A Personalized Recommender System for Writing in the Internet Age

Mari Carmen Puerta Melguizo*, Olga Muñoz Ramos*, Lou Boves*, Toine Bogers†, Antal van den Bosch†

*Department of Language and Speech, Radboud University.
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands.
M.Puerta, O.MunozRamos, L.Boves@let.ru.nl

†ILK/Language and Information Science, Tilburg University
P.O. Box 90153 NL 5000 LE Tilburg, The Netherlands
A.M.Bogers, Antal.vdnBosch@uvt.nl

Abstract

Writing is a complex task and several computer systems have been developed in order to support writing. Most of these systems, however, are mainly designed with the purpose of supporting the processes of planning, organizing and connecting ideas. In general, these systems help writers to formulate external visual representations of their ideas and connections of the main topics that should be addressed in the paper, sequence of the sections, etc. With the advent of the world wide web, writing and finding information for the written text has become increasingly intertwined. Consequently, it is necessary to develop systems able to support the task of finding relevant information during writing, without interfering with the writing process proper. In this paper we present the Proactive Recommender System: À propos. This system is being developed in order to support writers in the difficult task of finding appropriate relevant information during writing. We raise the question whether the tendency to interleave (re)search and writing implies a need for developing more comprehensive models of the cognitive processes involved in writing scientific and policy papers.

1. Introduction

Writing in a professional environment is a difficult task. Although writing has been practiced for more than 25 centuries, empirical research of the writing process only started some 50 years ago. The first broadly accepted model of the cognitive processes involved in writing was the one proposed by Hayes and Flower developed in the early 80s (Hayes and Flower, 1980). Because text processors were not widely available at that time, it comes as no surprise that they model the cognitive processes involved in writing with pen and paper. Furthermore, since almost all research of the writing process has been conducted in laboratory settings where subjects had to produce short essays, one may ask whether the model can also be applied to writing research papers and policy documents. Finally, the criteria to assess the quality of a short essay are probably very different from those used to assess professional papers. For one thing, writers of professional documents must include or refer to all relevant information that is known, while writers of a short essay are only supposed to cover some items they deem relevant for their exposé. Yet, virtually all software for supporting text production seems to build on the concepts developed in pen-and-paper research.

In this paper we first explain the Hayes and Flower model and its later additions. Then we will relate the model to what is considered good practice for writing in the Internet age, and analyze the ways in which existing writing tools facilitate the tasks and in which ways these tools can be improved. We will illustrate our arguments with a Proactive Recommender System: À Propos. This system is being developed in order to support writers in the difficult task of finding relevant information during writing.

2. The cognitive processes involved in writing

2.1. The Model of Hayes and Flower (1980)

Since the beginning of empirical research scholars have agreed that writing involves at least three different cognitive processes, usually called 'planning', 'translation' and 'review/editing'. Hayes and Flower (1980) propose that there is a recursive interaction between planning, translation and review/editing. The model defines three components: the writing process proper (which includes the three processes/stages mentioned above), the task environment and the writer's long-term memory (see Fig. 1).

Planning involves retrieving domain knowledge from the writer's Long-Term memory (LTM) and organizing it into a plan that specifies, among other things, the effects that the writer wants (or needs) the text to have. During the process of Translating writer's plans and goals are transformed into sentences. In the Reviewing stage the writer evaluates the relation between the text written so far and the linguistic, semantic and pragmatic aspects that best serve the goal.

The task environment includes everything existing outside the writers' mind and that can influence the writing task. The main elements included here are the text produced so far and the so called rhetorical problem (the writing assignment, the specification of topic and the audience).

In the writer's LTM are stored the writer's knowledge about the topic, the knowledge of sources based on literature search, the writing plans and the knowledge about the audience who will read the work.

