Using the web to explore scientific knowledge and extend the desktop information space

Fernando Figueira Filho, Brendan Cleary
Dept. of Computer Science
University of Victoria, Canada
ffilho,bcleary@uvic.ca

Wendy Mackay
LRI, Bâtiment 650
Universitè Paris Sud, France
wendy.mackay@lri.fr

Paulo Lício de Geus
Institute of Computing
University of Campinas, Brazil
paulo@las.ic.unicamp.br

ABSTRACT
We conducted a study on how academic researchers manage multiple documents acquired from the web for later retrieval. We interviewed 11 participants and identified their strategies when trying to re-find specific documents. We found that they often prefer web-based search for re-finding documents, despite knowing that the document of interest is stored on their computers. We argue that Web search engines can act as an extension of the desktop information space. We found that users choose keyword-based search not only when the document's location is unknown but also when the retrieval cost is very low: they do not bother about properly storing files because most files are easily found again with a web-based search engine. We close by discussing the implication of these findings for the design of future document management tools.

Categories and Subject Descriptors
H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*web-based interaction, collaborative computing, organizational design*

General Terms
Human factors

Keywords
Information management, information-seeking behavior, web search, digital libraries, scientific reading.

1. INTRODUCTION
Users manipulate large numbers of documents in their work environments and often have trouble finding them again [8]. Studies have argued that users prefer folder navigation over keyword search to find their documents again [4, 12, 7], even if an "ideal" search engine is available [15]. For instance, a predominant finding is that people would not bother recalling relevant keywords to retrieve a document if they know its exact location on their desktops [7].

However, the Web has been providing users with a great variety of information sources and powerful search mechanisms. The readiness of information provided by these mechanisms might potentially relieve users of the need for keeping documents previously retrieved from the web organized, since that same information might be easily found again using web-based search engines [11]. Therefore, the problem of keeping previously-found information available ultimately reflects a dilemma from the user perspective: should she keep on top of her own file organization or should she rely on web-based search engines to re-find information as needed? The question gains relevance as we observe work activities that typically require the simultaneous manipulation of multiple documents, part of which might need to be found again.

This article examines that dilemma from the standpoint of academic researchers. On one hand, academics can benefit from the widespread use of digital indexing and the emergence of interoperable ontologies within many scientific disciplines, which allows such users to concurrently browse articles that are available in distinct digital libraries on the web [13]. On the other hand, digital libraries can be regarded as general information management (GIM) systems, which are designed to cater the different needs of many users [5], and as such may fail at providing effective tools for users to create and manage their own document organization.

This article is structured as follows: we begin by reviewing the related literature in Section 2. We then introduce our research questions in Section 3 and describe our study in Section 4, followed by the presentation of our findings in Section 5 and a discussion in Section 6. We finish by envisioning the future of PIM in Section 7 and by drawing a conclusion in Section 8.

2. RELATED WORK
Previous work on PIM has studied users' information management strategies within the desktop environment. Boardman and Sasse [8] argued that the perceived value of information influences the selection of PIM strategies. As such, users are willing to take the time to organize what they have often invested significant time in authoring, i.e. personal documents. Bergman et al. [7] hypothesized that providing an improved desktop search mechanism would increase the use of the search function and found only a limited effect: users continued to prefer browsing over search to re-find information. None of these studies, however, considered the

role of web-based search mechanisms and web information sources for re-finding.

Other studies went beyond the desktop environment, thus investigating users' information management practices over web information [11, 3, 9]. Aula et al. [3] studied the search and re-access strategies of experienced web users, showing evidence that users often make use of search engines on the Web to re-find material. Jones et al. [11] suggested that search engines are essentially maintenance free and therefore have the potential to ease the burden of keeping files organized for later retrieval. Capra et al. [9] argued that people use a variety of tools-at-hand to augment what search engines and current browsing software support.

However, although these studies have made significant observations, they have heavily focused on users' strategies to re-find *web pages*. In particular, we expect users to behave differently with regard to other types of documents: while web pages are typically re-accessed, other documents may need to be stored and eventually organized for direct manipulation on one's desktop environment. This brings other possibilities for re-finding that are not limited to web search engines, but can also include desktop search mechanisms, desktop applications such as bibliography managers and navigation strategies such as browsing through folders.

