Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web

Stefano Mizzaro and Carlo Tasso

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

Abstract. We show how personalization techniques can be exploited to implement more adaptive and effective information access systems in electronic publishing. We distinguish persistent (or long term) and ephemeral (or short term) personalization, and we describe how both of them can be profitably applied in information filtering and retrieval systems used, via a specialized Web portal, by physicists in their daily job. By means of several experimental results, we demonstrate that persistent personalization is needed and useful for information filtering systems, and ephemeral personalization leads to more effective and usable information retrieval systems.

1 Introduction

Oversupply of information constitutes a well known phenomenon that is progressively becoming worse and that threatens Web usefulness. Other related issues are information waste (documents published on the Web do not always reach the appropriate readers, or reach them too late), and low information quality (the amount of available information is increasing, but its quality is decreasing). These problems are very general, and affect all kinds of Web contents, i.e., information to be accessed, products to be purchased, and services to be exploited. Personalization allows a more effective information access: end users can be delivered personalized information, tailored to their individual needs, and, more generally, it enables a more effective and efficient transfer of the published information from the authors to the most appropriate readers.

At the University of Udine, we have been investigating the issue of personalization in information access for several years [1, 8, 9, 10, 20, 22]. In this paper, we present the most recent results concerned with the application of adaptive and personalized information access to the electronic publishing field, and more specifically in scholarly publishing. We claim that personalization is needed and useful in information access, and especially in scholarly publishing, where users (i.e., researchers) are interested in it for two important reasons: (i) detecting newly published information which is relevant to their interests and preferences, and (ii) accessing stored information for satisfying specific information needs. However, this twofold situation requires a novel approach, in which two distinct and complementary

P. De Bra, P. Brusilovsky, and R. Conejo (Eds.): AH 2002, LNCS 2347, pp 306–316.

personalization techniques (i.e., ephemeral and persistent personalization) are applied together to meet user's requirements.

This paper is structured as follows. Section 2 briefly describes the world of scholarly publishing, and the heavy changes introduced in it by the Web. In Sect. 3 we present a short overview of information access approaches, and discuss how personalization techniques can be useful in this field. In Sects 4 and 5 we show the application of personalization techniques to information access to scholarly publications. Section 6 closes the paper and sketches some future developments.

2 Scholarly Publishing and the Web

The communication mechanism adopted by science today arose in the 17th Century, with the publication of the first scientific journals. Since about 1930, the dissemination of scholarly information is based on peer review, that usually assures a high quality of the published papers. Internet has changed, and is changing, this situation [3]. Now 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). The repositories usually contain *preprints*, i.e., electronic versions of submitted papers made publicly available before review, acceptance, and, possibly, publication.

The exploitation of the Web has also highlighted another essential characteristic of scholarly publications, i.e., their hypermedia nature. A rich hypertext structure is provided by both the citations across publications and the references (to chapters, sections, figures, and so on) within each publication. Multimediality is also important since it leads to a more effective communication, and even though still limited today, it will increase in the next years. Another aspect that further extends the richness of the hypermedia structure is the storage (easy obtainable on electronic media) of a publication as a multilayered document (dlp.cs.berkeley.edu), that includes the various versions of a document, the slides and presentations resulting from it, the referees' and readers' comments, and any other remark about the document. This provides a richer information on the topic at hand, and adds more hypermedia information as links among the various layers.

All these new means available to authors allow an ever growing rate of production of scholarly articles (see, e.g., the arXiv usage statistics: arXiv.org/show_stats). The new Web-based approach guarantees easier access, more powerful and richer means of information seeking, and better timeliness, but it features also some drawbacks (e.g., the quality of preprints is not assessed by a peer review process) and poses some new problems (copyright problems, social and legal acceptance, and so on). However, after a slow start [15], the impact of scholarly publishing is steadily increasing [11].

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, as shown in the next section. 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.

3 Personalized Information Access

Information access is the process exploited by a seeker who wishes to find and retrieve some data/information which satisfies an *information need*. It is common to distinguish between two kinds of information access:

- *Information retrieval* (*IR*) [2] is characterized by a static database of documents, a short term information need, and a query made up by a few (usually less than two) terms. Web search engines are the most known instance of IR systems. It is well known that IR is a difficult task [4], since users have to specify in a query something that they do not know (if they knew it, they wouldn't be searching for it).
- *Information filtering* (*IF*) [14], on the other side, is characterized by a dynamic database (actually an incoming flow of documents) and a long term and rather static information need. Users of IF systems are more motivated to express their information needs as more accurate and complete descriptions that will last for longer time. These descriptions are usually called *user profiles* (or *models*), and are made up by a lot of data: concepts, relationships among them, weights, etc.

