Write While You Search: Ambient Searching of a Digital Library in the Context of Writing

Anatoliy Gruzd
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, Illinois 61820 USA
agruzd2@uiuc.edu

Michael Twidale
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, Illinois 61820 USA
twidale@uiuc.edu

ABSTRACT
We consider ideas for a tighter integration of searching a digital library while writing a paper. A prototype system based on web services is described which allows us to explore the design space of ambient search tools to support and inspire the writing process.

Categories and Subject Descriptors
H.3.5 [**Information Systems**]: Digital Libraries;
H.5.2 [**Information Systems**]: User Interfaces

General Terms
Algorithms, Documentation, Design, Human Factors

Keywords
Writing support, ambient search, integrating search with writing

1. INTRODUCTION
We are interested in examining ways to more tightly integrate information search and use in the context of writing. We particularly focus on scholarly writing by students and researchers, but also consider related work involving the analysis and synthesis of information from a range of sources into some kind of written report. Conventionally this work has been divided into a number of somewhat distinct activities:

1) Searching for the information
2) Organizing, analyzing, systematizing, synthesizing, obtaining insights, planning the report
3) Writing the report

These three are typically done sequentially in the order shown. Digital Library (DL) systems and researchers in Information Retrieval have mostly looked to support 1 and parts of 2. Conventionally word processors and researchers in computers and writing have only looked to support 3 and parts of 2. We want to explore the development of tools that explicitly support all three, enabling people to do them in any order and switching between them as they like. That is, we want to explore re-contextualizing information search into the process of writing. We acknowledge that this is not universally appropriate or desirable, but we do believe there are circumstances when it might be very helpful.

Of course skilled writers have always used techniques to integrate searching, analyzing and writing more tightly. These can range from strewing papers around a desk and the floor to juggling multiple windows and applications on a PC. However these approaches can be hard to manage, meaning that because of the constraints of the available tools some people compartmentalize the actions rather than integrating and hence contextualizing them.

We do not have a particular solution to advocate for. Rather we are exploring a design space of possible functionalities and interfaces in order to more fully understand the dimensions of the problem. The prototype reported here is just a single data point in this space, developed to enable us to explore the space in a more purposeful manner. It has been developed in a context of rapid prototyping, exploiting a number of different web services, using a mashup style, linked together via Python and JavaScript.

2. INTEGRATING INFORMATION SEARCHING AND WRITING
"The more research you do, the more impossible it is to start writing" [5]

Writing, particularly academic writing, can be a challenge for experts and novices alike. Digital libraries have greatly improved the ease with which we can search for information, even from our desktop. It is easier to integrate searching and writing activities when both are done in different windows on the same PC. Nevertheless, the act of writing remains difficult. The very accessibility of so much information through ever more complete DLs with ever more sophisticated search functionalities can mean that searching and reading articles turns into something of a displacement activity, postponing the dreadful moment of starting work on the paper. Sadly this kind of problem is one more likely to be experienced by the more diligent, perfectionist student, a personality strait particularly evident at the graduate level.

There is a literature in writing studies that advocates for a tighter integration of writing and searching as a way to improve the quality of final papers. It is important to start writing at the early stages of one's research because the writing process can be considered as such that stimulates learning. Emig [7], for example, studying various definitions of the learning process by some of the most influential psychologists of the 20th century, discovered clear correspondences between writing and learning. Nelson & Hayes [12] found that more experienced writers were inclined to employ an issue-driven (writing down preliminary thoughts, looking for supportive sources, reading) rather than a content-driven (exhaustive information search, reading, and only

then writing) approach. We refer to the 'issue-driven' approach as '*write while you search*'. We wondered if a tool that encouraged a focus on writing, but provided a background, almost ambient search functionality would be helpful and encourage researchers to change their habits to this more productive approach.

