

WHAT A PROACTIVE RECOMMENDATION SYSTEM NEEDS: *Relevance, Non-Intrusiveness, and a New Long-Term Memory*

Keywords: Proactive recommendation systems, information seeking, writing stages.

Abstract: The goal of the project *À Propos* is to develop a proactive, just-in-time recommendation system for professional writers. While authors are writing, the proactive system searches for relevant information to what is being written, and presents this information to the writers in a manner that is perceived as timely, non-intrusive, and trustworthy. In this paper we present our ideas and the first steps performed in order to reach this goal. Writing a professional document is a complex and highly demanding task that can be seriously affected by interruptions from the environment. Consequently, a proactive system should be 1) able to present highly relevant information consequently, 2) able to identify in what stage of writing the author is involved, and what are the moments in which information needs are more important and less disruptive, and 3) serve as an external long-term memory for the writer. In this paper we describe the steps and first results of our project *À Propos* in order to develop a proactive recommendation system that covers these goals.

1. INTRODUCTION

Behind the process of writing professional documents lies a steady but intermittent need to check, validate, and add information. Search engines have become the primary tool for information access in both company-internal networks and the Internet. Still, broad keyword-based search is inefficient. Considerable time is spent interacting with low-precision search engines. The time in which the author is away or distracted from creating the document can have a negative impact on the time spent, and on the quality of the text. In addition, relevant information may be missed because the writer did not realize that the information exists and could be looked up. Furthermore, 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.

Proactive Recommendation Systems (PRSs) retrieve large quantities of documents, decide what available information is most likely relevant to the text to be written, 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. Only a few PRSs have been specifically developed to support writing. For example, the Remembrance Agent (Rhodes, 2000) suggests personal email and documents based on text being written or read. Watson (Budzik and Hammond, 1999) is another PRS (or IMA: Information Management Assistant as the authors called it) that performs automatic Web searches based on text being written or read. Finally, IntelliGent™ is a PRS currently being developed at the Search Expertise Centrum in The Netherlands. The system proactively submits queries to a potentially large number of search engines and presents the retrieved information while the user is writing a document. However, a serious problem with all of these PRSs is that they are developed as search support tools and, do not seem to take into account the specific characteristics of the task at hand. Writing professional documents is a complex and highly demanding task that can be seriously affected by any type of interruption from the environment.

To understand the factors influencing the process of writing, we briefly describe the cognitive model of writing proposed by Hayes and Flower (1980). We then summarize the results of an initial study

conducted with IntelliGent™ (Deshpande et al, 2006). The purpose of the study was to get a general impression of how users perceive and evaluate PIRs supporting writing. Based on the results of that study, we describe the steps of our project *À Propos* in order to develop a PIR able to present highly relevant information, just-in-time and in a non-disruptive manner.

2. THE STAGES OF WRITING

Our interest in studying the writing process derives from the fact that we think that the need for a PIR might differ from one moment to another during the process of writing a document. To understand the factors that play a role in writing, we adopted the model proposed by Hayes and Flower (1980). Although the original model has been extended (e.g. by adding the Working Memory; cf. Olive, 2004), the original model is fully adequate for our purposes. According to the model, the writing process happens in three stages: Planning, Translating, and Reviewing (see Figure 1).

Planning involves retrieving and selecting information from the Long-Term Memory (LTM) and the Task Environment. Planning is divided into three sub-processes. *Generating* involves retrieving domain knowledge from the LTM. *Organization* implies selecting the most useful material retrieved by the generating process, organizing it into plans and determining the sequence in which these topics will be presented. *Goal setting* involves the elaboration of criteria that will allow the writer judging the appropriateness of the written text relative to the writing intentions. Planning precedes the formal writing or translation and continues occurring during the entire writing process in the background.

During the Translating stage information is taken from the LTM in accordance to the writer's plans and goals and is formulated 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 would best serve the writing goal. Reviewing involves two sub-processes. *Reading* allows to detect errors or weaknesses and to evaluate the appropriateness of the written text in relation to the goals established during planning. *Editing* appears as a system of production rules that result in changes to the text.

