Assessing the Effectiveness and Usability of Personalized Internet Search through a Longitudinal Evaluation

Lex van Velsen [1], Florian König [2], Alexandros Paramythis [2]

[1] University of Twente, Dpt. of Technical and Professional Communication
P.O. Box 217, 7500 AE Enschede, the Netherlands
l.s.vanvelsen@utwente.nl

[2] Johannes Kepler University
Institute for Information Processing and Microprocessor Technology (FIM)
Altenbergerstraße 69, A-4040 Linz, Austria
{alpar, koenig}@fim.uni-linz.ac.at

Abstract. This paper discusses a longitudinal user evaluation of Prospector, a personalized Internet meta-search engine capable of personalized re-ranking of search results. Twenty-one participants used Prospector as their primary search engine for 12 days, agreed to have their interaction with the system logged, and completed three questionnaires. The data logs show that the personalization provided by Prospector is successful: participants preferred re-ranked results that appeared higher up. However, the questionnaire results indicated that people would prefer to use Google instead (their search engine of choice). Users would, nevertheless, consider employing a personalized search engine to perform searches with terms that require disambiguation and / or contextualization. We conclude the paper with a discussion on the merit of combining system- and user-centered evaluation for the case of personalized systems.

1 Introduction

Attaining automatically personalized system behavior is, in many cases, a process that can not be considered "complete" at a certain moment in time. On the contrary, personalization often becomes effective only after a certain period of user-system interaction and even after that can be subject to constant changes, as user characteristics and interests may change and expand. In order to explore how these long-term changes affect the (perceived) usefulness of personalized output, longitudinal studies with a partial user focus need to be conducted [1]. However, most evaluations of personalized systems are short term and do not focus on the effects of continued use [2]. Furthermore, most often these evaluations take either a system-centered or a user-centered focus, while a combination of both yields the most valuable evaluation results [3]. System-centered evaluations focus on the quality of system algorithms (to be assessed by means of quality metrics), while user-centered evaluations center on users' subjective experience of their interaction with the system.

This paper discusses a longitudinal evaluation of a personalized search engine, which combined a system- and user-centered approach. The goal of this paper is threefold. First, we want to assess whether personalized Internet meta-search, as provided by Prospector, the system under evaluation, is effective and perceived as useful. The second goal, which is related to the first, is that to determine *why* the system is (regarded as) effective or not, and determine its usability. Third, we want to provide future personalized system evaluators with information that can help them to design their evaluation setup, by reflecting on the experiences we gained in this study.

The rest of this paper is organized as follows. Section 2 describes Prospector, the system which was the subject of this study. Section 3 presents the evaluation setup, followed by the user-centered evaluation results in section 4 and the system-centered ones in section 5. We wrap up this paper with our conclusions in section 6.

2 The Prospector System

Prospector's personalization algorithm is based on the utilization of the Open Directory Project (ODP)[1] ontology, which provides semantic meta-data for classifying search results. Prospector uses taxonomies as overlays [4] over the ODP ontology for modeling user and group interests and bases the re-ranking of search results on said overlays. The operation of Prospector can be summarized as follows: the underlying search engine retrieves results for a user's query; results are classified into thematic topics using the ODP metadata; user- and group- models maintained by the system are used to determine an appropriate ranking of the results; users are presented with the re-ranked results, which they can rate on a per-result-item basis; the system uses these ratings to update individual and group models. User- and group- models are overlays containing the probability of an ODP topic being of interest to a user or group.

The version of Prospector discussed in this paper has been preceded by two other versions, described in [5, 6]. In this paper, we discuss the third version of the system, largely shaped by the results of an evaluation of the second version, reported in [7]. The most important new features in this version include a more stable ranking algorithm, better use of existing meta-data, and usability enhancements. The rest of this section provides a brief overview of interactive aspects of the system and the result re-ranking algorithm (for additional information please refer to [8]).

In order to get personalized search results users first have to register. At the first login they are asked to specify their interest in the 13 top-level ODP topics. It is explained to the users that this way they will benefit from the ratings of results by users with similar interests. Representative sub-topics are also listed for each topic, to help users form a better mental model of the area a topic covers.

