

Evaluating a Personal Information Assistant

Ioannis Psarras and Joemon Jose

Department of Computing Science
University of Glasgow
Glasgow, UK, G12 8QQ
{psarras, jj} @dcs.gla.ac.uk

Abstract

Personal Information Assistants that search on user's behalf aim to fetch relevant documents on a regular basis. We have developed such assistant system and evaluated it using a long-term, task-oriented approach involving real users. Current evaluation methodologies are inadequate and hence have resorted to a long-term real user study. This paper describes the design and development of our system and the results of the evaluation study.

1. Introduction

Web search engines, designed for discovering documents online, are very popular and are generally perceived to do an excellent job. However, recent studies, such as [Jansen et al, 2001; Jansen et al, 2000], have demonstrated the deficiencies of web search tools. Among other things, they have highlighted the fact that users often find it hard to transform their search needs to an appropriate set of query terms. Retrieval tools depend on the query terms or the quality of query representation in producing initial set of relevant results. In addition, these studies demonstrated that searchers continuously look for information on the same or similar topics, for example things that relate to their work.

Often such information requirements change by sliding into new topics, based on the changes of user interests. The only way to satisfy evolving needs is to search on a continuous basis [Bates, 1989]. However, performing this task manually is impossible within the time constraints of most users. Unfortunately, no search engines currently help searchers in satisfying their evolving information needs.

During the past years, a great amount of research has been carried out in the field of information assistants. Recommendation systems have become popular in the field of search assistants [Chen et al, 1998; Lieberman, 1995; Stefani et al, 1998], as well as in the area of e-commerce (e.g. Amazon). Most of them base their information filtering and recommendation services on complex machine learning algorithms and different architectures. However, none of these systems are evaluated properly, let alone in a real user-based environment. Also, due to lack of an established evaluation methodology, comparison of such systems is not possible.

In this paper, we describe the development and evaluation of a pro-active personalized information assistant, PIA, based on implicit feedback gathering techniques and collaboration algorithms to implement personalized document recommendations. We have employed a task-oriented, user-centered evaluation methodology, where 19 users used the system regularly for 7 days. A baseline system (Google) has also been used to demonstrate whether web search engines can benefit from the integration of personalization and user profiling features in their retrieval strategy. The experimental results indicate that the Personalized Information Assistant is effective in capturing and satisfying users' evolving information needs and providing additional

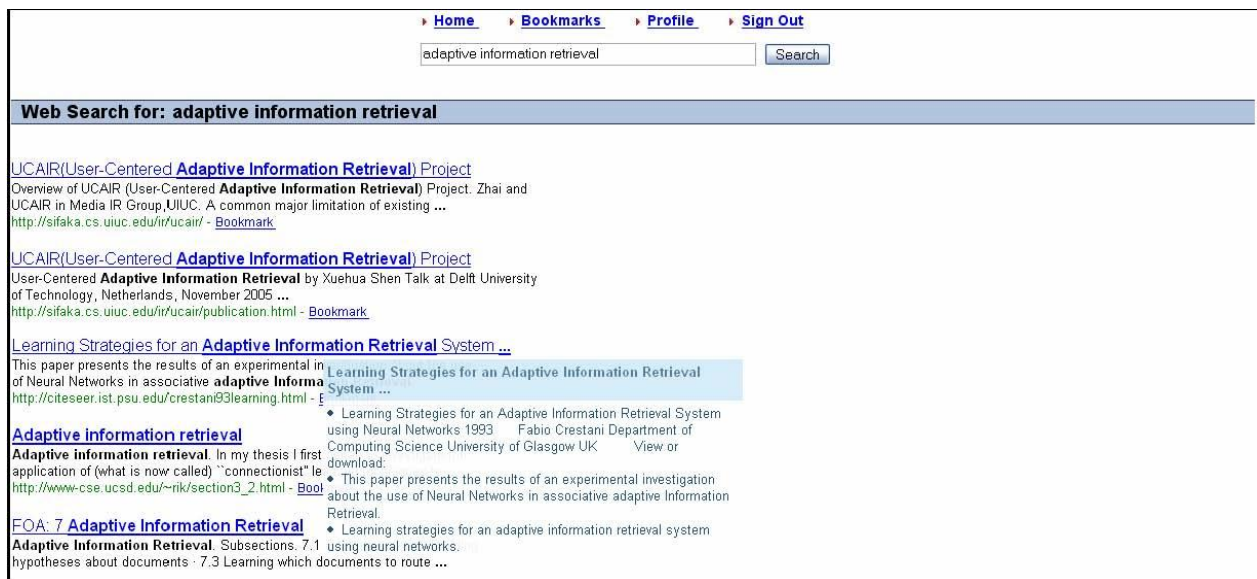
information on their behalf.