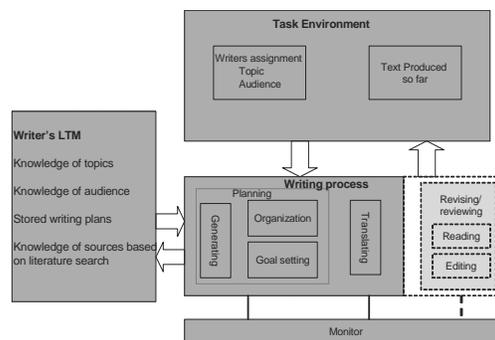


Figure 1: Cognitive process model of writing (Flower and Hayes 1980).

Writing is a recursive and non-linear process and all stages can occur rapidly, simultaneously and automatically. For instance, the sub-processes of generating and editing might interrupt other processes, and planning might occur as part of editing (Glynn and Britton, 1989).

The structure of the writing process, and especially the way in which that structure affects writer's behaviour, can have a large impact on the ways in which PRSs should interact with the user. To explore in which of the writing stages authors are most in need of additional information, Deshpande et al. (2006) performed an exploratory study with the proactive system IntelliGent™.

3. WRITING WITH AN INTELLIGENT™ PROACTIVE RETRIEVAL SYSTEM

The study conducted by Deshpande et al. (2006) had the goals of understanding how and when scientists use IntelliGent™ in a natural working environment and to what extent the PRS is supporting them during writing.

IntelliGent™ proactively submits queries based on a broadly defined user profile in combination with what the user is currently typing. The system presents the retrieved information to the user proactively and immediately. The results of the search are presented in a semi-transparent window located in the bottom right of the screen (see Figure 2). 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 from the digital library. The information in the window is changed depending upon the text that is being input and new queries are created. The information presented also changes as the user moves the cursor while reviewing previously written parts of the document.

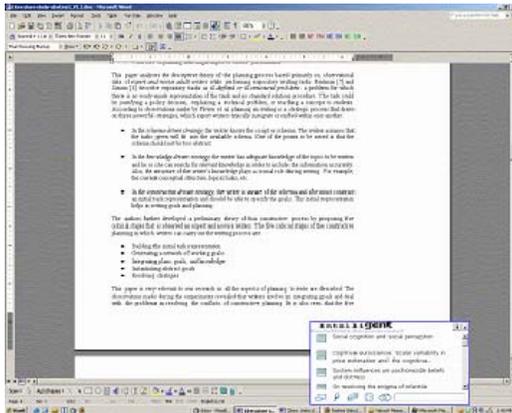


Figure 2: IntelliGen™ System for proactive information retrieval

During two months researchers from the department of Language and Speech Technology at Radboud University (The Netherlands) used IntelliGen™ whenever they were using MS-Word. The Scopus® database was linked to IntelliGen™ as the source for information. To investigate if there were different information seeking needs during the different stages of writing, several open and structured interviews and questionnaires were conducted. Also the issues of efficiency, effectiveness, and overall satisfaction with IntelliGen™ were addressed.

The main results of the study show that the system was not really efficient and did not improve participants' productivity. According to the participants, the main reason was that the system frequently presented irrelevant information, resulting in disruptions of the writing process. We concluded that better filtering techniques are necessary to improve the selection of relevant information, and avoid annoyance and frustration among writers. Actually, participants recognized that the use of the system would add value to their writing tasks if it were able to present really relevant information when needed.

We also found that, as the user shifts between different stages in the writing process, information requirements differ. In most cases, participants were not aware of doing any planning before starting the translating stage. However, in all cases, they found that the most important moment to search for information is before the translating stage starts. When asked if they also looked for new information during the stages of translating and reviewing, participants claimed not to do it frequently and only when some justification of their ideas while writing

was needed. These results are similar to the ones described by Dansac and Alamargot (1999).