Web personalization is the process of selecting, preparing and delivering Web contents for a given user, by taking into account his specific needs and preferences [23]. Personalization means delivering to the user the most relevant contents, in the most adequate way, and at the most appropriate time. A personalization system is based on three main functions, which all can be performed in a personalized way: selection, visualization, and delivery. In order to be personalized, all the three functions have to be supported by specific information about the user, which is included in a user profile and has to be available when the personalization process takes place. In this paper we deal only with the first of the three functions, i.e., selection of the most appropriate content.

Personalization techniques are very numerous and are ranging from simple user-controlled customization of Web content, to autonomous system-controlled adaptation [17, Reader's Guide, p.6]. We distinguish two types of personalization [23]: *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. Usually, but not necessarily, the information exploited for building the profile comes from various sources, it concerns different aspects of the user, and it is often extended by means of (possibly sophisticated) reasoning or learning processes. In ephemeral personalization, the information used to build the user profile is gathered during the current session only, and is immediately exploited for executing some adaptive process aimed at

personalizing the current interaction. At the end of each session, the user profile is lost, and no information about the user is stored in a persistent way for later use.

Information access systems should and can exploit both kinds of personalization [4, 12]. We propose here a twofold approach. 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, in IR persistent personalization is not feasible, 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 used in an effective way, with the goal of modeling the search session, rather than the information need, for immediately providing personalized support during the searching session. The idea of long and short term modeling in information access is not new (see, e.g., [7]), however it has been considered from the IF perspective only, i.e., it consists in building user profiles across a shorter or longer period of time (a limited number of sessions or very many sessions), and the profiles, in both cases, model only the topics interesting for the user. Our approach is innovative for two reasons: (i) short term modeling is performed through ephemeral personalization, restricting 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. This novelty allows to provide adaptive support to the user, as it will be shown in Sect. 5.2.

We have experimented this twofold approach in scholarly publishing portals for physics. We chose that community since the physics (especially high energy physics) field seems well ahead in exploiting the full potential of web publishing (no surprise, since the Web was born at CERN, one of the major physics institutions worldwide): the above cited arXiv repository (formerly known as xxx) is already a used, valid, and widely accepted media for physics and astronomy fields [11], and the SPIRES (www.slac.stanford.edu/spires/hep) citation index is almost three times more complete than the ISI well known database. In the next sections we present an application of persistent and ephemeral personalization within the Torii vertical portal (torii.sissa.it) on physics, which has been developed in the 5th FP IST project TIPS (Tools for Innovative Publishing in Science), see tips.sissa.it.

4 Persistent Personalization in Information Filtering

In previous work, we have developed and evaluated several content-based filters [18] for persistent personalization. Among them [1, 20, 22], the most effective has been the information agent *ifT* (*information filtering Tool*) [20], which is based on the user modeling shell *UMT* (*User Modeling Tool*) [10]. The work presented here concerns the exploitation of ifT in the Torii portal.

4.1 Content-Based Filtering through ifT

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]. In one of them, four subjects received

2000 documents (20 each day, for 100 days) on various computer science topics. Each subject was interested in some specific area(s) of computer science, and ifT was filtering and ranking the incoming documents according to their relevance. Initially the user profile was empty, and the user was allowed to 'explain' his interests through relevance feedback only. Throughout the experiment, standard precision and recall were measured. Fig. 1 shows the evolution of precision over time (100 sessions): dots represent the observed data, the irregular line represents the moving-average of order 5, and the regular line is an interpolation curve. The results show good learning capabilities (a precision of 80% is reached after 8 sessions), as well as a very high final precision value which saturates at about 92% in the interpolation model.