2. System Overview

Personal Information Assistant (PIA) is developed as an adjunct to the current web search engines. The system is situated on a server and interacts with users using Java's servlet technology. User queries and other interaction data are captured and processed at the server. The queries are forwarded to Google and the results are parsed and presented to the user. Also, we have employed a combination of implicit and explicit feedback to create and analyze user interests [Kim et al, 2000; Oard et al, 1998; White et al, 2002]. A set of user interests form the user's profile. Profile and interest representation are discussed in more detail in the next paragraph. A detailed analysis of PIA and the integrated algorithms has been presented in [Psarras et al, 2006].

2.1 The User Interface

The main user interface feature is an integrated search engine, which communicates with Google to obtain and present the results. A query-biased summarization system, described in [White et al, 2003], has been integrated to allow users to get a more clear idea of each result. Search result presentation is presented in Figure 1, while the effect of summaries is also visible through a mouse-over tooltip behavior.



The screenshot displays a web search interface. At the top, there is a navigation menu with links for Home, Bookmarks, Profile, and Sign Out. Below the menu is a search bar containing the text 'adaptive information retrieval' and a 'Search' button. The search results are displayed below the search bar, starting with a header 'Web Search for: adaptive information retrieval'. The results include several entries, each with a title, a brief description, and a URL. The first entry is 'UCAIR(User-Centered Adaptive Information Retrieval) Project' with a description and a URL. The second entry is 'UCAIR(User-Centered Adaptive Information Retrieval) Project' with a description and a URL. The third entry is 'Learning Strategies for an Adaptive Information Retrieval System ...' with a description and a URL. The fourth entry is 'Adaptive information retrieval' with a description and a URL. The fifth entry is 'FOA: 7 Adaptive Information Retrieval' with a description and a URL. The sixth entry is 'Adaptive Information Retrieval' with a description and a URL.

Figure 1: Search result presentation and the effect of summary generation.

In addition, users can view documents recommended by the system through their personalized home page. Through this page users can browse and access the documents recommended by the system, as well as bookmark the more important ones and delete those not related to their search needs. A user profile is presented as a set of interests, which essentially work as document collections for documents related to this particular topic/interest. Figure 2 illustrates this particular feature.

Welcome to your PIA homepage, demo	
retrieval information	adaptive recommendation
<p>Access percentage : 57% edit delete</p> <p>Integrated multilevel secure system for information retrieval in ... Integrated multilevel secure system for information retrieval in distributed computer systems. Source, Proceedings of the 12th international conference on ... http://portal.acm.org/citation.cfm?id=276088.276186&coll=GUIDE&dl=&CFID=15151515&CFTOKEN=6184619 bookmark delete</p> <p>Freenet: A Distributed Anonymous Information Storage and Retrieval ... Freenet: A Distributed Anonymous Information Storage and Retrieval System. In Proc. of the ICSI Workshop on Design Issues in Anonymity and Unobservability, ... http://citeseer.ist.psu.edu/clarke00freenet.html bookmark delete</p>	<p>Access percentage : 14% edit delete</p> <p>Active Recommendation Project Adaptive Recommendation Project for the Library Without Walls ... (ARP) is developing research on recommendation systems for large databases and the WWW, ... http://arp.lanl.gov/ bookmark delete</p> <p>An Adaptive Recommendation System without Explicit Acquisition of ... AN ADAPTIVE RECOMMENDATION SYSTEM. 175. the confidence values. These confidence values are corrected by the GA-based learning ... http://infolab.usc.edu/DocsDemos/YodaExtend.pdf unbookmark delete</p> <p>Project-Team-AxiS: Supporting Information Retrieval with adaptive ... A major challenge in the field of recommender systems design is the following: How to produce adaptive recommendations of high quality minimizing the effort ... http://www.inria.fr/rappportsactivite/RA2005/axis/uid41.html bookmark delete</p>

Figure 2: The personalized home page, showing recommended documents discovered by the system, as well as the interests in a user's profile.

