

# À Propos: Pro-Active Personalization for Professional Document Writing

Toine Bogers

ILK / Language and Information Science  
Tilburg University, P.O. Box 90153  
NL-5000 LE Tilburg, The Netherlands  
Tel: +31 13 466 2451  
Fax: +31 13 466 2892  
A.M.Bogers@uvt.nl

## 1. INTRODUCTION

Information Management Assistants (IMAs) are software agents that “automatically discover related material on behalf of the user by serving as an intelligent intermediary between the user and information retrieval systems” [3]. IMAs monitor the user’s interaction with everyday applications and then infer and anticipate the user’s information need from the context. They then try to fulfill these needs by accessing the traditional search engines and retrieval systems for the user, fusing the results, and presenting them to the users. Some examples of IMAs are Watson [3], Syskill & Webert [6], Letizia [5], and the Remembrance Agent [7].

The majority of these systems, however, do not take into account the interests and characteristics of different users or groups of users, which means the results presented are independent of the user or the user’s workgroup. In this application for the IliX Doctoral Forum we describe ongoing and proposed research in the À Propos project<sup>1</sup>, which aims to develop a personalized, pro-active IMA that takes into account a user’s preferences and interests, as well as the dynamics of the workgroup or community the user belongs to.

## 2. À PROPOS

The goal of the À Propos project is to develop an IMA that supports users in the daily process of writing professional documents, such as scientific articles and technical or business reports. It aims to reduce the time users spend searching for information and analyzing search results by pro-actively searching for relevant, personalized, and trustworthy information. Furthermore, information should be presented to the user in a non-intrusive and timely fashion so as to minimize disturbing the user’s writing process. In this paper, however, we will only focus on the information retrieval part of the project.

À Propos aims to develop methods for generating *search profiles*

<sup>1</sup>See also <http://ilk.uvt.nl/~toine/apropos>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.  
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

that enable effective, trustworthy, and high-precision information retrieval with regard to the user’s current information need. This information need is influenced not only by the document the user is writing, but also by his or her context: personal characteristics and those of the workgroup the user belongs to.

Search profiles are generated on the basis of a collection of documents previously written by the user and his or her workgroup. The profiles must also be able to adapt to the information needs of individual users and of their workgroup. 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. Figure 1 shows the different stages of the À Propos agent.

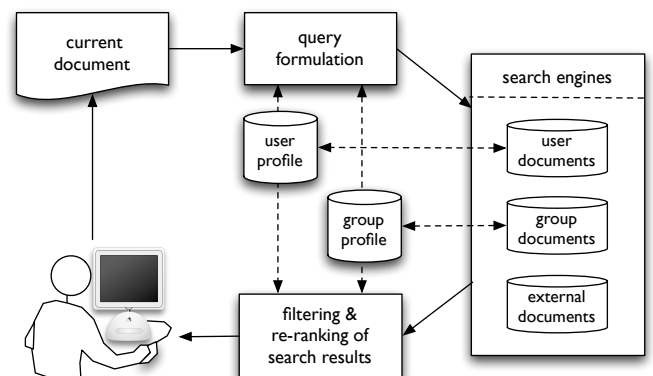


Figure 1: The À Propos workflow

### 2.1 Query formulation

An essential step in locating new and relevant information is understanding the contents of the text, to enable À Propos to match that content against all the knowledge and information sources available to it. The À Propos agent analyzes sentences on the level of terms and factoids, which are then used to formulate the appropriate queries for finding relevant information. During the writing process the active document and the user’s information need can change considerably, which means that queries derived from it and the relevant documents also change. The following research questions are relevant in the query formulation stage:

- How do we formulate a query to find relevant documents?
- How much context should be used in formulating a query?

## 2.2 Fusion & filtering

After the query has been submitted to the 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. Filtering and re-ranking the list of results also depends on the search profiles and on other characteristics of user's workgroup. The big research question here is how much influence search profiles should have on filtering and re-ranking search results?

## 2.3 Search profiles

Search profiles are generated on the basis of a collection of documents previously written by the users of a workgroup. The À Propos agent 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. The group profile can be a generalized combination of all the member profiles. Another option is to first extract the group profile and initialize all individual profiles with it. 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 when the user provides positive feedback on recommended documents.

These search profiles are used to guide the retrieval and recommendation process for the user: terms and phrases in the user profile can be used to expand formulated queries and to filter or re-rank search results. The group profiles serve the same purpose and can, additionally, 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. Relevant research questions about the search profiles are:

- What is the best way to construct a search profile?
- Which type of profile should be derived from the other type? Should group profiles serve as the basis for personal profiles or should personal profiles be combined to form a group profile?
- Which source is more important for effective personalization: positive feedback from the user on recommended documents or the user's own authored documents?

