Personalization techniques in the TIPS Project: The Cognitive Filtering Module and the Information Retrieval Assistant

Stefano Mizzaro and Carlo Tasso

Artificial Intelligence Laboratory
Department of Mathematics and Computer Science
University of Udine
{mizzaro, tasso}@dimi.uniud.it

Abstract. Persistent and ephemeral personalization techniques can be exploited to implement more adaptive and effective information access systems in electronic publishing. Within the TIPS project, we effectively applied both these techniques in information filtering and retrieval systems used, via the specialized Torii portal, by physicists in their daily job.

1. Introduction

Internet has changed, and is changing, the standard communication mechanism adopted in science. Nowadays, a peer reviewed journal can be distributed by electronic means, and the peer reviewing can take place completely electronically, drastically reducing time and money for publishing (see, e.g., JHEP at jhep.sissa.it or Earth Interactions at EarthInteractions.org). Many publishers now allow their subscribers to electronically access the full text of the papers published on standard journals. Beyond modifying the standard scholarly journals and proceedings, the Web has also introduced a new way of disseminating scholarly knowledge: *e-prints*, i.e., open online repositories of scholarly papers (see, e.g., arXiv.org, mainly about physics, or cogprints.soton.ac.uk, about disciplines concerning cognition).

As a result, the scholar is nowadays overloaded by a large amount of highly structured hypermedia information, in the form of scholarly publications, online repositories, commentaries, and so on. In this scenario, it is important to allow the scholar: (i) to stay up-to-date, being notified when new information on some topics of interest is published, and (ii) to quickly and easily find, on demand, information on specific topics. Both goals can be approached by advanced personalization techniques. At the University of Udine, we have been investigating the issue of personalization in information access for several years [1, 7, 8, 9, 17, 19]. Within the 5th FP IST project TIPS (Tools for Innovative Publishing in Science, contract number IST-1999-10419), see tips.sissa.it, we applied adaptive and personalized information access techniques to the electronic publishing field, and more specifically in scholarly publishing.

Personalization is needed and useful in information access, and especially in scholarly publishing. Users (i.e., researchers) are interested in it for two important reasons: (i) detecting newly published information relevant to their interests and

preferences, and (ii) accessing stored information satisfying specific information needs. However, this twofold situation requires a novel approach, in which two distinct and complementary personalization techniques (i.e., ephemeral and persistent personalization) can be applied together to meet user's requirements. Personalization plays indeed a fundamental role not only for the highly subjective nature of the information seeking process, but also because the job of a researcher is highly innovative, it does not conform to any standard behavior, and it is therefore quite different for each researcher.

Personalization techniques are very numerous and are ranging from simple user-controlled customization of Web content, to autonomous system-controlled adaptation [14, Reader's Guide, p.6]. We distinguish two types of personalization [20]: *persistent* (or *long term*), i.e., based on a user profile which lasts over time and is stored in a persistent information structure; and *ephemeral* (or *short term*), which is not based on a persistent user profile. The main differences are the temporal features of the process aimed at building and managing the user profile: in persistent personalization, the user profile is incrementally developed over time and at the end of each session it is stored in order to be used later on in subsequent sessions. These two personalization techniques (ephemeral and persistent) nicely match with the two classical kinds of information access [6, 10]: information retrieval (IR) [2] and information filtering (IF) [12].

On the one side, personalization in IF means capturing the long term information interests and preferences of the user, in order to tailor the selection process to the specific personal characteristics. On the other side, persistent personalization is not feasible in IR, since in that context information needs have a short term nature and are different, for the same user, in the different sessions. However, ephemeral personalization can be effectively exploited, with the goal of modeling the search session, rather than the information need, for immediately providing personalized support during the searching session. The resulting approach to the personalization in IR systems is innovative for two reasons: (i) a short term modeling is performed through ephemeral personalization, which restricts the scope of observation to the current session only, and (ii) we do not build a model of the information need (difficult, if not impossible, during just one session), but rather a session model.

In the remaining part of the paper, we briefly describe the two Torii modules in which we have applied this approach. A longer description is available in [18].

2. The Cognitive Filtering Module

In previous work, we have developed and evaluated several content-based filters [15] for persistent personalization. Among them [1, 17, 19], the most effective has been the information agent *ifT* (*information filtering Tool*) [17], which is based on the user modeling shell *UMT* (*User Modeling Tool*) [9].

ifT exploits lightweight natural language processing and co-occurence-based semantic networks for building long term user profiles and for evaluating the relevance of text documents with respect to a profile. The main mechanism for building user profiles exploits explicit relevance feedback provided by the user on both positive and negative examples. The learning capabilities of this mechanism have been evaluated by means of several laboratory experiments [1].