For each search users may choose the underlying search engine to use by selecting the corresponding tab (see Fig. 1): Web (i.e., the www.etools.ch meta-search engine), Yahoo and MSN. Google was not included for technical reasons. When issuing a query this engine is accessed, its results are retrieved and classified (per the ODP ontology). The classification paths are displayed for each result, and the tree control on the left side of the results page lets users filter results by these topical categories.

[1] For information on the Open Directory Project please refer to: http://www.dmoz.org

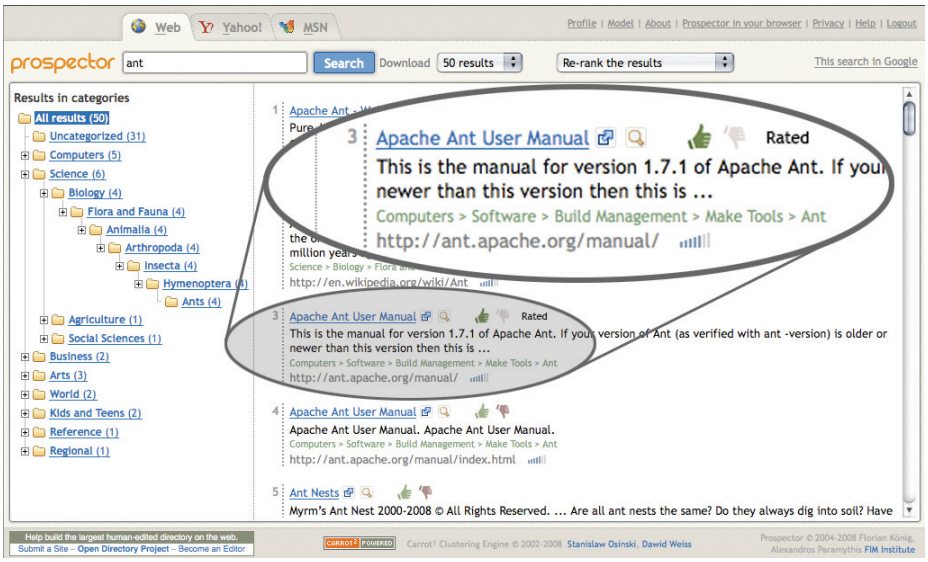

Fig 1. Prospector's main interface.

Re-ranking of results works as follows: The first step is to calculate a relevance probability for each result item, composed from the interest probability of each ODP topic in which the result has been classified. If the user model does not contain an interest probability for a topic, the value is derived from more general topics in the user model, and from the group models. In a second step, the calculated relevance probabilities for each topic are combined into a weighted average. The affinity of the user to the corresponding group is used as the respective weight. In a third step, the relevance probability of each result is combined with its rank as returned by the underlying search engine (both normalized in the value space [0..1]. The normalized rank and score values are then combined by a weighted extension [9] of the *CombSUM* method [10]. Prospector uses this final value for re-ranking the result list accordingly.

By rating individual results positively (thumbs up) or negatively (thumbs down) users implicitly express their preference for certain topics. Quickly evaluating a result is possible by toggling an embedded preview below the result with the magnification glass icon; opening a result in a new window is facilitated by the arrow icon. When previewing or returning from an opened result, the user is notified / reminded of the possibility to rate that result, by pulsating the thumbs a few times.

Each rating modifies the appropriate user- and group- models, thus affecting the calculation of relevance probabilities of classified results in subsequent searches. To give users a way to quickly assess the system-calculated relevance of a result, its probability is visualized next to the URL by means of a "connection quality" indicator, as used in mobile phones. Hovering over the bars with the mouse shows the exact relevance percentage in a tool-tip.

For logged in users the ranking is by default personalized with respect to their user model and the models of associated groups. In addition, users can request that results be re-ranked using a particular group model (e.g., "re-rank for people interested in

arts"). This feature is intended to let users focus on the specific high-level topic represented by the group, and is also available for anonymous, not logged in users.

The user models in Prospector are scrutable [11], allowing users to inspect, correct and fine-tune them, while at the same time strengthening their confidence in the system. Affinities to groups, as set when registering, can be changed at any time. The interests inferred from the ratings can be viewed and changed as well. Finally, entire topic "branches" can also be removed from the model, which gives users the means to purge old or invalid interest information.