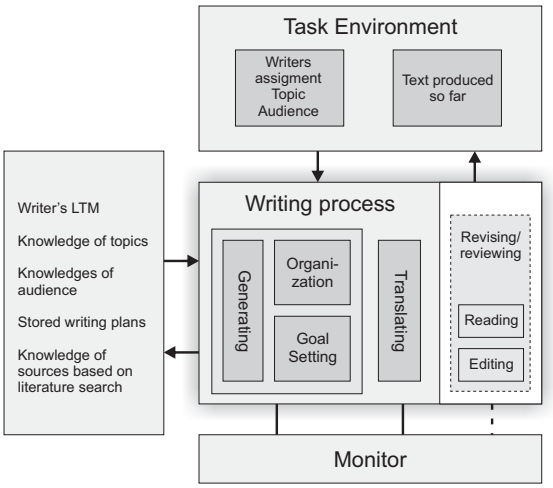

Figure 1: The Model of Writing proposed by Hayes and Flower (1980)

2.2. The Revised Model by Hayes (1996)

Later on, Hayes (1996) revised the model and emphasized the role of working memory, as well as socio-cultural and motivational aspects in writing. The main components are now the task environment and the individual (see Fig. 2). In the new model the task environment is divided into social and physical contexts. The social environment needs to be considered because the way a text is written ought to be affected by the audience it is meant for.

In the physical environment the composing medium or tool used to write is included. Variations in the medium seem to lead to differences in the way people carry out the writing task. For example, Haas (1996) found that writers tend to plan more when they write on paper than with a word processor, presumably because it is easier to sketch, draw and interconnect ideas using pen and paper. Haas also found that writers tend to revise documents on general level (i.e. modifying the structure) when using a pen, and more on a local level (i.e. revision of syntax, semantics, vocabulary) when working on screen. These results suggest that the introduction of computer tools as a medium for writing entice users to change the processes, rather than making the original processes easier or more effective. However, it should be noted that the revised model of Hayes still seems to deal only with writing short essays with pen and paper or with a stand-alone text processor and without the need to look for external information.

The components motivation/affect, cognitive processes, working memory and long-term memory are included in the part that models the writing individual. The main cognitive processes of writing are now text interpretation, reflection and text production. Text interpretation refers to the creation of internal representations based on linguistic and also graphic input. Planning has been replaced with the more general cognitive function of reflection that includes processes of problem solving, decision making and inference. Text production refers of course, to the act itself of producing written texts.

The working memory is also included in the model and in addition to its storage function, it performs a set of control functions. In a similar way, the model of Kellog (1996) integrates working memory into writing. During the writing process information from LTM is retrieved and deposited into working memory.

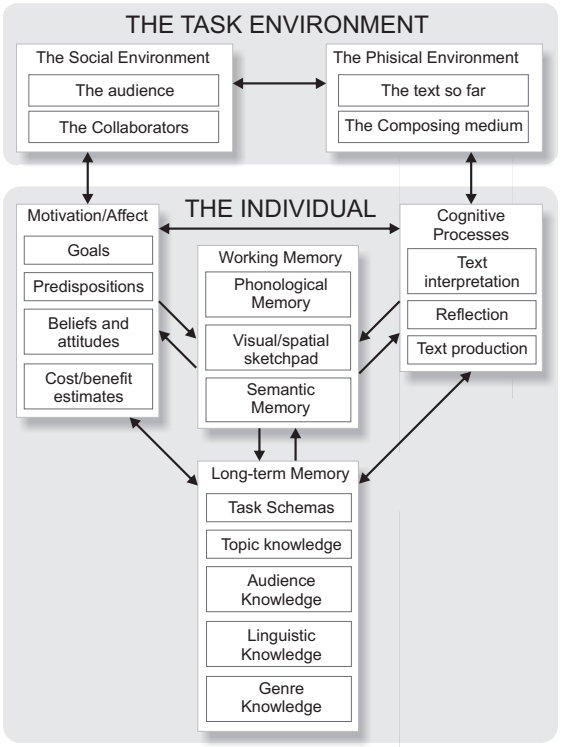

Figure 2: The Model of Writing proposed by Hayes (1996)