Given the performance reached by ifT, we decided to adopt it as the filtering engine of the Torii portal. More specifically, the problem approached with persistent personalization has been the high (and currently increasing) rate of incoming documents: about 100-200 new e-prints are submitted every day and included in arXiv, which is accessible through Torii. Normal users (researchers in high energy physics) were used to start the working day by browsing the long list of new e-prints. By adding a personalized filtering engine to Torii, each user can now define one or more profiles related to his interests, and all the new incoming information is automatically filtered. In this way, Torii displays (in the first positions) only the documents which best match user's interests. Information overload is then reduced, as well as the cognitive load of analyzing many documents every day.

Torii has undergone a validation phase through field testing in July 2001. Twenty users were using the system for 29 days. All their sessions have been monitored and tracking logs of all actions acquired. Final interviews were also delivered. Cognitive filtering was working well and judged well by the users, who proposed to extend the system with the possibility to rank any set of documents (possibly coming as the result of a search in one of the available collections) by means of ifT.

3. The Information Retrieval Assistant

The interactive nature of IR is advocated since years [13] and is now widely accepted: between the user and the IR system a dialogue takes place [5], during which the user should receive adequate support [3]. The help should be provided proactively by the system and suggestions should be given "on the background", with the user retaining the control of the interaction [4]. A basic kind of support is *terminological* help, which identifies and suggests to the user terms that improve the query [11, 16]. Another kind of support is *strategic* help, which provides to the user useful hints on how to improve the strategy adopted for organizing the searching process (see a survey of this issue in [8]).

We use ephemeral personalization techniques to provide both strategic and terminological support to IR users. We have been doing research on this issue for several years. We implemented the FIRE and SAM prototypes [7, 8] that, by means of thesauri and of a detailed conceptual model of the session, are capable of suggesting to the users of a boolean IR system alternative terms and strategies to better (re)formulate their information needs. After some laboratory experiments involving several participants, we had evidence that terminological and strategic help are useful and nicely complement each other [7]. Following this positive evaluation, we applied ephemeral personalization to the IR system deployed in the Torii real setting: we implemented the Information Retrieval Assistant (IRA), a system providing various kinds of suggestions to users that are searching the paper and e-print database available in the Torii portal.

The most important innovative features in IRA concern the new models on which ephemeral personalization, i.e., both terminological and strategic suggestions, is based. Terminological help is obtained by a new spreading activation algorithm capable of browsing an heterogeneous, dynamically generated, and integrated thesaurus, starting either from the last inserted search term, or from the set of all the

search terms used by the user so far. Terminological help is then presented to the user by means of a ranked list of suggested terms.

The basic reasoning process exploited for ephemeral personalization is described in the following. Each user *action* (i.e., any operation performed by the user) on the user interface is notified to IRA. IRA monitors these time-stamped actions and builds a model of the session history, that is made up by a sequence of interleaved actions and states. A *state* is a set of parameters describing the current state of the system, like number of terms in the query, number of retrieved, read, and judged (as relevant or not relevant) documents, etc. At each state, i.e., after each action, a new set of situations is inferred. A *situation* is a history pattern, or an abstract description of the session history. Situations can be very simple, like 'insertion of a zero posting count term in the query' (a term that is not contained in any document), or they may concern a longer time interval, like 'two consecutive searches with no changes to the query'. Moreover, they can be more abstract and difficult to infer certainly, like 'user not reading the content of the retrieved documents'. The derivation of new situations is triggered by the last user action, but takes into account the whole session history.

From each situation, a set of *suggestions* is derived. One of the most important suggestions is terminological help, but IRA suggestions also include simple *hints*, that merely make aware the user of alternative actions (like reminding the user to have a look at the full text of the documents, or to judge, by clicking on the appropriate button, the relevance of the read documents), and more complex *advices*, i.e., a set of operations which are carried out collaboratively by the user and IRA (like author search, that suggests to look for documents written by the same author as the documents already judged relevant by the user). IRA suggestions are always contextual and are provided in two kinds of situations: *critical* (i.e., the user is experiencing some problem, as repeatedly retrieving no documents, or not making progress) and *enhanceable* (i.e., when the user could follow other – possibly more – appropriate alternative routes). Finally, IRA suggestions are ranked and proposed to the user.

