually* losing the information, while still retaining the data somewhere, should the user happen to need it in the future. Information can appear lost to the user without actually being removed from the computer.

A benefit to actually losing information that virtual loss would not provide is that it creates additional disk space. This is only a problem if disk space is constrained. It currently appears that this will not be a problem, but it could be an issue with the capture of large amounts of data about the user (e.g., continuous video feed of the user's life).

Conclusion

With respect to all of the assertions that we have posed, it is absolutely essential that there be a means by which they can be evaluated, and by which theories and techniques for understanding and addressing these issues can be tested. The most successful mode of evaluation to date in information retrieval research has been the use of test beds which allow many different investigations to be performed and compared (e.g. the TREC and MUC programs). Test beds for the evaluation of systems that support the re-finding of personal information will need to be substantially different from those that are currently available, since they will require knowledge of context, specification of task, and some record of interaction with information objects within task context. But such a resource would be enormously important for promoting real scientific research in this area, testing hypotheses, comparing approaches, and building on previous results.

Resources needed: An evaluation framework and methodology, and a testbed.

The validity of each of the above assertions is an open research question, and we throw this report into the future in hopes that researchers will catch it and be inspired to shed light on the statements.

References

1. Cool, C. and Belkin, N.J. (2002). A Classification of Interactions with Information. In *Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4),* 1-15.
2. Jones, W. (2005). Introductory Remarks, *PIM Workshop*, Seattle, WA, January 2005.

3. Lieberman, H., Rosenzweig, E. and Singh, P. (2001). Aria: An Agent for Annotating and Retrieving Images. *IEEE Computer*, July 2001, 57-61.
4. Miller, M.J. (2005) Google, Yahoo!, and MSN: The Search Continues. *PC Magazine*, March 22, 2005, 5.
5. Rhodes, B. and Starner, T. (1996). The Remembrance Agent: A Continuously Running Automated Information Retrieval System. In *Proceedings of 1st International Conference on the Practical Application of Intelligent Agents and Multiagent Technology (PAAM '96)*, 487-495.
6. Teevan, J., Alvarado, C., Ackerman, M.S. and Karger, D.R. (2004). The Perfect Search Engine is Not Enough: A Study of Orienteering Behavior in Directed Search. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '04)*, 415-422.
7. Yee, K-P., Swearingen, K., Li, K. and Hearst, M. (2003). Faceted Metadata for Image Search and Browsing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '03)*, 401-408.