2.3. The Role of Long-term Memory and the Need to Look for New Information

Current models of writing assume that knowledge about the topic of the text is available in the writer's neural LTM. The reality of writing professional texts, however shows that writers almost invariably need to look for additional external information. Sometimes it is just very difficult to remember and retrieve information that is already stored in the writer's LTM. The need for searching or verifying detailed factual data is especially important during the reviewing phase. At other times the writer does not know about a specific bit of information and needs to consult external sources.

Of course, the need to do research before writing a paper existed before the advent of digital computers and text processors. Most probably, the idea that Planning should precede Translation derives from that practice. However, the advent of the Internet has caused a dramatic change in the process of writing. More often than not, writing is now interleaved with searching for information. Searching while writing is so easy that few writers can resist it. The question remains, however, whether continuously switching between the tasks of writing and searching is efficient, and whether it tends to result in the best possible quality of the texts. Because it is now common practice to interleave writing and searching, it is not only necessary to design tools

that support the cognitive processes involved in writing, but also tools that help users in finding information that writers cannot retrieve from their own memory.
Most computer writing environments (Britton and Glynn, 1989) seem to have been designed with the purpose of supporting the process of reflection. In general, systems help writers with grammar and spelling or to formulate external representations of their ideas and connect the main issues that should be covered, the ordering of the sections, etc. The Writing Environment (WE) developed by Smith et al. (1987) was one of the first to support the processes of organization and editing by means of a network mode, a tree mode, an editing mode, and a text mode. Another example is the tool developed by Neuwirth and Kaufer (1989), which offers visual representations that help to organize notes and graph summaries.

3. Writing in the internet age: finding related information during writing

The World Wide Web is a rich source of information about virtually any topic and looking for information on websites has become the most popular way to access information. Search engines have become the primary tool for information access in the Internet and company-internal networks. However, searching and navigating is often not as efficient and easy as users might like. Users frequently feel "lost in hyperspace". Feeling lost or the tendency to lose one's sense of location is the most common problem users experience while navigating and strongly affects user's performance and satisfaction (Conklin, 1987; Gwizdka and Spence, 2007; Puerta Melguizo et al., 2006). Ironically, the very potential strength of hypertext, namely improving the management of loose collections of relatively unstructured information, turned out to be a major potential weakness (Neuwirth and Kaufer, 1989).
Furthermore, keyword-based search is inefficient and relevant information may be missed because the writer did not realize that the information exists and could be looked up. Considerable time is spent interacting with low-precision search engines. Consequently, the time in which the author is away from creating the document can affect the total time spent, and the eventual quality of the text. Switching from the text editor to the search engine imposes extra demands on the user's cognitive capacities. A system that can relieve authors from explicit search and switching between applications by means of searching information accurately and recommending this information in a proactive manner would be most welcome.

4. À Propos: a proactive recommender system

The main goal of Proactive Recommendation Systems (PRSs) is to consult large quantities of documents, decide what available information is most relevant for the task, and offer that information without user requests. The decision about what information to offer is mainly based on the text that is currently being written, in combination with personal profiles and profiles of the person's working group. A few PRSs for writing have been developed. For example, the Remembrance Agent (Rhodes, 2000) suggests personal email and documents. Watson (Budzik and Hammond, 1999) performs automatic Web searches based on text being written or read. A problem with these PRSs is that they are developed as search support tools and do not seem to take into account the specific characteristics of the writing task, which can be seriously affected by any type of interruption from the environment.
The goal of the project À Propos is to develop a proactive, adaptive, personalized, just-in-time knowledge management environment for writers in a professional environment. The architecture of À Propos is inspired by other PRSs such as Watson (Budzik and Hammond, 1999) and Stuff I've Seen (Dumais et al., 2003). A detailed description of the system's architecture can be found in (Puerta Melguizo et al., 2007) where the role of the different components of the system such as observers, filters and gatekeepers is explained.
Deshpande found that two main issues need to be addressed if a PRS for writing is to assist rather than distract the users (Deshpande et al., 2006). First, proactive suggestions must be extremely accurate. Second, procedures to identify the different writing stages and related information needs must be created in order to design an appropriate user's interface.