We performed a laboratory evaluation that highlighted some positive qualitative results: the sample users that used IRA were satisfied with the adequacy, timeliness, comprehensibility, and usefulness of the suggestions. Moreover, as foreseen, terminological help has been especially appreciated.

4. Conclusions and Future Work

In this paper we have shown how persistent and ephemeral personalization techniques can be exploited to implement more adaptive and effective information access systems. More specifically, the research presented here approaches two problems of the user of a scholarly publishing system: the need to be timely and accurately updated about new relevant information and the request for adequate, effective and easy-to-use support during search of archive information. Several experimental results show that persistent personalization is useful for information filtering systems, and ephemeral personalization leads to more effective and usable information retrieval systems.

References

1. F.A. Asnicar, M. Di Fant, C. Tasso, User Model-Based Information Filtering, in M. Lenzerini ed. *AI*IA 97: Advances in Artificial Intelligence – Proc. of the 5th Congress of AI*IA*, LNAI 1321, Springer, Berlin, D, 1997, 242-253.
2. R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, New York, NY, USA, 1999.
3. N.J. Belkin, Helping People Find What They Don't Know, *Comm. of the ACM* 43(8), 2000, 59-61.
4. N. Belkin, C. Cool, D. Kelly, S.-J. Lin, S.Y. Park, J. Perez-Carballo, C. Sikora, Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval, *Information Processing and Management* 37(3), 2001, 403-434.
5. N. Belkin, C. Cool, A. Stein, U. Thiel, Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems, *Expert Systems with Applications* 9(3), 1995, 379-395.
6. D. Billsus, M.J. Pazzani, User Modeling for Adaptive News Access, *User Modeling and User-Adapted Interaction Journal* 10(2-3), 2000, 147-180.
7. G. Brajnik, S. Mizzaro, C. Tasso, Evaluating User Interfaces to Information Retrieval Systems: a Case Study on User Support, *Proc. of the 19th Annual International ACM SIGIR Conference*, Zurich, CH, 1996, 128-136.
8. G. Brajnik, S. Mizzaro, C. Tasso, F. Venuti. Strategic help in user interfaces for information retrieval, *J. of the Am. Soc. for Information Science and Technology*, 53(5):343-358, 2002.
9. G. Brajnik, C. Tasso, A shell for developing non-monotonic user modeling systems, *International Journal Human-Computer Studies* 40, 1994, 31-62.
10. W.B. Croft, S. Cronen-Townsend, V. Lavrenko, Relevance Feedback and Personalization: A Language Modeling Perspective, *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001, www.ercim.org/publication/ws-proceedings/DelNoe02/.
11. E.N. Efthimiadis, Query expansion, *Annual Review of Information Science and Technology (ARIST)*, M. E. Williams ed., vol. 31, 1996, 121-187.
12. U. Hanani, B. Shapira, P. Shoval, Information Filtering: Overview of Issues, Research and Systems, *User Modeling and User-Adapted Interaction* 11(3), 2001, 203-259.
13. P. Ingwersen, *Information Retrieval Interaction*, Taylor Graham, London, UK, 1992.
14. A. Jameson, C. Paris, C. Tasso eds., *User Modeling – Proc. of the 6th Intl. Conference UM97*, Springer-Verlag, Wien New York, 1997.
15. T. Malone, K. Grant, F. Turbak, S. Brobst, M. Cohen, Intelligent information sharing systems, *Comm. of the ACM* 43(8), 1987, 390-402.
16. R. Mandala, T. Tokunaga, H. Tanaka, Query expansion using heterogeneous thesauri, *Information Processing & Management* 36, 2000, 361-378.
17. M. Minio, C. Tasso, User Modeling for Information Filtering on Internet Services: Exploiting an Extended Version of the UMT Shell, *UM96 Workshop on User Modeling for Information Filtering on the World Wide WEB*, Kailua-Kona, Hawaii, USA, January 1996.
18. S. Mizzaro and C. Tasso. Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web. In Paul De Bra and Peter Brusilovsky editors, Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, In press. Malaga, Spain, 29-31 May 2002.
19. C. Tasso, M. Armellini, Exploiting User Modeling Techniques in Integrated Information Services: The TECHFINDER System, in E. Lamma and P. Mello eds., *Proc. of the 6th Congress of the Italian Association for Artificial Intelligence*, Pitagora Editrice, Bologna, I, 1999, 519-522.
20. C. Tasso, P. Omero, *La personalizzazione dei contenuti Web: e-commerce, i-access, e-government*, Franco Angeli, Milano, I, 2002.