Another significant application of ifT has been developed within the *ifWeb* system for filtering Web documents [22]. The system includes the information agent *ifSpider*, aimed at the autonomous navigation of the Web for searching documents relevant to a specific user profile. The navigation performed through hyperlinks is opportunistic: only the paths including documents which feature relevance scores above a given threshold are considered.

ifWeb has been evaluated in several laboratory experiments. In one of them, devoted to assess its ranking capabilities, each subject was initially defining a profile through relevance feedback given on 4-6 documents, and then he was performing a series of nine sessions with ifWeb. After each session, the subject was requested to provide the correct ranking of the documents given by ifWeb, and human and system rankings were compared. Fig. 2 shows precision (continuous line) and the ndpm measure [24], which evaluates the difference between the two rankings (good performance is indicated by decreasing values). After the first sessions, as precision reaches good values, the ndpm starts to decrease, indicating the capability to produce a better ranking.

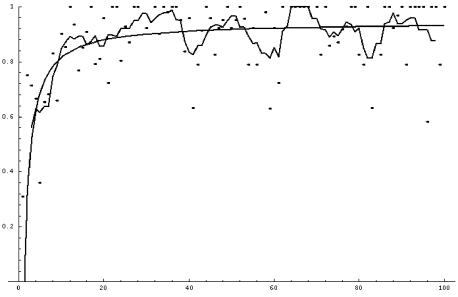

Fig. 1. Precision of ifT over 100 sessions.

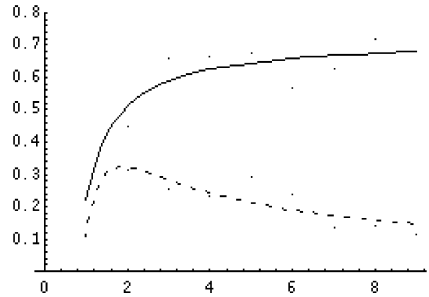

Fig. 2. Precision and ndpm measures of ifWeb over 9 sessions.

4.2 Exploiting Content-Based Filtering in Electronic Publishing

Given the performance reached by ifT, we decided to adopt ifT 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 available in Torii. Normal users (researchers in high energy physics) are 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 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. Fig. 3 shows a snapshot of the Torii portal.

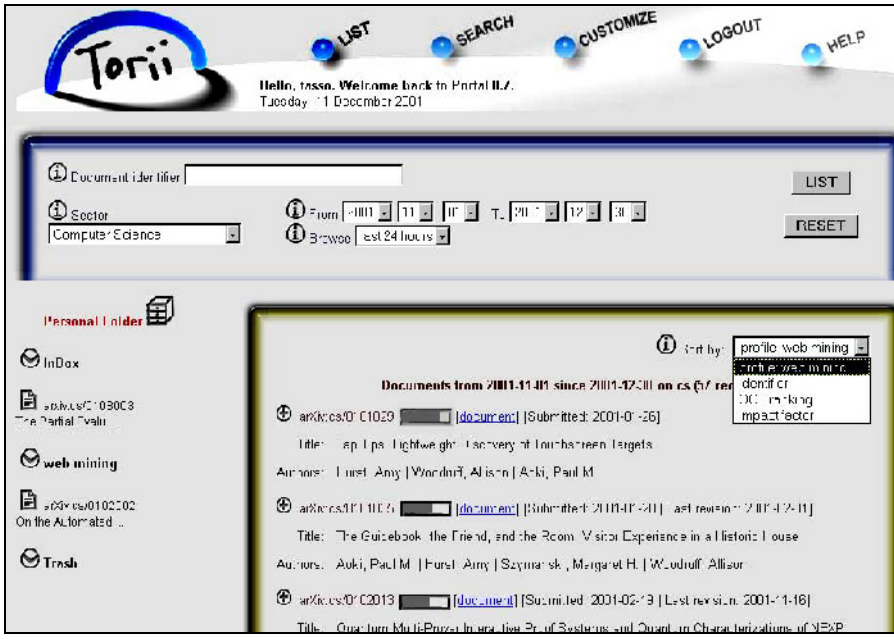

Fig. 3. Torii portal: documents ranked by the ifT filtering engine.

The relevance measure produced by ifT and exploited in Torii is a unique figure (see the bars shown in Fig. 3) which combines the document *topicality* value, i.e., a measure of how much concepts relevant for the user are present in the considered document, and the *conceptual coverage* value, i.e., a measure of how many of the concepts relevant for the user are present in the considered document.

Torii has undergone a first 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. SISSA (the managing institution of Torii), has decided to use the filtering engine ifT as a standard tool available to all users of their portal.