4.1. Selecting and Presenting Relevant Information

Recommendations should be both on topic and personalized. To increase the topicality of suggestions one can use detailed personalized taxonomies integrated in an easily expandable, yet robust IR model (such as the Vector Space model). Ideally, personalization should go so far that two users with different interests writing or reading the same document should get different personalized recommendations. We are investigating two different types of personalization: on the user level and on the group level.

4.1.1. User personalization

From the user perspective we only consider evidence of the user's interests and expertise. From these data we build a personal profile of terms which is used to re-rank the initial recommendations list. Three different sources of information are considered for this purpose. First, the important terms of previously selected documents are added to the user's personal profile. Second, important terms in the user's past documents are given different weights, dependent on whether they were written by the user or merely read. Finally, the PRS allows users to enter informational queries manually; the important query terms entered explicitly are also be included into the user profile.

4.1.2. Group personalization

Group personalization is done on the basis of the expertise of the members of a group. Not every group member has an equal level of expertise or interest in the specific topic of the document being written by an active user. À Propos performs group personalization by identifying the expertise of the group members in different topics. The user-level profile can be seen as an expertise fingerprint of that user, with terms that describe his or her interests. The user's own documents are an effective source for obtaining important

expertise terms (Balog et al., 2007). We can then use taxonomies, such as the ACM hierarchy, to represent the topics for which we want to quantify a group members' expertise. By collecting an adequate number of documents for each topic we can construct topic fingerprints.
The next step is to match these topic fingerprints with the user's expertise profile by calculating the term overlap. A higher overlap indicates more expertise in the subject. This way we can calculate the expertise of each group member on the different topic areas and also find out which group members are experts in the topic of the user's active document. Figure 3 shows an example of such a distribution.
Knowledge of the distribution of expertise over the group is then used for personalization. The recommendation of a document by an expert on the topic should be considered as more reliable and this can have a significant influence on the final re-ranking (Bogers and Van den Bosch, 2006). Group personalization can be used to recommend highly regarded documents that were not in the initial recommendation list. Finally, the expertise fingerprints can also be compared to each other and used to suggest related topics to provide for a more serendipitous experience. Serendipity is especially important in the earliest phases of planning and composing a document.

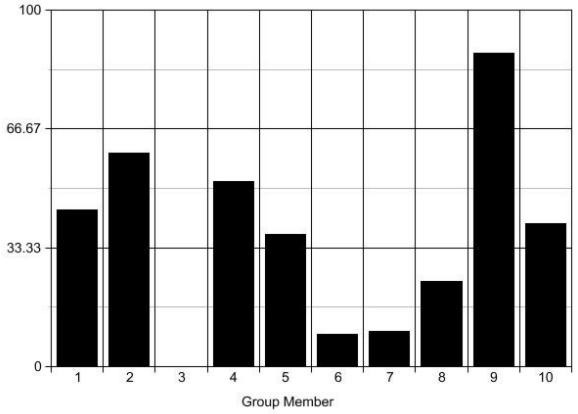

Figure 3: Distribution of expertise in a group. The vertical axis represents overlap between the fingerprints of individual users and experts.

4.2. The Current User's Interface

À Propos proactively submits queries based on the user profile, in combination with what the user is currently typing or reading. The retrieved information is presented proactively and immediately. In the present version of the system, search results are presented in a semi-transparent window located in the bottom right of the screen (see Fig. 4).
The window contains URLs related to what the user is typing. As the user moves the cursor over the references, the URLs become fully visible and active. On clicking the required URL, the user accesses the corresponding paper. The information presented also changes as the user moves the cursor while reviewing previously written parts of the document, on the basis of queries created from the text in the paragraph in which the cursor is located.