There is little theoretical foundation to help design such a tool. To the best of our knowledge, no one model explicitly describes information searching and writing processes within the same framework of the user's information behavior. At first glance, Kuhlthau's model of Informational Seeking [11] may seem to cover the integrated approach to searching and writing. However, while this model provides some initial explanation to how task and information seeking behavior may interrelate, it does not model a work task generally [9] or the writing process specifically [6]. On the other hand, those models that do attempt to consider a work task such as Vakkari's task-based Information Retrieval model [14] are too broad to cover each individual task like writing.

3. RELATED WORK

In recent years Digital Library researchers have called for the support of users' actual information need(s) and their primary work task when designing DLs. Examples include a Dynamic Review Journal framework to "[assist] authors in collating and analysing experimental results, organising internal project discussions, and producing papers" [4] and Garnet, a spatial hypertext interface to DLs which "provides an integrated environment for both seeking and organising information" [2].

More specifically, aspects of the issues noted above are supported by Query-Free Proactive Retrieval Agents, also known as Reconnaissance or Just-in-Time Information Retrieval agents. These tools aim to support the finding of new relevant information based on contextual information about what the user is currently doing, typically reading. Some of them exploit the context of what the user is currently writing as an input source (e.g. [3], [10], [13]).

Our system is related to these systems but with more of a focus on encouraging and inspiring writing rather than being chiefly a powerful information retrieval agent requiring minimal effort from the user.

4. SCENARIO OF USE

The following illustrates our ideas about integrating ambient search of a digital library into exploratory writing activities, using PIRA (Personal Information Research Assistant).

John is a graduate student. Three weeks ago he received a writing assignment for his Information Retrieval class. After some preliminary literature review in the XX Digital Library, he decides to write a paper on the use of Natural Language Processing (NLP) in Information Retrieval (IR) systems. The deadline for this paper is rapidly approaching, but John has not started writing yet.