3. MOTIVATION AND RESEARCH QUESTIONS

We intended to better understand re-finding behavior in the context of knowledge-intensive domains. In this context, the work of full-time academic researchers involves multi-session searches across different search portals and is not limited to the manipulation of information on web pages, which brings novelty to the present work. We thus elicit the following research questions:

1. How do academic researchers evaluate the trade-off between (a) keeping documents organized for later retrieval or (b) giving up their own organization and relying on web-based search for re-finding?

2. How do academic researchers assess the usefulness of keyword search in comparison to location-based strategies for re-finding?

4. STUDY

4.1 Participants

We interviewed 11 participants (10 men, 1 woman) about their strategies for re-finding documents. All participants were full-time researchers in Informatics and worked in the same research laboratory. All were familiar with search engines and knew how to operate desktop tools and the file hierarchy. Nine used MacOS (Leopard and Snow Leopard), with the Finder as the file browser and Spotlight for desktop search. One participant used Windows Vista, with Windows Explorer as the file browser and the Windows Search Box for desktop search. One used Ubuntu Linux, with Nautilus as the file browser and Beagle for desktop search.

4.2 Procedure

We conducted semi-structured interviews in which we asked each participant about their recent experiences re-finding documents. Our aim was to reconstruct events of search failure and characterize potential issues that could arise from the interplay between desktop- and web-based strategies to re-find documents. For example, we asked them to remember a time when they had problems finding again a document. Each interview took approximately 25 minutes and was conducted in the participant's work place or in a separate room, if they shared an office. We asked participants to show us, on their computers, the steps they used to find particular documents, to help them remember the details of each search. We took notes and recorded each interview with a video camera.

4.3 Data analysis

The interviews were transcribed and analyzed using an open coding strategy [14]. As a result, we identified a central phenomenon of interest: when it comes to scientific articles, academics often prefer to use web search to re-find information, even though they may have stored the information at some point in the past on their personal computers. We then engaged in the axial coding process to identify specific coding categories that may be related or explain our central phenomenon. There was no pre-defined categories or theory prior to analysis.

5. FINDINGS

Participants in our study often found themselves struggling with the following decision: (a) store and organize documents acquired from the web for later retrieval or (b) re-finding these documents directly on the Web using web-based search engines. Some participants have developed their own strategies to deal with those situations in their routine work. In doing so, participants considered a balance between *costs* and *benefits* that were inherently associated with their strategies, as illustrated in Fig. 1.

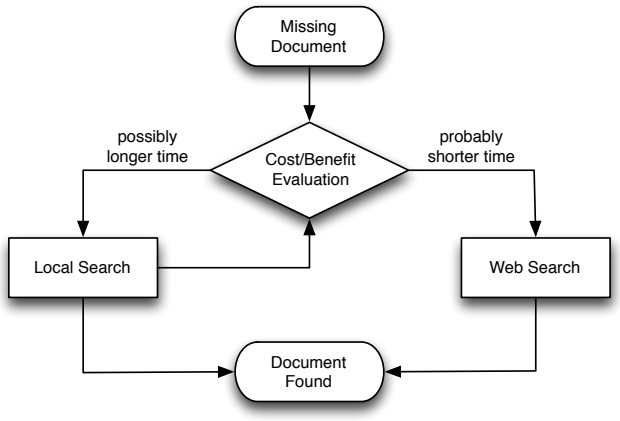

Figure 1: Re-finding process for academic researchers

5.1 Trade-off between organizing for later retrieval and direct re-finding

Research articles are inherently difficult to store because they contain multiple properties that do not map straight

away to the strict hierarchy found in nested folders. Participants thus evaluate the *organizing cost* when deciding how to organize documents for future retrieval: *"I don't store [papers] on my computer... I could rename them, but I'm lazy to do that and don't know how to rename them because [at retrieval time] maybe I don't know the title, I just know the content, sometimes I know the title, not the content."* Because it is sometimes difficult to store and keep track of large collections of documents using folders, many participants maintain clean file structures and avoid keeping things that can be downloaded again:*" I don't like having files locally on my computer, I prefer to get them online."*