4.3. Writing Stages and No Intrusiveness

One problem with presenting proactive information is that it can interrupt the ongoing writing task. Interruptions can be more disturbing and distracting in specific stages of the writing process. This needs to be considered in order for the system to recognize what are the most opportune moments to present the information in a non-intrusive and timely fashion.
Furthermore, replicating other studies (Dansac and Alamargot, 1994; Jones, 2003) it has been found that writers look for extra information especially during Planning and Reviewing (Deshpande et al., 2006; Puerta Melguizo et al., 2007). We studied the effects of presenting proactive information during Planning and Reviewing.

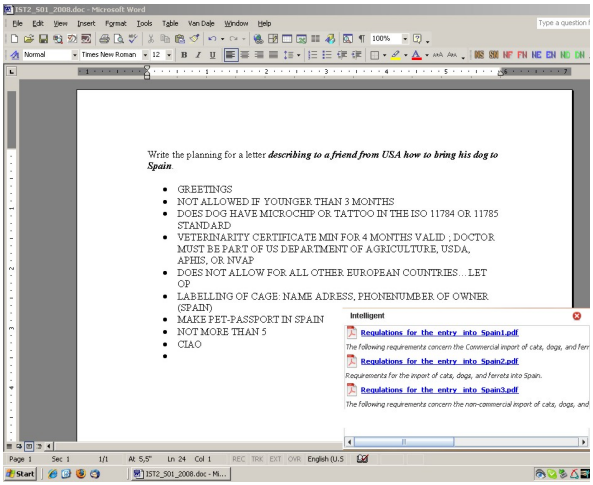

Figure 4: The user's interface.

4.3.1. Presenting Proactive Information during Planning Tasks

Planning involves creating and organizing ideas, and setting goals during composition. In order to simulate the stage of Planning, we used a procedure similar to the one described by Berninger et al. (1996). Participants were told that they had to write short essays about different topics, but before starting to create actual sentences they had to plan and write an outline of the major points of the text with supporting details and the order in which they would be introduced.
A total of 32 participants wrote the planning outlines under four information seeking conditions: 1) without PRS and no option of looking for extra information, 2) without PRS and with the option of getting information by actively searching information in the Web (active search), 3) with presentation of relevant information by our PRS, and 4) with presentation of non-relevant information by our PRS.
We found that the presentation of information by the PRS did not seriously impair time performance during planning. Furthermore, when relevant information was presented proactively, the quality of the resulting writing plan was significantly better, in the sense that participants introduced more information and included more correct ideas in their planning than in any of the other conditions. Concluding, the presentation of proactive relevant information by the PRS improves the quality of the planning process. At

the same time, the interrupts caused by presenting newly retrieved information does not disturb the task, since it does not take more time to write a plan in the condition where the PRS presented relevant information, compared to the other conditions. The results of this experiment also show that active search initiated by the user resulted in a lower quality of the information found and the written text.