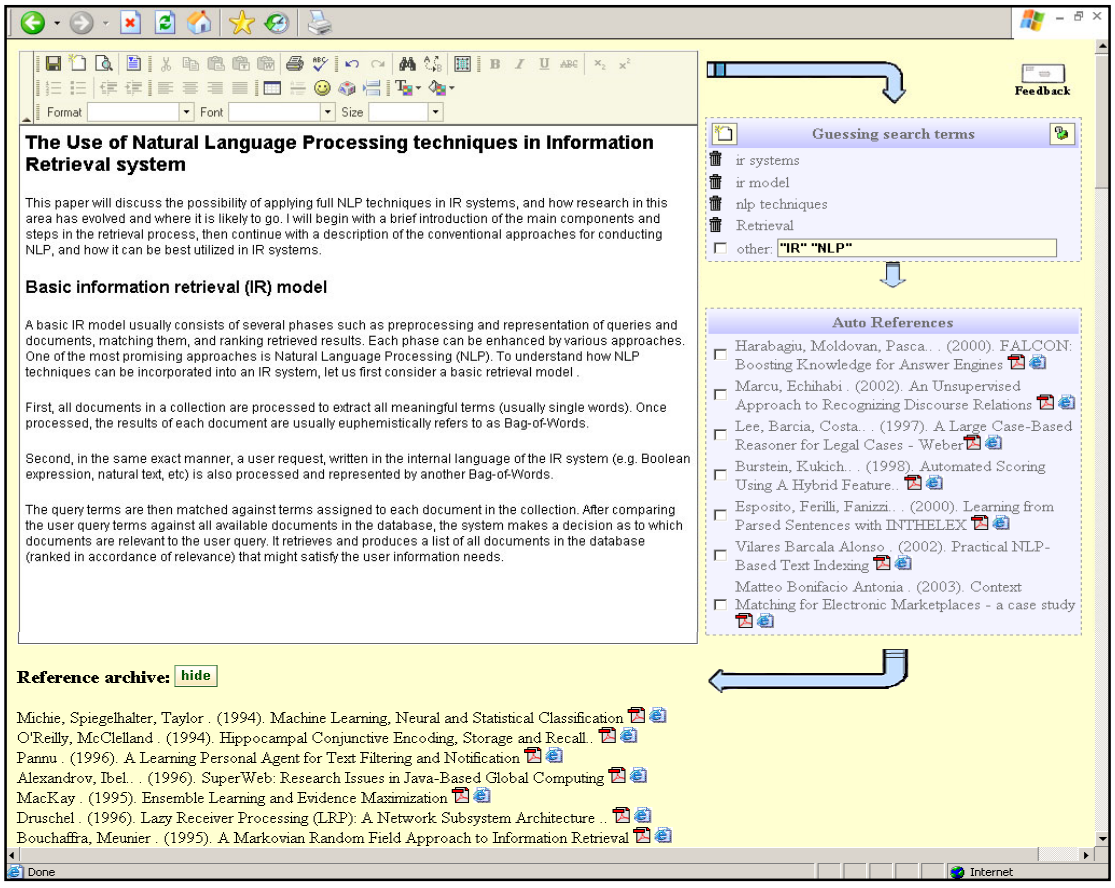

Figure 1. PIRA main display showing integration of writing and ambient searching of CiteSeer.

John decides to try PIRA. He opens up Internet Explorer (the browser it currently works best with) and goes to http://www.writeNcite.com (Figure 1). He sees a page divided into two parts: a text editor with an MS Word-like interface and a much smaller part – a search pane. The arrows indicate the information flow and John realizes that he needs to begin writing first to get some recommendations from PIRA. He starts free writing about the topics he thinks his paper should be about. In his own words, he says he wants to investigate the advantages NLP techniques may provide to IR systems. Then he stops. He does not know what to say next.

John notices that PIRA has already compiled a list of some articles related to his topic. As there are only 7 retrieved articles (and not much detail – just abbreviated author and title information), it is easy to quickly look over them. Each article can be considered to be in one of four distinct groups:

1) Irrelevant;
2) Interesting but not sure if relevant;
3) Relevant but for later use;
4) Relevant and can be used right away.

For example, a title "Learning from Parsed Sentences with INTHELEX" looks interesting, but after John hovers over it and reads the abstract (Figure 2) it becomes evident that although it is about NLP techniques it is not specifically applied to an IR task, so he quickly mentally rejects this one. He doesn't have to do anything with the search interface; that reference will eventually fall off the end into the archive, and something new will appear. Some background papers seem to be very relevant, but John does not want spend his time reading them right now. So using checkboxes, he selects them for later use. They will now stick around and not move on into the archive. Finally, the other three articles on the list could be used in the paper right away. Furthermore, they give John an idea how to develop his argument. He may begin his paper by talking about how different IR applications benefit from using NLP techniques.

PIRA provides direct links to full-text articles in the DL in either PDF or HTML formats. While skimming through the text of articles, John copies and pastes descriptions of each system into his paper. This process is easy as PIRA automatically adds quotation marks and citations information of copied quotes. If John doesn't like one of the search terms, or considers he has enough information specifically relating to that one, he can delete it and PIRA will bring in another one. In all cases if John does nothing, he will get a gradual stream of suggestions that might be helpful or inspirational. If he later realizes that he had noticed something that was there once, he can recover it by clicking on the archive button at the bottom of the screen.

5. PIRA: GENERAL DESCRIPTION

One of the main motivations of designing PIRA is to allow users to focus on their primary writing task, using access to a DL as a source for inspiration as the writing proceeds. To achieve this, PIRA takes some initiative and conducts information search in the background while the user is working on something else. When the user needs some additional information to support her writing, something is already available without the need to conduct a search. We call this passive mode ambient search. In