Finally, as explained in section 2.2, explicit feedback is not required to return high quality document recommendations, in terms of accuracy and relevance to the subject. However, explicit profile manipulation can be satisfied through an interface illustrated in figure 3. Through this page, users can add/delete interests or remove/append terms to existing interests. Figure 3 also gives an idea of how user profiles are represented.

Welcome to your profile manager, demo	
Your Interests	
adaptive recommendation	adaptive recommendation systems <input type="checkbox"/>
retrieval information	adaptive computer conference distributed info <input type="checkbox"/>
Want to add more interests?	
Interest name :	<input type="text"/>
Interest terms :	<input type="text"/>
<input type="button" value="Save Profile"/>	

Figure 3: The profile management page allowing users to manipulate the interests in their profile.

2.2 Relevance Feedback Gathering

Our information assistant gathers information from a number of sources implicitly, aiming to reduce the cognitive load imposed by explicit user ratings. Explicit feedback can optionally be used, through the user interface presented in figure 3, to amend the interests created in each user's profile. Click-through data collected from user interaction with search results are analyzed to gather user's implicit interest in some documents, which are taken into account in user profile modeling. Users initially assess the relevance of search results from the title and snippets of search result. Query-biased summaries [White et al, 2003] have also been employed and implemented in PIA. As demonstrated in [White et al, 2003], incorporating summaries into search results presentation, facilitates more interaction with the system and allows users to assess the relevance of documents more accurately.

Furthermore, each bookmarked item is implicitly denoted as an item of interest and is further processed during the profile modeling process. Also, such documents can be further accessed in the bookmark portal, where users can browse past bookmarks. On the contrary, document deletion does not provide implicit knowledge for the future. In other words, a deleted item is only removed from the document collection without affecting the relevance feedback gathering

process.

2.3 Profile/Interest Representation

In the past, several recommender systems have attempted to represent user needs as a single keyword vector, [Lieberman, 1995; Stefani et al, 1998]. Such systems do not recognize the multiple facets of user interests. A query about “*business administration*” may denote the latest news in this field, but it may also denote the education and career prospects of this area. PIA attempts to overcome this issue by representing a user’s profile as multiple weighted keyword vectors. The weight of each keyword is calculated as a result of the profile modeling algorithm. Essentially, a profile is a set of user interests, which in turn are represented as a weighted keyword vector.

2.4 Profile Modeling

Modeling user profiles effectively continues to be an active research area in information retrieval. Information needs and interests are volatile, thus tools that aim to capture user profiles must evolve and adapt rapidly and efficiently. The profile modeling algorithm implemented in Personalized Information Assistant uses implicit data collected during search iterations and user interaction with the system, to amend user interests accordingly. Search results that have been implicitly denoted as interesting, via click-through behavior, are parsed and summarized to a set of representative terms for this query. We use term frequency (TF) to discover the most informative terms with respect to the user’s query, but alternative strategies were tested, such as a Binary Voting model and a TF-IDF scheme. At this point, user interests can be amended or created depending on the similarity between the current user profile containing interests related to the extracted set of terms. More specifically, a single-pass clustering algorithm is applied, using cosine coefficient as the similarity matching function, in order to detect various facets’ of user interests.

Finally, similar to [Martin et al, 2003], the document recommendation algorithm is invoked regularly, which formulates a new query based on the keyword vectors in each interest of the user profile. The top-ranked results for the query issued are recommended as part of the respective user interest.

3 Evaluation Issues

Evaluation of recommendation systems and search assistants is a challenging task due to the variance of possible scenarios that can occur. The current methodologies in IR have failed to address the evaluation of such personalized systems [Borlund, 2000; Harman, 1992; Jose et al, 1998]. Variations in recommendation algorithms and system implementations make evaluation even harder. As Konstan et al (1999) indicate, existing approaches to evaluating recommender systems can be divided into two categories. In off-line evaluation techniques the performance of a recommendation mechanism is evaluated on existing datasets, while in on-line evaluation methodologies, the performance is evaluated on users of a running recommender system.