3 Evaluation Setup

We asked 130 persons whether they were willing to use Prospector as their primary search engine for 12 days, to have their use with the system logged and, finally, to complete three questionnaires. Because the study could be considered privacy infringing, potential participants were informed they would remain anonymous at all times. Twenty-one persons responded positively, including nineteen men and two women, with an average age of 25.8 years (SD = 2.8). Most were students at the Johannes Kepler University in Linz, Austria. They rated their computer skills as high and used the Internet on a daily basis. All participants used Google as their primary search engine, with one third of them performing more than 15 searches a day, one third 2 to 15 searches a day, and the remaining third just a few searches a week.

Besides logging all actions performed with Prospector (as input data for the system-centered evaluation), we distributed a pre-questionnaire, a mid-questionnaire (after five days of use) and a post-questionnaire (after 12 days of use) as input for the user-centered evaluation. For economy of space we will refer to these questionnaires as "preQ", "midQ" and "postQ" respectively. They addressed the following issues by means of open-ended questions (unless specified otherwise):

1. **Demographics**. In preQ we questioned participants' demographics, internet use, experience with personalized systems and use of search engines.
2. **Expectations**. The preQ asked for the participants' expectations of using Prospector and the expected usefulness of a scrutable user model.
3. **Perceived usefulness**. We asked the participants to score their agreement on three statements (7-point Likert scales) on the usefulness of Google (preQ) and on the usefulness of Prospector (midQ and postQ). The statements were based on the perceived usefulness scale of a search engine by Liaw and Huang [12].
4. **Comparisons between Prospector and Google**. In midQ and postQ, we asked the participants to compare the perceived quality of the results they received from Google and Prospector.
5. **Incidents**. We twice (midQ and postQ) asked the participants to describe incidents that made them satisfied or dissatisfied with Prospector.
6. **User modeling**. The transparent user model allowed us to 'break up' the evaluation of the personalization done by the system in two parts: user modeling and application of the algorithm. This layered evaluation approach makes it possible to pinpoint the cause of incorrectly personalized output more specifically [13]. Therefore, we asked the participants to inspect their model and questioned its clarity (midQ) and correctness (midQ and postQ).

7. **Usability**. In the postQ, we inquired the participants' experience in relation to usability issues of high relevance for personalized systems, as listed by Jameson [14]. Examples include system predictability, controllability and privacy.

4 User-Centered Evaluation Results

In this section we discuss the user-centered results according to the questionnaire elements, listed in section 3.

Expectations. Most participants expected Prospector to outperform Google, by reducing search times (six participants) and giving better results than Google (six participants). As one participant put it: "Hopefully I will quickly find information that in other search engines is only on the second or third page." Twelve participants were initially positive about the possibility to view and alter their model.

Perceived usefulness. The scale we used to measure perceived usefulness appeared to be very reliable (Cronbach's $\alpha = .95$). On a scale from 1 to 7 (where 7 is very useful), Google scored 6.05 (SD = .83). Prospector scored 3.71 (SD = 1.66) in midQ and 3.70 (SD = 1.55) in postQ. There was no significant difference in Prospector's perceived usefulness between midQ and postQ ($t = .12$; $df = 20$; n.s.). To determine whether Google was perceived as more useful after the cold-start problem was overcome, we compared the Google score with the postQ Prospector score. This difference is significant ($t = 6.29$; $df = 20$; $p<.01$): Google was perceived as more useful.

Comparisons between Prospector and Google. Halfway through the study, nine participants preferred Google for searching and one person preferred Prospector. Of particular interest were the answers by six participants that stated their preference depended on the nature of the search task. They liked Google better for searching for simple facts, but thought Prospector had an added value when conducting searches related to their personal interests or study. After 12 days, 19 participants preferred Google. However, several participants apparently did so because Prospector did not offer results in German (the mother tongue of all participants). As one person stated: "I prefer Google, because it provides the possibility to search nationally. With Prospector one doesn't find regional results. The program doesn't like German words."

Incidents. From midQ we could derive two causes that led to dissatisfaction with Prospector: irrelevant search results (mentioned 9 times) and illogical re-ranking of search results (mentioned 6 times). A positive incident that was mentioned more than once regarded Prospector's particular helpfulness when submitting a query containing words with ambiguous meanings. When we asked for these incidents in the postQ the same picture emerged. However, this time more participants mentioned specific searches for which Prospector was useful, like product reviews or scientific articles.