5 Ephemeral Personalization in Information Retrieval

5.1 Supporting Users of IR Systems

The interactive nature of IR is advocated since years [16] and is now widely accepted: between the user and the IR system a dialogue takes place [6], during which the user should receive adequate support [4]. 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 [5]. A basic kind of support is *terminological* help, which identifies and suggests to the user terms that improve the query [13, 19]. 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 [9]).

We propose to 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 prototype [8] that, by means of thesauri, is capable of suggesting to the users of a boolean IR system alternative terms to better (re)formulate their information needs. After a carefully designed laboratory experiment involving 45 participants, we had evidence that terminological help alone is useful, but needs to be complemented by strategic help.

On the basis of these results, we added to FIRE a Strategic Aid Module (SAM) capable of providing to the users suggestions on which strategies are more likely to be effective in a certain situation [9]. SAM is based on a collaborative (between the user and the system) view of the session: users know their needs, judge the relevance of the retrieved documents, select the terms to be added to the query, and retain the control of the session; the system monitors users' actions and provides contextual suggestions, proposing alternative routes, emphasizing mistakes (e.g., term spelling), and so on. SAM is based on a detailed conceptual model of the session, made up by representing user actions, the current situation of the session, and the set of feasible and more appropriate suggestions. By exploiting a knowledge base, the current situation of the session is inferred from the actions made by the user, and personalized suggestions are selected on the basis of the current situation. We performed two laboratory experiments (one in which we simulated the activity of the users of the previous FIRE experiment, and one that involved six new participants), both of which showed that strategic support is useful, well accepted, and it allows users to learn the best strategies.

5.2 Supporting Users of IR Systems on the Web

Following the positive evaluations of the two prototypes mentioned above, we decided to apply ephemeral personalization to an IR system deployed in a 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. IRA exhibits some innovative features with respect to the previous two prototypes. It fully integrates terminological and strategic suggestions. The underlying IR system is a probabilistic one (Okapi, see web.soi.city.ac.uk/research/cisr/okapi/okapi.html) in place of a boolean one, and it works on an underlying

full text database, containing almost 200,000 scholarly documents about physics (as opposed to the bibliographic, and smaller, database used in the previous experiments). IRA is designed to be deployed in a real life environment, and used by physicists in their daily job. IRA can also be easily tailored to be used with other IR systems.

However, the most important innovative features in IRA are on the conceptual side, and concern the new models on which ephemeral personalization, i.e., both terminological and strategic suggestions, is based. The sorted term lists suggested in terminological help are obtained by a new spreading activation algorithm capable of browsing heterogeneous, dynamically generated, and integrated thesauri, starting either from the last inserted search term, or from the set of all the search terms used by the user so far. This new version of terminological help has shown, by means of an experimental evaluation involving six participants, significant improvements with respect to the terminological help previously used in FIRE: more terms are suggested (since more term sources are used), they are more adequate to the current context and ranked in a better way (mainly for two reasons: the synergy among the different term sources and the new spreading activation algorithm).

The enhanced reasoning process for suggestion generation is represented in Fig. 4. Each user *action* (i.e., any operation performed by the user, such as term insertion/removal/modification, search in the database, document reading, relevance judgment, etc.) on the Okapi user interface is notified to IRA by Okapi. 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 can concern a longer time interval, like 'two consecutive searches with no changes to the query'. Moreover, they can be more

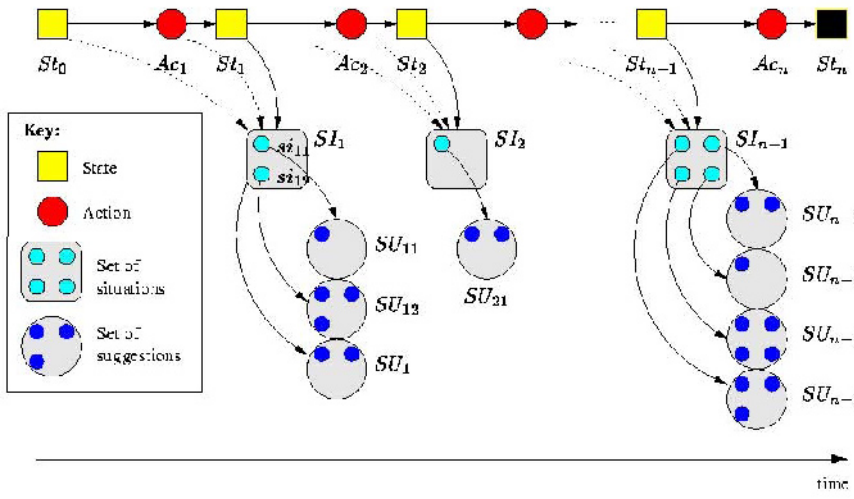

Fig. 4. IRA reasoning process.