According to [Konstan et al, 1999], the main issue with on-line evaluation is the need to field a fully engineered system and build up a community of users. There are problems in getting enough participants, for long-term studies, who are eager to use a prototype system frequently. But as long as the system is in a working condition and the participants have been found, on-line evaluation techniques are inarguably better than alternatives. However, such methodologies developed for the evaluation of systems like music and film recommendations are unsuitable for

evaluating personalized information systems.

The current evaluation methodology applied to IR systems, as used in TREC, exclude the users from the evaluation stage [Harman, 1992]. Similarly, the interactive evaluation methodologies, largely laboratory-based, do not allow the evaluation of systems dealing with long-term information needs.

In the next section we will present an approach towards the evaluation of our Personalized Information Assistant and discuss the results gathered after a long-term experiment with real user base.

4 Evaluation Methodology

In order to evaluate the system as realistically as possible, we employed a user-centered, task-based evaluation methodology [Borlund, 2000; Jose et al, 1998]. In our experiments users were asked to use the system regularly for 7-10 days. Since most people participated in the experiments for 7 days, we used the results only up to 7 days.

4.1 Participants

A total of 19 users were recruited for our experiment. Participants were mainly students of various nationalities and backgrounds contacted through email. Furthermore, due to the length of these experiments, the system was used by each of them from their own environment, such as their home or office, so they had no additional guidance. Apart from this, our aim was to conduct the experiments in a real, everyday environment, and not a laboratory-based setting, where user search behavior would be affected.

Our recruitment was specifically aimed at targeting non-professional computer users and mostly those classified as inexperienced. Overall our subjects had an average age of 22 with a range of 10 years (youngest 19 years; oldest 29 years). A total of 14 participants were categorized as inexperienced, based on the data collected through the entry questionnaire, against only 5 experienced users. The classification between experienced and inexperienced searchers was made on the basis of the subjects' responses to questions about the level of their computing, Internet and web searching experience. More experienced searchers were those who used computers and searched the web on a regular, often daily basis. Inexperienced searchers were those who searched the web and used computers but on a more infrequent basis.

4.2 Tasks and Data Collection

In our experiments each subject was asked to complete at least two search tasks. These tasks were used to investigate the effectiveness of the two systems in the following aspects: Interest modeling and overall retrieval performance.

A common task was given to all participants, in order to allow us to benchmark various system aspects and compare user behavior. The details of this task are given below:

“Assume you are a politician invited to give a talk at the local university about your views on middle east politics. You have decided to talk about recent conflicts in Iraq. Try to find more documents about this subject.”

In addition, we encouraged participants to use the system for their own search tasks with the

requirement that at least one of their personal interests was investigated. Throughout the whole evaluation, we collected a combination of qualitative and quantitative data, through continuous, extensive background logging. Also, we used questionnaires to collect data at various points of the experiment. The results from all completed tasks have been summarized and grouped per user. Thus, no differentiation between task results have been made.

Since this paper attempts to presents the results obtained from the evaluation of Personal Information Assistant, data collected for all search tasks are analyzed together. An analysis of results for the search tasks individually would have allowed us to discover whether a difference was observed between the participants' given tasks and their own tasks, but this was outside the scope of this evaluation. The aim of this research is to study the effectiveness of PIA in retrieving additional documents and not the variance in participants' behavior, with respect to the search task currently investigated.

4.3 Procedure

The user study was carried out in the following manner. Initially, participants were informed on the features and capabilities of our system. Upon agreement of participation, they were asked to complete an entry questionnaire with respect to their background as well as their search experience on the web.

Then, users were asked to find information on at least two topics as described in section 4.2. After using our retrieval tool for two to three days, each user was asked to complete another questionnaire, allowing us to capture their views on more important aspects. The questionnaire consisted of a combination of Likert scales, semantic differentials and open-ended questions focusing on the effectiveness of the profiling scheme, the relevance of recommended documents and their overall satisfaction from our system. All questionnaires were completed by each user individually and the experimenter was not present in these occasions.

Since the experimentation period for each user ranged between 7-10 days, we had to capture user viewpoint more than once. Therefore, after 4 days, where all participants have used the system more thoroughly, users were asked to complete the same questionnaire again. This allowed us to observe variances in user behavior between their assessments.