User modeling. When we questioned the participants halfway about the visualization of the user model, 16 participants commented they understood what they were looking at, two 'thought they did', and, finally, three persons stated they did not completely understand what was displayed. Next, we asked whether the user model was a correct representation of their (search) interests. Nine participants stated it was, and six said this was mostly, or for a larger part the case. Three participants answered that they could not judge this as they had not performed enough searches or ratings for a complete user model to be generated. Finally, two participants stated that the user

model was not a good reflection of their (search) interests. In the postQ, we asked the participants to judge the correctness of their user model again. Eleven participants said it was correct, three said it mostly was, and one person said it partly was. This time, four persons said they had not provided Prospector with enough feedback to generate a correct user model and two participants considered their user model incorrect. Unfortunately, the data logs cast doubts over the participants' answers. Even though all participants gave their opinion about the user model, the data logs show that only 11 participants inspected their user model before day seven and only 8 participants inspected it between day seven and twelve (with only 3 users making any changes at all). Therefore, the results regarding user modeling remain inconclusive.

Usability. The usability issues predictability, comprehensibility, unobtrusiveness and breadth of experience received mixed results: half of the participants were positive about these issues and half were not. Other issues received more uniform feedback. When asked about controllability, most participants stated they thought they were fully or for a larger part in control over the system. Privacy was not considered a barrier to using Prospector – 16 persons said the search engine does not infringe on their privacy. The last question addressed system competence. A majority believed that Prospector could deliver the results they desired. Interestingly, six participants commented that the system had the potential to deliver relevant search results, but conditionally (offering as an example the inclusion of results in German).

5 System-Centered Evaluation Results

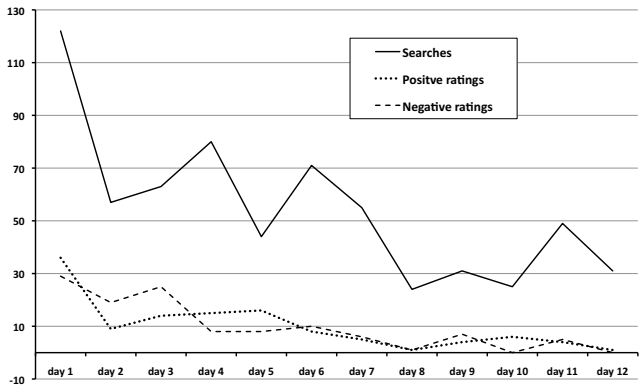

Fig 3. Number of searches, positive and negative ratings

The number of searches, positive and negative ratings over the duration of the evaluation are displayed in Fig. 3. It shows a decrease in all cases. Of note is the fact that there is no significant difference between positive and negative ratings over time.

As a means of determining whether personalization has positively affected the ranking of search results, we examined whether the results participants clicked on were ranked higher than in the original result set. Specifically, for all days, we calculated the distance between the personalized and original ranks of viewed results. This

distance was positive if the result had been promoted, negative if it had been demoted, and 0 if the result had retained its rank after re-ranking.

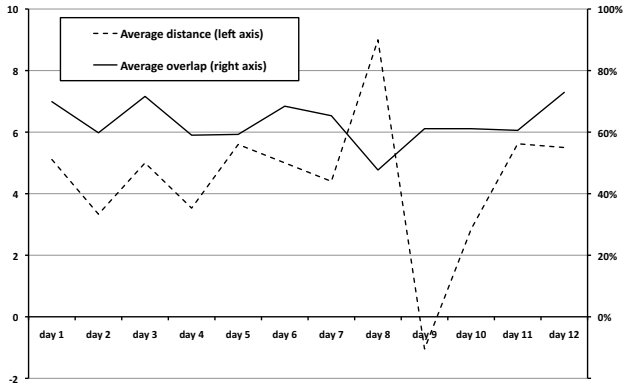

Fig 4. Distances between original and re-ranked results, and percentage of original results still ranked between 1 and 12 after re-ranking ("overlap")

Fig. 4 displays the average distance between these ranks for each day. It shows that for most days, the difference was positive for the personalized rank. The exception is day 9, during which only few searches were performed, while, at the same time, a number of results with a high negative distance were previewed or opened. This combination distorted the overall number for this particular day. For all 12 days, the viewed pages had been promoted by, on average, 4.75 ranks (SD = 11.52). To test the re-ranking effect, we compared the average rank distance for each viewed result to 0. This difference is significant ($t = 9.14$; $df = 490$; $p<.01$): Search results that participants viewed were, on average, placed higher up, due to personalization.