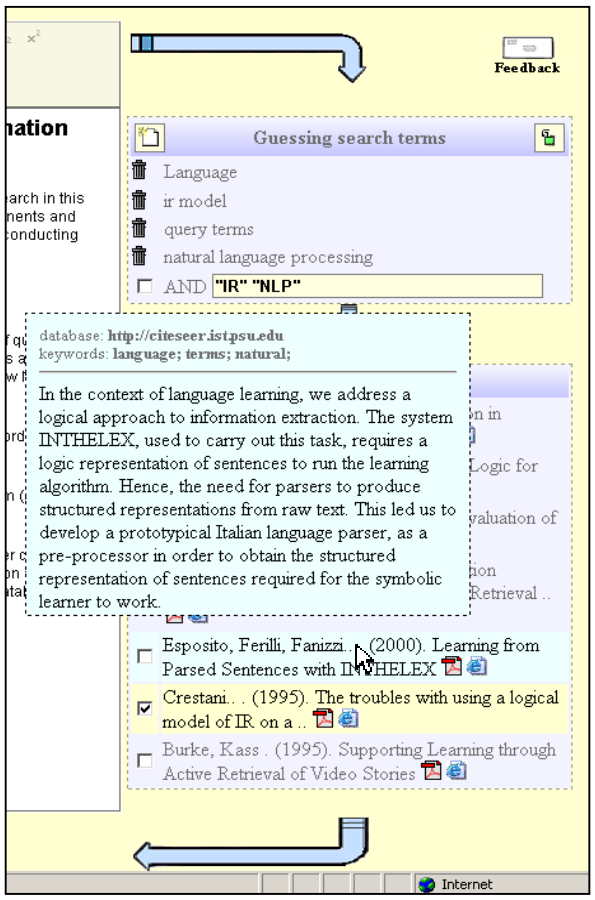

Figure 2. PIRA search pane with popup abstract.

some ways it is a bit like a news crawl or updated weather information on various desktop displays.

Due to the complexity in interpreting exactly what a particular user might be searching for and for what purpose(s), it is inevitable that complete automation is bound to fail. Keeping this in mind, PIRA provides a list of the search terms that it has guessed and is using in its searching of the DL, and allows the user to veto them. To minimize effort, the interaction is optional and mostly negative – the user just rejects a term by clicking on the bin icon – rather than having to think of and enter new ones. (There is also an option for adding terms, but it gets rather close to active search for which it may make more sense to switch to the conventional DL search interface.) We call this mode of a system involvement as an active mode that is initiated by the user.

In short, PIRA operates in both modes of system involvement in a search process. To formally characterize these two modes, we use the scale developed by Bates [1] to describe the degree of user versus system involvement in the search. She defines five levels of system involvement (from 0-no system involvement to 4-fully autonomous system). According to this scale, PIRA is working on both a level 2 involvement (search conducted by the user's request) and 3 involvement (monitor and suggest).

6. HOW PIRA WORKS

While the user is working on her paper, PIRA periodically passes the text to the Yahoo! Term Extraction Web Service (developer.yahoo.com/search/content/V1/termExtraction.html). This provides a set of concept terms. Those extracted from the current paragraph (defined by cursor position in the text) receive higher weights than those from the rest of the available text, and as a result are more likely to appear in a search query. The list of concept terms is expanded by adding related terms retrieved by submitting concept terms to a clustering engine (clusty.com).

Next, the three groups of terms (from the current paragraph, from the other paragraphs, and related terms) are combined into one list. The top four terms from this list are sent to CiteSeer (citeseer.ist.psu.edu). CiteSeer first searches for documents matching all submitted terms. If no matches are found, it attempts to retrieve some results by performing a simple keyword search. The remaining terms in the combined list await their turn. When a user manually removes one of the top four active terms, PIRA draws a new term from the waiting list.

During writing, the number and nature of ideas and arguments in a draft is constantly changing. Every new idea or argument may stimulate a new information need(s). To make PIRA more receptive to the appearance of new ideas and arguments in the paper, we adopted a query term aging mechanism similar to the history algorithm used by Henzinger et al [8]. As a paper evolves and new ideas appear, each active term gradually ages and is replaced by a new one from the waiting list. The aging rate depends on how fast new concepts are introduced in the paper. PIRA takes the top seven results ranked and returned by CiteSeer, and displays them. If a user sees something relevant, she can either access the full-text article right away or save a desired item for later. If neither option is selected, the item becomes a candidate to be moved to the reference archive. The reference archive (a temporal list of results that had been displayed) can be accessed at any time. CiteSeer does not provide access to papers in HTML format, so PIRA (when possible) converts PDF files to HTML.