On the 7th day, we asked participants to discover information on the common evaluation task, section 4.2, by using a popular web search engine (Google). This acted as a kind of a baseline system for our study. Also, the idea behind this is that, from their previous search interaction, users have a clear idea of what information they need and hence they will be able to formulate an effective query. Finally, they were asked to complete another questionnaire summarizing their thoughts on the system and indicating their preference between our own search assistant and the baseline system.

4.4 Hypothesis

The research hypothesis is that our personalized information assistant can capture evolving user information needs and pro-actively fetch relevant information. User satisfaction must also be investigated. Our main hypothesis is divided into sub-hypotheses and is structured as follows:

1. Personalised information assistant can capture users' evolving needs effectively
 - a. Combination of implicit and explicit feedback techniques can model evolving user

needs effectively

b. Users avoid editing their profiles very often

2. Personal Information Assistant can satisfy user information needs effectively by fetching additional relevant documents

5. Evaluation Results

This section presents the experimental results of our study collected from 19 users during an evaluation period of 1 month, where each participant used our assistant regularly for 7 days. All results presented in this section were collected from semantic differentials and Likert scales, grading from 1 to 5. Where applicable, the standard deviation of the mean values is given in square brackets. As for the statistical tests, we opted for the non-parametric tests due to the lack of the normal distribution assumed in our data set (c.f., Hull, 1993). The Wilcoxon Signed Ranks Test was run to establish the statistical significance.

This section is structured as follows. Firstly, we present statistics on the system usage during the evaluation period. Then, the experimental results that are related to the user modeling performance are demonstrated. Thirdly, we present the results regarding the relevance of recommended documents with respect to user interests and needs. Finally, the participants' system preference is illustrated. Although no questions have been omitted from this analysis, they are presented in different order compared to the questionnaires. All questions in parts of this section are essentially grouped into sections to demonstrate and support a single aspect of the system's evaluation.

5.1 Evaluation Statistics

Analyzing the data recorded during the evaluation, we can investigate the system usage by the participants. Also, by investigating the overall usage of the system, during the evaluation period, we can observe the user behavior and implicitly record their interest in our system. Evaluation participants executed a total of 354 queries ranging from a maximum of 29 to a minimum of 9 queries per user. On average, each participant issued 18.6 queries throughout the evaluation period.

Similarly, table 1 demonstrates the average usage of the system, in minutes and in queries, on a daily basis. This was measured by monitoring the time each participant used the system every day and averaging this figure over all users. The format of each cell is *minutes spent per day / queries issued per day*. It is illustrated that, after the 3rd evaluation day, the usage rate is mostly increasing. Taking into account that, the first days of this experiments, participants needed to familiarize with PIA, such an increase in system usage indicates a trend of an implicit user interest for our system. At the same time, it can be observed that after a couple of days, when users have issued a lot of queries and the system has a finer idea of their interests, they spend more time accessing recommendations, than performing queries. This is also illustrated in the evaluation analysis of the recommended documents, presented in section 5.2.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Min	5.7 / 1	7.9 / 2	3.5 / 1	8.1 / 0	6.2 / 0	7.3 / 1	10.2 / 1
Max	15.2 / 10	13.3 / 8	19.8 / 6	14.7 / 3	19.5 / 3	22.3 / 6	20.7 / 8
Avg	12.0 / 3.9	10.8 / 4.7	13.3 / 2.7	11.5 / 1.7	13.2 / 2.0	16.1 / 2.8	17.2 / 2.6

Table 1: Average daily system usage (in minutes) per user during the evaluation period.
The format of each cell is minutes spent per day / queries issued per day.

5.2 Modeling evolving user needs

The evaluation of a personal information assistant constitutes the main contribution of this study. Evaluation data collected from the questionnaires and log file analysis are presented in tables 2 and 3.

Table 2 shows participants' subjective assessment on the profile modeling scheme. Questionnaire data were captured in two stages (after 3 and 7 days) by a 5-point scale with a high score representing a more positive perception in the analysis. As it can be seen, participants were highly satisfied from the profiling scheme, question *"How would you rate the system in profiling your interests in general?"*. The results are improved after the system has been used for more time, but not significant according to the Wilcoxon statistical test.

Similarly, question *"How often did you edit your profile to improve the system's performance?"* in table 2 illustrates another angle of the profile modeling algorithm. Participants indicated that there was no apparent need to amend their profile explicitly. The difference between the results gathered on the 3rd and those gathered on the 7th days are not statistically significant (Wilcoxon test).