In conclusion, we found that a PRS to support professional writing will only be appreciated 1) if the recommended information is highly relevant and 2) is offered in the right stage of the writing process. From these results the main goals of our project *À Propos* were developed.

4. *À PROPOS*: A PROACTIVE, PERSONALIZED, JUST-IN-TIME RECOMMENDATION SYSTEM

The goal of the project *À Propos* is to develop a proactive, adaptive, personalized, and just-in-time knowledge management environment for writers in a professional environment. The resulting environment is depicted in Figure 3.

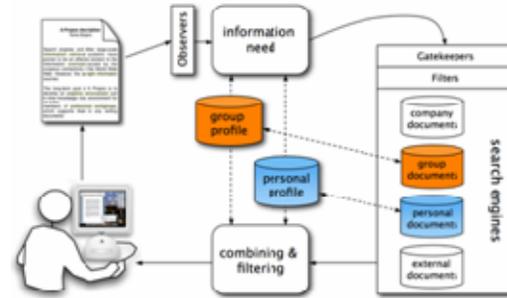


Figure 3: The *À Propos* architecture

The architecture of *À Propos* is inspired by IntelliGen and other PIRs such as Watson (Budzik and Hammond, 1999) and Stuff I've Seen (Dumais et al, 2003). *À Propos* is based on a client-server architecture. The client runs on the user's computer and monitors user's activity constantly. The server handles all the incoming requests for information, consults the relevant information sources, performs load balancing, and returns the search results to the *À Propos* client.

According to our initial study, two main issues need to be carefully investigated. First, in order to present highly relevant information, appropriate filtering techniques need to be developed. Second, procedures to identify the different writing stages and related information needs must be created. Our plans for these equally important sides of *À Propos* are discussed in the next subsections.

4.1 Relevance

The acceptance of any PRS hinges on the relevance of the suggested information: users will simply stop using the agent if they are presented with irrelevant information too often. Therefore, À Propos aims to develop methods for generating *search profiles* that enable effective, trustworthy, and high-precision information retrieval with regard to the user's current information need.

Search profiles are generated on the basis of a collection of documents previously written by the user and the workgroup. The profiles must also be able to adapt to the specific information needs of the user while writing a specific document. The À Propos agent integrates these search profiles with a parallel interface to public domain and proprietary internal search engines, as well as the user's own pool of documents. The final step is fusing and filtering the search results from all the different sources. The next sections go into more detail about the different steps in the architecture.

4.1.1 Information Need

The first job of the À Propos agent starts whenever a user is writing or reading a document. Determining the information need is aided by the *Observers*, software agents that monitor user's activity in different applications such as Internet browsers, word processors and email clients. The Free Search Observer monitors explicit searches for information in the À Propos search window.

Observers collect the paragraph the user is currently writing or the text displayed on screen if the user is reading a document. In a later stage the À Propos agent uses this context to estimate the user's information need by formulating appropriate queries. Query formulation is dependent on the search engine used, so it is not performed at this stage.

We plan to extend the query component by experimenting with different context sizes for the Observers and considering different context extraction constraints for each type of Observer. Currently, there is almost no difference between context extractions for different Observers, which is suboptimal. In addition, the frequency with which context is extracted and subsequent queries are launched, should be tuned to the stage in the writing process. Another factor we wish to investigate is the probable beneficial influence of using the personal and group search profiles in the extraction stage.

4.1.2 Search

Searching for documents relevant to the user's information need is the second step in À Propos. The context extracted in the previous stage is sent to the

À Propos server and distributed to the search engines in the form of specific queries. Distilling an appropriate query from the user's context is done by spotting key terms and phrases in the context using domain-specific taxonomies and heuristics for ranking query terms. The extracted terms and phrases are then used to generate queries, enabling À Propos to perform searches without user-formulated queries. This approach is similar to the query-free approaches to news search (Henzinger et al, 2003), and expert diagnostics (Hart and Graham, 1997).