However, in the case of working information [4], some participants reported a desire to organize documents, especially if these documents were related to a particular project: *"I save them all together, for example, if they are related to a particular project."* In this context, Boardman and Sasse [8] pointed out that file organization is more worthwhile since the cost of filing is offset by predicted benefits at retrieval time. We found that this argument is not entirely true in the case of academic researchers. For instance, the benefits at retrieval time eventually will not pay off the costs of management, and some participants reported to adapt their organization strategies to deal with large collections over time: *"When I'm starting a project or when I need related work, I install the pdf files in folders, I use colors, to know the ones that I have already read... the problem is that this [organization] never goes really far, at some point I just forget to keep it updated."*

Managing multiple projects simultaneously might lead to project fragmentation problems [6], thus decreasing the usefulness of file organization and increasing the effort to navigate the file hierarchy as the collection of documents becomes larger. In particular, navigation requires covering a distance between two locations within a given information space, in general by following some sort of orienteering strategy [15], such as browsing through a folder hierarchy. Even in the case when the user knows the document's location in advance, longer distances among folders take more time to navigate: *"This is something very important... you have to save... or to load the image from a particular directory and it takes a long time to navigate the whole [file] hierarchy to go to this directory."* To overcome this problem, participants preferred to use web search to re-find documents: *"In fact I would say that many times I try to look on the web first, so I don't save them [papers] on the computer, unless it is a group of things... Because I won't use bookmarks."*

5.2 Usefulness of keyword search for PIM

As an alternative to navigation, users also consider using desktop tools to search over their collections. Previous work on PIM argued for the limited usefulness of keyword search on the desktop environment [15, 7]. However, in our study, we found that, despite both desktop and web search tools sharing the same basic functionality—one enters a query and receives back a result list—their usefulness is perceived differently by each participant according to some factors. Search efficiency is one of them: *"The search function in Mac is really bad... Because it gives you everything. So, you know, by using the exact keywords, you find everything related to [the paper], but not the one you want."*

To overcome this problem, some users reported using the keyword search function typically available on bibliography managers: *"I store [scientific articles] in the same folder and then when I import [a new article] to my bibTeX, I also associate the new entry with its respective PDF file, so that I keep track of where it is."* However, as opposed to desktop-based tools, digital libraries available on the web provide users with more than just an indexing mechanism. Participants reported using various features that are not readily available on their desktops, such as browsing through links to related references: *"I'm used to research them [scientific articles] on the ACM [Library], rather them finding them in my personal files. Because I'm used to brainstorm over the subject using keywords, you know, when you are looking for [scientific] papers and then you can get the related ones."*

6. DISCUSSION

Our results indicate several points that should be carefully considered when drawing conclusions upon users' practices to manage information in the workplace setting. First, today's widespread use of web applications have modified the way people work with digital documents. As such, focusing the investigation exclusively on desktop-based tools such as folder hierarchies seems to ignore a variety of practices that are carried out outside one's own file organization. Second, new PIM tools are emerging within the web environment. Our results indicate that re-finding digital documents elicits a different behavior when compared to re-finding web pages: the former is explicitly manipulated in users' personal computing environments and may require some organization effort, as opposed to the latter. Third, our study population was chosen among people with a distinct feature: academics have a need to manage multiple documents simultaneously to get their work done. Although this brings threats to the validity of our results, it helped us to highlight a phenomenon of increasing significance: the boundary of today's desktop environment cannot be considered in the same way it was a couple of decades ago.

However, issues of integration between desktop and web-based tools might arise when users need to re-find documents that were previously manipulated. For instance, digital libraries are general information management (GIM) systems, in which information professionals, e.g. librarians, manage the available information for a range of other people [5]. This poses a significant gap between the personal computing environments and GIM systems. While in the former users have total control over their documents' organization, in the latter this control is delegated to others. As a result, although many participants in our study reported the need to keep their file hierarchies tidy, e.g. for a project, none could report an efficient way to do the same on the web. Furthermore, our participants considered web-based search engines and digital libraries as a means to re-find information efficiently, but the state-of-the-art of these tools does not provide users with features to organize the retrieved information. This is an indication that future tools for information management would require the power of today's web-based search mechanisms to re-find information, along with efficient ways to categorize retrieved information into collections that meet the needs of users with distinct requirements.