abstract and difficult to infer certainly, like 'user not reading the content of the retrieved documents'. Situation derivation 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*, that 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 as two types of textual descriptions, a short one and a longer one displayed on demand. They are shown in IRA own window on the background, thus allowing the user to maintain the control of the interaction with the IR system. The user can either accept the suggestions received (e.g., he can insert into the query some new relevant term provided by terminological help), or can ignore them. IRA knowledge bases by now contain 33 actions, 28 situations, and 20 suggestions, and are still being extended.

We performed a first laboratory evaluation that highlighted some positive qualitative results: the sample users that used IRA are satisfied with the adequacy, timeliness, comprehensibility, and usefulness of the suggestions. Moreover, as foreseen, terminological help is especially appreciated. IRA is now being deployed and used by real end-users, and another more extensive evaluation of it will take place in the next months.

6 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.

So far, we have kept separated the two approaches, but they naturally complement each other. Therefore, we plan to integrate them in various ways: the long term user profile can be used in IR, e.g., to rank the retrieved documents in a more personalized way; vice versa, the suggestions can be useful during the initial construction of the profile, or during feedback iterations. We also believe that these personalization techniques can be fruitfully applied also outside the scholar community, for instance in the more general context of electronic publishing, where various media such as newspapers, magazines, news agencies, and so on are continuously fed with new information. Finally, the quality of information is another important issue, that we

have not considered in this paper, and that we are approaching with a collaborative work approach [21].

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. C.W. Bailey, Jr., Scholarly electronic publishing bibliography - Version 40: 12/12/01, http://info.lib.uh.edu/sepb/sepb.html, visited 5/1/02.
4. N.J. Belkin, Helping People Find What They Don't Know, *Comm. of the ACM* 43(8), 2000, 59-61.
5. 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.
6. 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.
7. D. Billsus, M.J. Pazzani, User Modeling for Adaptive News Access, *User Modeling and User-Adapted Interaction Journal* 10(2-3), 2000, 147-180.
8. 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.
9. 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*, 2002, in press.
10. G. Brajnik, C. Tasso, A shell for developing non-monotonic user modeling systems, *International Journal Human-Computer Studies* 40, 1994, 31-62.
11. C. Brown, The E-volution of Preprints in the Scholarly Communication of Physicists and Astronomers, *J. of the Am. Soc. for Information Science and Technology* 52(3), 2001, 187-200.
12. 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/.
13. E.N. Efthimiadis, Query expansion, *Annual Review of Information Science and Technology (ARIST)*, M. E. Williams ed., vol. 31, 1996, 121-187.
14. 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.
15. S.P. Harter, Scholarly communication and electronic journals: An impact study. *J. of the Am. Soc. for Information Science* 1998, 49(6), 507-516.
16. P. Ingwersen, *Information Retrieval Interaction*, Taylor Graham, London, UK, 1992.
17. A. Jameson, C. Paris, C. Tasso eds., *User Modeling – Proc. of the 6th Intl. Conference UM97*, Springer-Verlag, Wien New York, 1997.
18. T. Malone, K. Grant, F. Turbak, S. Brobst, M. Cohen, Intelligent information sharing systems, *Comm. of the ACM* 43(8), 1987, 390-402.
19. R. Mandala, T. Tokunaga, H. Tanaka, Query expansion using heterogeneous thesauri, *Information Processing & Management* 36, 2000, 361-378.
20. 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.
21. S. Mizzaro & P. Zandegiacomo Riziò. An automatically refereed scholarly electronic journal: Formal specifications. *Informatica - An International Journal of Computing and Informatics* 24(4), 2000, 431-438.
22. 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.
23. C. Tasso, P. Omero, *La personalizzazione dei contenuti Web: e-commerce, i-access, e-government*, Franco Angeli, Milano, I, 2002.
24. Y.Y. Yao, Measuring retrieval effectiveness based on user preference of documents, *J. of the Am. Soc. for Information Science* 46(2), 1995, 133-145.