## 3. PAST WORK

Past work in the À Propos project focused on combining re-ranking with search profiles. We developed a method for constructing search profiles that model the expertise of workgroup members based on the documents each member has authored. These individual user profiles contain expertise-related terms for each workgroup member and are then used to rank all workgroup members on their expertise on a topic. In addition, we successfully used those expertise rankings to improve the performance of a baseline IR system with statistically significant gains ranging from 1.5% to over 34%. Our approach is described in greater detail in [1] and [2].

We also created a special workgroup IR test collection to better test and evaluate the performance of the À Propos agent. This ILK collection<sup>2</sup> contains 147 document titles and abstracts of publica-

<sup>2</sup>Publicly available via <http://ilk.uvt.nl/~toine/apropos/>. See [1] for more information.

tions of current and former members of the ILK workgroup and 80 natural language queries.

## 4. FUTURE WORK

The effectiveness of the À Propos agent will be thoroughly investigated in actual office environments, as well as the usability and the exact role an IMA such as À Propos plays in the writing process [4]. Before these tests can take place, however, we need to experimentally validate our ideas and solutions and we propose the following setup.

We wish to construct a specialized version of the ILK collection where we select one or two papers of each current member of the ILK workgroup. We ask each main author to identify for each paragraph in the document which of the references were used to write that passage, i.e. which were relevant to the writing process. This way we hope to simulate which documents our IMA would have had to recommend, had the agent been monitoring the active document's writing process.

We realize that usually publications are referenced directly in the text itself where they were used, but this does not always have to be the case. The ideas in a reference might have been used in more than one section with the author simply referencing the paper only once. Furthermore, not all references need to be of equal importance in writing the document in question. In addition to this, we also wish the annotators to signal which ILK publications that were not referenced might have been relevant for which document passages.

Using these passages to construct queries determines, at the very least, which documents should be returned when writing that section. This is close to the ideal À Propos situation. We realize that using a collection such as this does not guarantee that all unseen, relevant documents would also be returned. However, it does provide us with a decent performance estimate. Using this collection we wish to explore, among other things, the following:

- The effect of context size in query formulation: does using more text as source material for query construction lead to better results?
- Different forms of formulating queries: simply selecting terms based on TF-IDF, using a thesaurus to guide term selection, or using the terms and factoids in the search profiles.
- What is more effective: normal query expansion vs. expansion based on the search profile terms?
- How search profiles should be used to re-rank search results.

## 5. MOTIVATION

My motivation for participating in the IiX Doctoral Forum lies in both the focus of the IiX conference and that of our ILK workgroup. The IiX focus on information interaction and retrieval in context is a direct match with the scope of the À Propos project which encompasses many of the symposium's themes.

In contrast, the ILK workgroup focus is more on NLP and machine learning. This means that feedback on specific IR-related issues is more difficult to obtain in my workgroup. This makes IiX an excellent opportunity for feedback, especially on the setup we proposed in section 4 to investigate our research questions.

## 6. REFERENCES

- [1] T. Bogers and A. van den Bosch. Authoritative Re-ranking in Fusing Authorship-based Subcollection Search Results. In

- F. de Jong and W. Kraaij, editors, *Proceedings of the Sixth Belgian-Dutch Information Retrieval Workshop, DIR-2006*, pages 49–55, Enschede, March 2006. Neslia Paniculata.
- [2] T. Bogers and A. van den Bosch. Authoritative Re-ranking of Search Results. In *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, vol. 3936 of *Lecture Notes on Computer Science*, pages 519–522, Berlin, April 2006. Springer Verlag.
- [3] J. Budzik and K. Hammond. Watson: Anticipating and Contextualizing Information Needs. In *62nd Annual Meeting of the American Society for Information Science*, Medford, NJ, 1999.
- [4] A. Deshpande, L. Boves, and M. C. P. Melguizo. À Propos: Pro-active Personalization for Professional Document Writing. In *To appear in SIGWriting '06: Proceedings of the 10th International Conference of the EARLI Special Interest Group on Writing*, September 2006.
- [5] H. Lieberman. Letizia: An Agent That Assists Web Browsing. In C. S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 924–929, San Mateo, CA, 1995. Morgan Kaufmann publishers Inc.
- [6] M. J. Pazzani, J. Muramatsu, and D. Billsus. Syskill & Webert: Identifying Interesting Web Sites. In *AAAI/IAAI, Vol. 1*, pages 54–61, 1996.
- [7] B. J. Rhodes. *Just-In-Time Information Retrieval*. PhD thesis, MIT Media Laboratory, Cambridge, MA, May 2000.