7. FUTURE OF PIM

The combination of benefits from web- and desktop-based tools might be found in the next generation of bibliography managers. Tools such as Zotero [2] and Papers for Mac [1] provide an integrated environment to retrieve, organize and re-find scientific articles. However, none of our participants reported to use them. It might be the case that those tools still cannot replace both desktop and web-based tools simultaneously. Or perhaps the strategies used by academics with traditional tools, e.g. folders and digital libraries, are so well integrated in their work routines, that they do not feel the need to change their current methods.

Either way, we envision a future with greater integration between social networking tools and personal information management tools. Social networks may work as an alternative to keyword search [10], by distributing the problem of information filtering and content analysis among peers within the scope of a social network, i.e. social search. The approach also takes advantage of the fact that colleagues might share tacit knowledge to some degree, which might help to build content upon subjective information needs that cannot be easily expressed using plain queries and keywords.

Other possibilities in terms of personal information management are emerging with the popularization of smaller devices, such as the iPad. Although these technologies facilitate the integration across different information silos and devices, their utility for academic work is not yet understood. The popularization of web-based storage tools and cloud applications might eliminate the need for organizing documents across distinct devices and personal computing environments. The increasing adoption of these technologies calls for the revisitation of previous findings in the PIM literature. Further empirical research is needed to characterize the effects of those technological changes in the context of personal information management.

8. CONCLUSION

We present the findings of a study that aims to bridge the gap in the literature by (i) investigating how people manage multiple documents acquired from the web for later retrieval and (ii) understanding the role of keyword search by considering web search engines as potential retrieval tools. We found that keyword search tools are useful not only when the document's location is unknown, but also when retrieval costs are so inexpensive that users do not bother storing documents that could be easily retrieved using web-based search engines. Web-based search is a tool that can be used across applications and activities, which adds to its power. However, we found that users still have a need to organize certain types of documents, especially if they are related to a project. In this case, web-based search mechanisms and digital libraries lack efficient ways for users to organize information.

9. REFERENCES

[1] Papers for mac. Available online: `http://www.mekentosj.com/papers/`. Last access: 11/25/2011.

[2] Zotero. Available online: `http://www.zotero.org`. Last access: 11/25/2011.

[3] A. Aula, N. Jhaveri, and M. Käki. Information search and re-access strategies of experienced web users. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 583–592, New York, NY, USA, 2005. ACM.

[4] D. Barreau and B. A. Nardi. Finding and reminding: file organization from the desktop. *SIGCHI Bull.*, 27(3):39–43, 1995.

[5] O. Bergman, R. Beyth-Marom, and R. Nachmias. The user-subjective approach to personal information management systems. *J. Am. Soc. Inf. Sci. Technol.*, 54:872–878, June 2003.

[6] O. Bergman, R. Beyth-Marom, and R. Nachmias. The project fragmentation problem in personal information management. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 271–274, New York, NY, USA, 2006. ACM.

[7] O. Bergman, R. Beyth-Marom, R. Nachmias, N. Gradovitch, and S. Whittaker. Improved search engines and navigation preference in personal information management. *ACM Trans. Inf. Syst.*, 26(4):1–24, 2008.

[8] R. Boardman and M. A. Sasse. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 583–590, Vienna, Austria, 2004. ACM.

[9] R. Capra, G. Marchionini, J. Velasco-Martin, and K. Muller. Tools-at-hand and learning in multi-session, collaborative search. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 951–960, New York, NY, USA, 2010. ACM.

[10] F. M. Figueira Filho, G. M. Olson, and P. L. de Geus. Kolline: a task-oriented system for collaborative information seeking. In *Proceedings of the 28th ACM International Conference on Design of Communication (ACM SIGDOC 2010)*, pages 89–94, September 2010.

[11] W. Jones, H. Bruce, and S. Dumais. Keeping found things found on the web. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 119–126, New York, NY, USA, 2001. ACM.

[12] W. Jones, A. J. Phuwanartnurak, R. Gill, and H. Bruce. Don't take my folders away!: organizing personal information to get things done. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1505–1508, Portland, OR, USA, 2005. ACM.

[13] A. H. Renear and C. L. Palmer. Strategic reading, ontologies, and the future of scientific publishing. *Science*, 325:828–832, August 2009.

[14] A. L. Strauss and J. M. Corbin. *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage Publications, 1998.

[15] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422, Vienna, Austria, 2004. ACM.