To prevent overloading the search engines with too many, irrelevant, and/or redundant search requests, each information source has separate Gatekeepers and Filters.

Gatekeepers generate queries for their respective information sources and determine whether or not to execute the queries suggested by the Observers. For each information source, the Gatekeepers use a domain-specific taxonomy of which a certain number of query terms need to match for the query to be executed. In addition, the Gatekeepers compare new queries to queries that were submitted recently to prevent redundancy.

Filters guard the relevance of the documents and filter out data that is not relevant for inclusion in a query. Among others, function words that have little lexical meaning such as 'and', 'the', 'of', etc. are suppressed. The filters also transform a query into the format required for the different information sources.

The *information sources* can be divided into four different types. External documents are typical electronic document stores such as Scopus®, ACM, Springer, etc., but also the Internet (Google Scholar, CiteSeer). Company-internal document stores include Intranet databases and other in-company information resources such as patent databases and technical reports. The group documents cover all the work done by the workgroup the user is a part of. Finally, the user's personal document collection, consisting of self-authored documents and other downloaded papers, is another information source.

We plan to extend this component by using personal and group profiles to enhance the performance of the Filters and the Gatekeepers. These profiles could also be used for expanding queries that are submitted to public search engines.

4.1.3 Combining and Filtering

After a query is submitted to different search engines, the different sets of search results need to be combined and filtered to present a single list of recommendations to the user. The first step in this third stage of À Propos is *deduplication*, a well-researched problem in distributed information

retrieval (e.g. Callan et al, 1995). Bibliographic screening techniques similar to those developed in the CiteSeer project (Giles et al, 1998) are used for deduplication in À Propos.

The next step is filtering the results—essential in guarding the user against the system recommending irrelevant information. Depending on the stage of the writing process, documents that the user has seen before may have to be filtered out. In the end, only documents whose ranking scores exceed a strict relevancy threshold are recommended to the user.

We plan to test the influence of personal and group profiles on this filtering step. Filtering and re-ranking the list of results depends on these search profiles and on other characteristics of user's workgroup. The search profiles could also be used to re-rank the results, giving preference to documents that match better with the user's personal profile.

4.1.4 Personalization

The relevance of suggested documents is strongly affected by the topic of the user's current text and by the user's research interests. Personalization is handled in À Propos by generating and applying search profiles that are generated on the basis of a collection of documents previously written by the workgroup. À Propos distinguishes between individual user profiles and the workgroup profile.

A *user profile* is created by using a combination of questionnaires and important terms extracted from the documents written by the user. Another source for profile information is supplying À Propos with a list of in-company and public information resources that should be consulted. In either case, initially, the À Propos agent will give more weight to terms and factoids used in documents authored by the user and the questionnaire answers. Search profiles are updated as the user works on new documents or whenever the user provides positive feedback on recommended documents. These search profiles are used to guide the retrieval and recommendation process for the user as terms and phrases in the user profile can be used to expand queries and to filter or re-rank search results.

The *group profiles* serve the same purpose and can also contain information about group dynamics such as trustworthiness or expertise. For instance, documents recommended by experts on the active document topic should receive a higher weight than documents recommended by laymen.

Future work on personalization includes comparing different methods of constructing and combining the search profiles and investigating their influence on the relevancy of the recommendations made by À Propos.

4.2 Writing Stages and Non-intrusiveness

The structure of the writing process has a large impact on the ways in which a PRS should interact with the user.

On the one hand, it seems that there is a strong need for searching information during the planning stage, but often most of the planning occurs before any substantial writing is done. Although a writing strategy that involves explicit planning is often recommended in formal writing courses, few scientists seem to do it. The challenge to designers is then to make the PRS so effective and powerful that professional writers experience the added value of adhering to a strategy that involves explicit planning. In other words, the PRS should be able to motivate users to change their writing procedures in such a manner that the system can help them to find information in the appropriate moment. For example, if users would make their writing plans explicit by typing section headers and short summaries of what should go into each section before they set out to create the full text, a PRS might be in a much better position to search for potentially relevant information than if users start typing the full text of the first paragraph, and continue in this manner from one paragraph to the next. The big benefit to writers would be that they receive recommendations in a proactive manner, shortening considerably their task of seeking for information, and minimizing the risk of missing essential information.