The last question in table 2, *"How would you rate the system in capturing your temporal needs?"*, demonstrates the effectiveness of our system in adapting to users' evolving needs. As we can observe, participants indicated that evolving information requirements can be captured effectively. The questionnaire results gathered at the end of the evaluation illustrated an increase in user rating, but the statistical test indicated no significance between the results (Wilcoxon test).

	After 3 days	After 7 days
How would you rate the system in profiling your interests in general?	4.00 [0.74]	4.10 [0.73]
How often did you edit your profile to improve the system's performance? (smaller is less often)	1.43 [1.07]	1.34 [1.09]
How would you rate the system in capturing your temporal needs?	4.06 [0.70]	4.15 [0.68]

Table 2: The average of questionnaire data related to the effectiveness of Personal Information Assistant's profile modeling scheme. All figures are presented in two time intervals (after 3 and 7 days).

Table 3 shows data collected by analyzing the system logs, in the aspect of interest management and user satisfaction with the profile modeling scheme. As before, all figures are presented such that we can investigate data at two time intervals. So, each table column is split into two sub-columns, where the left represents data collected on the 3rd day, while the right shows data collected on the 7th day of the evaluation.

It can be seen that participants initially had an average of 6.57 interests in their profile, but the variance is quite high with some people having only 2 interests, while others had significantly more. After 7 days, interest creation has stabilized to 7.73. Also, table 3 demonstrates that the profiling algorithm initially introduced 94.8% of user interests compared to only 5.3% of explicit additions, while near the end, 94.5% of them were created implicitly against 5.5% of explicit additions. Moreover, experimental subjects interfered with their interests by an average of 18.4% during the first days of the evaluation, by either deleting or modifying interests. As it can be seen, the percentage of interests explicitly amended has dropped from 6.3% to 5.4%, which constitutes another indication of the algorithm’s effectiveness. Therefore, we can conclude that implicit feedback gathering work effectively, since the vast majority of interests were accepted by the user without any further changes.

	Interests per user		Created by system		Created by human		Explicitly deleted		Explicitly modified	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Min	2	3	2	3	0	0	0	0	0	0
Max	12	17	10	15	1	2	4	7	2	2
Avg	6.57	7.73	6.23 94.8%	7.31 94.5%	0.42 5.3%	0.43 5.5%	0.80 12.1%	0.92 11.9%	0.42 6.3%	0.42 5.4%

Table 3: Data collected from the system logs, illustrating the interests per user and the effect of the profile modeling scheme. All figures are presented in two time intervals (after 3 and 7 days).

5.3 Recommending relevant documents

This section illustrates the experimental results gathered with respect to the recommendation algorithm.

Table 4 shows statistics, gathered by analyzing the log files, regarding the document recommendations in our system. According to table 4, recommendations range from 9 to 45 documents, on a user basis. Actually, the profile-based top-ranking sentences helped users on deciding whether a document is interesting or not. It was observed that the vast majority of participants decided to visit the suggested documents only when these appeared to be very close to their search needs. This subtle point makes the analysis of results even harder, since there is moderate interaction with suggested links. Finally, according to table 4, an average of 22.7% of all documents recommended was accessed by the users. Document deletion demonstrates a low interest in the retrieved document and the fact that it is kept at relatively low levels (14.6%) is satisfactory.

	Recommended	Accessed	Deleted
Minimum	9	0	0
Maximum	45	24	19
Average	23.30	5.31 (22.7%)	3.42 (14.6%)

Table 4: User statistics on document recommendations illustrating the number of recommended documents and how many of those were accessed or explicitly deleted by the users.

Similarly, table 5 illustrates in detail document recommendations on daily basis. Taking in consideration that at first the system makes no suggestions since it collects data to build each user’s profile, we can observe that there is a steady increase in information access through each user’s personalized home page.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Min	0	0	0	0	0	0	0
Max	0	0	6	5	7	6	7
Avg	0.0 [0.0]	0.0 [0.0]	1.8 [1.6]	2.4 [1.4]	2.5 [1.8]	2.7 [1.8]	3.2 [2.5]

Table 5: Average number of recommended documents accessed on daily basis.