4.3.2. Presenting Proactive Information during Reviewing Tasks

When reviewing previously composed text, writers read and edit their written text whenever errors or weaknesses are detected in the text. In this stage writers normally look for factual information in order to correct and/or justify written ideas. To explore the effects of presenting proactive information during the review phase, we asked participants to perform two different editing tasks: spelling corrections and filling in factual information that was left to be specified exactly when the text was drafted. These two tasks are similar to the ones described by Iqbal et al. (2005). Twenty participants performed both reviewing tasks under three different information seeking conditions: 1) without PRS but with the option of getting information by actively searching information in the Web (active search), 2) with presentation of proactive relevant information by our PRS 3) with presentation of non-relevant information by our PRS. Again, the main results of this experiment show that the presentation of proactive information does not seriously impair the time performance in editing tasks in comparison with the conditions in which the user was not interrupted by the system. Furthermore and very importantly, the time spent in looking for new relevant information was shorter when the PRS presented relevant information than in the cases in which participants had to search for the information actively. The information seeking time was even longer when non-relevant information was presented proactively. In this case, after assessing that the information by the PRS could not help in completing the editing task, participants started an active search. This result emphasizes the importance of developing appropriate search profiles and filters as described above. Finally, the quality of the editing tasks was also significantly better when proactive relevant information was presented showing once more, that active search initiated by a user is less effective.

5. Conclusions

Several systems have been developed to support the process of writing. Most of these systems, however, are mainly designed with the purpose of supporting the processes of reviewing and planning, organizing and connecting ideas. We think that it is also necessary to develop systems able to support author's knowledge about the topic. Or in other words, to develop systems that offer an external LTM memory to the writer and that can be accessed whenever the need to enter new information is needed. In this paper we presented the Proactive Recommender System: À propos. This system aims at supporting writers in the task of finding appropriate relevant information during writing.

First, we presented our approach at developing group and personal profiles that make sure the information presented by the system is relevant to the writer and to the specific piece of text is being written. We also describe the studies we performed in order to explore the effects of presenting proactive information when writers are planning and reviewing text. From our experiments we conclude that the user's interface of the PRS does not negatively affect the task of writing. And even more important, when relevant information is presented, the quality of the resulting text significantly improves in comparison with the situations in which the user actively seeks for information.

Futhermore, the results of our experiments with proactive presentation of information suggest that professionals are willing to accept unsolicited pop-up windows and similar interrupts if the information that they are alerted to by those interrupts is relevant for the completion of their (writing) task. Yet, more research is needed to better understand the human factors issues related to these interrupts.

5.1. An External Long-Term Memory

One of our goals is to develop the PRS in such a way that it can be used as an addition to the writer's neural Long-Term Memory (LTM). So far, virtually all writing research has been conducted in settings in which the LTM was limited to the writer's own brain (Berninger et al., 1996; Neuwirth and Kaufer, 1989). The advent of extremely powerful search systems already has had a large effect on the way people consider LTM. Students are no longer trained to memorize facts and information; rather, they are trained in efficient and effective search techniques. Thus, it is becoming more important to know how to find information than to memorize information in the first place. However, access to the virtually unlimited information in the Internet is not without problems. Knowing less, while searching more will make it more difficult to assess the importance of newly found information and to integrate it in a coherent framework. While computer tools may not be able to facilitate the integration process any time soon, they may be able to support the decisions about the relevance of the results returned from a query. The personalization and expert ranking that we are investigation holds the promise that it can help pofessionals to avoid getting lost in hyperspace and cyberspace. Also information retrieved in the form of documents or text snippets may have a different impact on how one decides to organize the information in a coherent text than when the information is retrieved from one's own experience. Consequently, we think it is neccesary to develop a new model of cognitive writing processes in which the external LTM that the WWW and other databases conforms, needs to be included as an important part of the physical environment. In this new model it would be also neccesary to include the need to look for information outside our brain during writing as another important cognitive process to consider. Currently there are a few models that try to represent writing and information searching process. The most relevant is the model of Kuhlthau (2004). In this model however, the writing starts as the search is finished. The real situation however is that both processes interact continuosly while writing and consequently a model that takes this issue into account is neccesary.

6. References

K. Balog, T. Bogers, L. Azzopardi, M. De Rijke, and A. Van den Bosch. 2007. Broad expertise retrieval for sparse data enviroments. In *Proceedings of the 30th Annual International Conference on Research and Development in Information Retrieval (SIGIR, 2007)*, pages 551–558, Amsterdam, The Netherlands. ACM Press.