7. CONCLUSION

The use of web services has allowed us to begin to explore the design space of features and interfaces that might support aspects of ambient searching. As noted, we do not believe we have already found a particular or even a stable solution. But we do believe that there is a category of tools, of which PIRA is an exemplar, that support a certain kind of searching and using a digital library that is more tightly integrated with certain activities than a more generic search interface. We have been considering the area of writing support and writing while searching, where the main power of a digital library is to act as a source of inspiration without diverting attention completely away from the act of writing. Much work remains to be done in ongoing iterative design and evaluation to understand the impact of the parameters of the algorithms used and the effect of interface and functionality elements on usability and usefulness. We are well aware that there are circumstances where a PIRA-like resource will be confusing, distracting, intrusive, of liable to reinforce bad habits. However we do believe that a refined version has great potential in certain circumstances and that the speed and power of developing using web services makes it more feasible to consider these more niche uses than has been feasible when developing more generic, monolithic search interfaces to Digital Libraries.

REFERENCES

[1] Bates, M.J. Where Should the Person Stop and the Information Search Interface Start? *Information Processing and Management* 26, 5 (1990), 575-91.

[2] Buchanan, G., Blandford, A., Thimbleby, H. W. and Jones, M. Supporting Information Structuring in a Digital Library. *Lecture Notes in Computer Science* (2004), 464-475.

[3] Budzik, J., Bradshaw, S., Fu, X., and Hammond, K. Supporting Online Resource Discovery in the Context of Ongoing Tasks with Proactive Software Assistants. *International Journal of Human-Computer Studies* 56, 1 (2002), 47-74.

[4] Carr, L., Miles-Board, T., Wills, G., Power, G., Bailey, C., Hall, W. and Grange, S. Extending the Role of the Digital Library: Computer Support for Creating Articles. In *Proceedings of the 15th ACM conference on Hyptertext and Hypermedia* (2004), 12-21.

[5] Elbow, P. *Writing with Power: Techniques for Mastering the Writing Process*. 2nd ed. New York: Oxford University Press, 1998.

[6] Elmborg, J. K., and Hook, S. E. *Centers for Learning: Writing Centers and Libraries in Collaboration.* Chicago: Association of College and Research Libraries, 2005.

[7] Emig, J. Writing as a Model of Learning. *College Composition and Communication 28*, 2, (1977), 122-128.

[8] Henzinger, M., Chang, Bay-Wei, Milch, B. and Brin, S. Query-Free News Search. In *Proceedings of the 12th international conference on World Wide Web*. Budapest, Hungary: ACM Press (2003), 1-10.

[9] Ingwersen, P. and Järvelin, K. *The Turn: Integration of Information Seeking and Retrieval in Context*. Dordrecht, The Netherlands: Springer, 2005.

[10] Jones, S. and Staveley, M.S. Phrasier: A System for Interactive Document Retrieval Using Keyphrases. In *Proceedings of the 22nd ACM SIGIR conference on information retrieval* (1999), 160-167.

[11] Kuhlthau, C.C. *Seeking Meaning: a Process Approach to Library and Information Services*. 2nd ed. Westport, CT: Libraries Unlimited, 2004.

[12] Nelson, J. and Hayes, J.R. *How the Writing Context Shapes College Students' Strategies for Writing from Sources.* (Technical Report No. 16). Berkeley, CA: Center for the Study of Writing, UC Berkeley, 1988.

[13] Rhodes, B.J. and Maes, P. Just-in-Time Information Retrieval Agents. *IBM Systems Journal 39*, 3-4 (2000), 685-704.

[14] Vakkari, P. Changes in Search Tactics and Relevance Judgments when Preparing a Research Proposal: A Summary of the Findings of a Longitudinal Study. *Information Retrieval 4*, 3-4, (2001), 295-31.