Questionnaire data was also collected with respect to the relevance of document recommendations. As table 6 illustrates, the majority of participants were satisfied from the system’s recommendations. It can be observed, that as they got more familiar with the system, they discovered that their personal page contains even more relevant documents. In fact, the system design suggests that its retrieval performance is directly affected by the overall utilization of the information assistant. Relying on a single search to model a user's interest could confuse the profile management. Therefore, the system becomes significantly better in terms of user profiling and retrieval performance as participants were seeking more information (Wilcoxon test $p=0.003$).

	After 3 days	After 7 days
Complete	3.84 [0.60]	4.05 [0.72]
Relevant	4.21 [0.85]	4.26 [0.66]
Adequate	3.68 [0.58]	4.16 [0.73]

Table 6: Questionnaire analysis on document recommendations with respect to three different factors. All figures are presented in two time intervals (after 3 and 7 days).

5.4 User Preference and comparison with the baseline system

The baseline system, identical to a modern search engine (Google) both in terms of look and feel and retrieval performance, was used to allow users to rank the two systems in order of preference. Table 7 shows the most major points. All questions were performed in 5-point scales with a high score representing a more positive perception in the analysis.

In the direct comparison between the two systems, question “*What was your personal preference between the two systems?*”, all experimental subjects ranked our system above the normal search engine as their personal preference which was further supported by appropriate statistical tests (Wilcoxon Test $p=0.002$).

As it can be observed, the vast majority of participants placed Personal Information Assistant

above the normal search engine in the question “*How effective do you think the system is?*”. This was backed up by significance tests (Wilcoxon Test $p=0.002$) in order to strengthen our original hypothesis.

Finally, all participants considered that Personal Information Assistant was very helpful in discovering additional documents regarding their interest. Although, the difference between PIA and Google in this aspect was quite minor (0.21), statistical tests (Wilcoxon Test $p=0.002$) indicated statistical significance, which constitutes another factor in the evaluation of our system.

Therefore, the results gathered after directly comparing the two systems together further enforce our original hypothesis that Personal Information Assistant is more effective in retrieving and recommending relevant documents. Finally, almost 90% of the participants illustrated the need to make available a system with similar personalization and recommendation features, mainly to assist less experienced searchers to locate information on the web.

	PIA	Google
How effective do you think the system is?	4.58 [0.50]	3.82 [0.71]
How much did the system help you to find relevant documents?	4.05 [0.65]	3.84 [1.08]
What was your personal preference between the two systems?	88.80%	11.20%

Table 7: Questionnaire data from the comparison between Personal Information Assistant and the baseline system.

6. Hypothesis Discussion

The objective of these experiments was to study the effectiveness of PIA from a user’s perspective. The fact that users accessed the system on their own for 7-10 days shows their interest, while it would also be interesting to note that the experimenter was not present when they used the system and completed the questionnaires.

In this section we will discuss what was learned from this long-term experiment and relate the outcomes of our evaluation to our original hypothesis, section 4.4.

1.a Combination of implicit and explicit feedback techniques can model evolving user needs effectively

Taking into account the results obtained and demonstrated in tables 2 and 3, it appears that the system can model user needs in an effective way. According to the results gathered and presented in table 2, the profile manager algorithm can model and adapt to users interests very successfully (4.10 out of 5.00). Although the results between the 3rd and 7th evaluation day didn’t show any statistical significance, we can still observe an increase in user ratings.

Table 3 demonstrates that the profiling algorithm introduced the vast majority of user interests (~95%) compared to only ~5.0% of explicit additions. At the same time, it illustrates that participants amended their interests by an average of 18.4%, which is very satisfactory. These results combined with the data from table 2, prove the first part of our hypothesis.

1.b Users avoid editing their profiles very often

As table 3 illustrates, explicit interest additions and deletions are very low, which suggests that PIA's profile manager works quite effectively. On the other hand, 5.4% of the interests added were explicitly amended in order to become more representative of the user's needs. This figure is very low, considering that the profile generation algorithm relies solely on the previous interactions of the user with the system.

Additional evidence that our hypothesis hold can be provided by table 2 - question 1, where users indicated that they amended their profile quite rarely.