V. Berninger, D. Whitaker, H.L. Yuen Feng Swanson, and R.D. Abbott. 1996. Assessment of planning, translating, and revising in junior high writers. *Journal of School Psychology*, pages 23–52.

T. Bogers and A. Van den Bosch. 2006. Authoritative reranking of search results. In *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, volume 3936, pages 519–522.

B.K. Britton and S.M. Glynn. 1989. *Computer Writing Environments: Theory Research and Design*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

J. Budzik and K. Hammond. 1999. Watson: Anticipating and contextualizing information needs. In *Proceeding of the 62th Annual Meeting of the American Society for Information Science*, pages 727–740.

J. Conklin. 1987. Hypertext: An introduction and survey. In *IEEE Computer Magazine*, volume 20, pages 17–41.

C. Dansac and D. Alamargot, 1994. *Accessing referential information during text composition: when and why?*, pages 76–97. Amsterdam University Press.

A. Deshpande, L. Boves, and M.C. Puerta Melguizo. 2006. À propos: Pro-active personalization for professional document writing. In *SigWriting, 10th International Conference of the EARLI Special Interest Group on writing*.

S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. 2003. Stuff ive seen: a system for personal information retrieval and re-use. In *Proceeding of the 26th Annual Int. Conference on Research and Development in Information Retrieval (SIGIR 2003)*, New York. ACM Press.

J. Gwizdka and I. Spence. 2007. Implicit measures of lostness and success in web navigation. *Interacting with Computers*, pages 357–369.

C. Haas. 1996. *Writing Technology Studies on the Materiality of literacy*. Lawrence Erlbaum Associates, Hillsdale New Jersey.

J.R. Hayes and L. S. Flower, 1980. *Cognitive processes in writing*, chapter Identifying the organization of writing processes, pages 3–30. Lawrence Erlbaum Associates, Hillsdale New Jersey.

J.R. Hayes, 1996. *The science of writing: Theories, methods, individual differences and applications*, chapter A new framework for understanding cognition and affect in writing, pages 76–97. Lawrence Erlbaum Associates, Hillsdale New Jersey.

S.T. Iqbal, P. D. Adamczyk, X. S. Zheng, and B. P. Bailey. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of ACM Conference on Human Factors in Computing Systems (ACM CHI 2005)*, pages 311–320.

P.H. Jones. 2003. Distributed information seeking in research collaboration: An extended economy of resources, memory and cognition. In *CSAPC 2003 Workshop*, Amsterdam.

R.T. Kellogg, 1996. *The science of writing*, chapter A model of working memory in writing, pages 57–71. Lawrence Erlbaum Associates, Mahwah, New Jersey.

C.C. Kuhlthau. 2004. *Seeking meaning: a process approach to library and information services*. Libraries Unlimited, Westport CT.

C.M. Neuwirth and D.S. Kaufer. 1989. The role of external representations in the writing process: Implications for the design of hypertext-based writing tools. In *Hypertext 1989*, pages 319–341. ACM Press.

M.C. Puerta Melguizo, V.R. Lemmert, and H. Van Oostendorp, 2006. *Current Research in Information Sciences and Technologies: multidisciplinary approaches to global information systems*, volume 1, chapter Lostness, Mental Models and Performance, pages 256–260.

M.C. Puerta Melguizo, L. Boves, A. Deshpande, and O. Muñoz Ramos. 2007. A proactive recommendation system for writing: Helping without disrupting. In W-P. Brinkman, D-H. Ham, and Wong W., editors, *ECCE 2007: European Conference on Cognitive Ergonomics*, pages 89–95.

B. J. Rhodes. 2000. *Just-in-time Information Retrieval*. Phd thesis, Massachusetts Institute of Technology, Massachusetts, USA.

J.B. Smith, S.F. Weiss, and G.J. Ferguson. 1987. A hypertext writing environment and its cognitive basis. In *Hypertext 1987*, pages 195–214.