On the other hand, proactive information recommendation does interrupt the ongoing task, and it may well be that these interruptions are more disturbing and distracting in specific stages of the writing process. Consequently, the possible different effects of interruptions during different writing stages need 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. We are currently conducting HCI experiments to investigate the effects of interrupts in the Planning and Reviewing Phase.

In addition, we are investigating several issues related to the interface of the PRS. Here, the goal is to design the interface and interaction procedure in such a way that it is easy for writers to observe that potentially relevant information has been retrieved, while at the same time it is easy to ignore the messages of the system if they are involved in a part of a task that would be difficult to resume after having been interrupted by À Propos.

4.3 A New Long-Term Memory

Another goal of À Propos is to develop the PIRs in such a way that it can be used as an addition to the writer's neural LTM. So far, virtually all writing research has been conducted in settings in which the LTM from which participants could 'get information' was limited to their own brain (e.g. Olive, 2004). The advent of extremely powerful search systems will have a large effect on the way people will consider and use LTM. In the future it may be more important to know how to find information than to memorize information in the first place. 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.

5. CONCLUSIONS

Current proactive recommendation systems do not take into account the various writing stages and different information searching needs in their design. The goal of the project À Propos is to develop a proactive, just-in-time recommendation system for professional writers that does take these issues into account. The idea is that while authors are writing, the proactive system searches for relevant information to what is being written, and presents this information to the writers in a manner that is perceived as timely, non-intrusive, and trustworthy. In this paper we present our ideas and the first steps performed in order to reach this goal.

ACKNOWLEDGEMENTS

We would like to thank Frank Hofstede of Search Expertise Centrum, Utrecht, The Netherlands for providing us with IntelliGent™. We would also like to thank IOP-MMI, administered by SenterNovem, for funding this project.

REFERENCES

Budzik, J. and Hammond, K., 1999. Watson: Anticipating and Contextualizing Information Needs. *Proc. 62nd Ann. Meeting Am. Soc. for Information Science*, 727-740.

Callan, J.P., Lu, Z. and Croft, W.B. 1995. Searching Distributed Collections with Inference Networks. In *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research*

and Development in Information Retrieval, pages 21-28, New York, NY, ACM Press.

Dansac, C. and Alamargot, D., 1999. *Accessing referential information during text composition: when and why?* In M. Torrance and D. Galbraith (Eds.). *Knowing what to write: Conceptual processes in text production*, pp.76-97. Amsterdam University Press.

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

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R. and Robbins, D. C. 2003. Stuff I've Seen: a System for Personal Information Retrieval and Re-use. In *SIGIR'03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 72-79, New York, NY, ACM Press.

Giles, C. Bollacker, K and Lawrence, S. 1998. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of Digital Libraries 98*, pages 89-98.

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

Hart, P.E. and Graham, J. 1997. Query-free Information Retrieval. *IEEE Intelligent Systems and Their Applications*, 12(5):32-37.

Hayes, L.S. and Flower, J.R., 1980. Identifying the organization of writing processes. In L.W. Gregg, and E.R. Steinberg. (Eds.) *Cognitive Processes in Writing*, Hillsdale, NJ, Lawrence Erlbaum, 3-30.

Henzinger, M., Chang, B., Milch, B. and Brin, S., 2003. Query-free News Search. In *Proceedings of the Twelfth International Conference on World Wide Web (WWW '03)*, pages 1-10, Budapest, Hungary, 2003.

Olive, T., 2004. Working memory in writing: Empirical evidence from the dual-task technique. *European Psychologist*, 9, 32-4.

Rhodes, B.J., 2000. *Just-in-time Information Retrieval*, Phd Thesis, MIT.