2.a Personal Information Assistant can satisfy user information needs effectively by fetching additional relevant documents

The claim above ascertains to what degree users perceive recommended documents to be relevant, an important factor in any information retrieval system. By investigating the documents recommended by the system, but classified as not relevant by the users, we can gain further insight in this aspect. As table 4 suggests, participants deleted 14.6% of the documents retrieved by the system, while accessing over 22%. This, in fact, indicates that the rest of the documents retrieved were kept, maybe for future access. Finally, table 5 demonstrates a constant increase in information access through each user's personalized home page, which further enhances our original hypothesis.

7. Conclusion

We have designed, deployed and evaluated a system aiming to supply users with up-to-date information regarding their personal needs [Psarras et al; 2006]. By using an implicit information gathering model we eliminate the necessity of having each user create his profile explicitly. Also, the assistant recommends additional documents that might be of interest to the users by formulating queries based on their interests and automatically seek more information on the web. We also presented techniques to keep up with users' changing needs effectively such as our profile management algorithm.

Also, we conducted and demonstrated a task-based, user-centered methodology to be able to investigate the system's strengths and weaknesses in a real environment. Although direct comparison between such systems or between assistants and web search engines is not possible, we introduced a possible way to achieve this effect. Finally, the methodology of this long-term experiment has been presented in detail and the results were analyzed against our retrieval hypothesis.

Acknowledgments

The work reported in this paper is partly funded by the Engineering and Physical Research Council project ADAPT(Ref EP/C004108/1) as well as the IST MIAUCE project (FP6- 033715)

References

- Bates, M.J., The design of browsing and berry picking techniques for the online search interface. *Online Review*, (1989) 13(5):407-424.
- Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1), 71-90.
- Chen, L. and Sycara, K. 1998. WebMate: a personal agent for browsing and searching. In *Proceedings of the Second international Conference on Autonomous Agents* (Minneapolis, Minnesota, United States, May 10 - 13, 1998). K. P. Sycara and M. Wooldridge, Eds.

- AGENTS '98. ACM Press, New York, NY, 132-139.
- Harman, D. 1992. Evaluation issues in information retrieval. *Inf. Process. Manage.* 28, 4 (Aug. 1992), 439-440.
- Konstan, J. A. and Riedl, J. (1999). Research resources for recommender systems. *CHI' 99 Workshop Interacting with Recommender Systems* (1999).
- Jansen, B. J. and Pooch, U. 2001. A review of web searching studies and a framework for future research. *J. Am. Soc. Inf. Sci. Technol.* 52, 3 (Feb. 2001), 235-246.
- Jansen, B. J., Spink, A., and Saracevic, T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.* 36, 2 (Jan. 2000), 207-227.
- Jose, J. M., Furner, J., and Harper, D. J. 1998. Spatial querying for image retrieval: a user-oriented evaluation. In *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, 232-240.
- Kelly, D. and Belkin, N. J. 2001. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 408-409.
- Kuhlthau, C. Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science.* JASIS 42(5): 361-371 (1991)
- Kim, J., Oard, D.W., and Romanik, K., Using implicit feedback for user modeling in internet and intranet searching. University of Maryland CLIS Technical Report 00-01 (2000).
- Lieberman, H. (1995). Letizia: An Agent That Assists Web Browsing. In *1995 International Joint Conference on Artificial Intelligence.* Montreal, CA (1995).
- Martin, I. and Jose, J. M. 2003. A personalised information retrieval tool. In *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM Press, New York, NY, 423-424.
- Oard, D., Kim, J., Implicit Feedback for Recommender Systems. In *Kautz, H., (ed.), Papers from the 1998 AAAI Workshop on Recommender Systems.* Technical Report WS-98-08, The AAAI Press, (1998), 80-82.
- Psarras, I. and Jose, J. A System for Adaptive Information Retrieval. In *Proceedings of Adaptive Hypermedia conference* (2006) 313-317
- Stefani, A. and Strappavara, C., Personalizing Access to Web Sites: The SiteIF Project. In *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia,* HYPERTEXT'98, June 1998.
- White, R., Ruthven, I., and Jose, J. M. 2002. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in information Retrieval* (March 25 - 27, 2002). F. Crestani, M. Girolami, and C. J. Rijsbergen, Eds. Lecture Notes In Computer Science, vol. 2291.
- White, R. W., Jose, J. M., and Ruthven, I. 2003. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Inf. Process. Manage.* 39, 5 (Sep. 2003), 707-733.