Because participants might tend to consult search results ranked highly, regardless of their relevance, we examined whether the first 12 results contained a disproportionately high percentage of items brought there by Prospector. We chose 12, as on a average-sized screen a user would see 6 results in one screen-full and most people do not look beyond the first 10 [15] – we rounded that number up to two full screen. Fig. 4 displays the daily average percentage of results among the first 12 that were originally there. Over 12 days, the mean percentage is 65.10%. This implies that users had a choice between (interspersed) original or re-ranked results, but chose the latter on purpose and not because they were conveniently placed at the top of the list.

In addition to these analyses, the two metrics "Rank scoring" [16] and "Average Rank" [17] were employed. Rank scoring shows how close to the optimal ranking a set of search results is, whereby 'optimal' denotes a situation in which all the consulted results appear at the top of the list. In this metric, the importance of the ranks decreases exponentially (e.g., a correct rank 1 creates a better score than a correct rank 10). We performed a paired samples t-test between the original rank score average (M = 5.05, SD = .59) and the personalized rank score average (M = 6.75, SD = 1.19). The averages were calculated from the rank score values of the 12 days. This difference is significant ($t = -6.92$; $df = 11$; $p<.01$): Personalized rank scores were higher than the original ones (see Fig. 5).

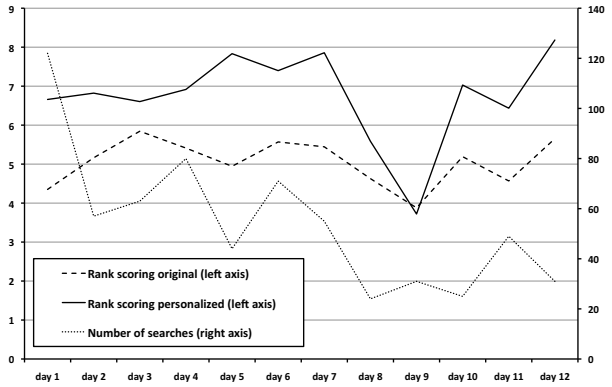

Fig 5. Rank scoring of the personalized and original ranks of viewed results

The average rank measure was calculated for the original and the personalized ranking of consulted results on a per day basis (see Fig. 6). The personalized results had a lower average rank in all cases, except on day 9. A lower average rank means that the consulted results appeared higher up in the result list. The significance of the difference between the average ranks can be derived from the significance of the average distance measure described above.

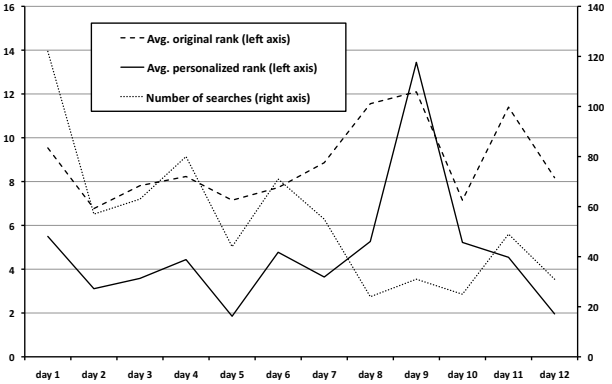

Fig 6. Average rank of the personalized and original ranks of viewed results

6 Conclusions

In this paper we have presented a longitudinal evaluation of the Prospector system, which combined a system- and a user-centered focus. The data that resulted from the system-centered evaluation has shown that Prospector effectively promotes items with relevant informational value in a list of search results. However, user perceptions on system usefulness were not in favor of Prospector: the participants thought their primary search engine (Google) was more useful. Their comments led us to think that this opinion was partly due to missing features popularized by Google (specifically,

localized search, spelling suggestion). Prospector offered a different interface with different features which may have biased the participants' perception of its usefulness, regardless of the system's actual value for searching, as people highly value the appearance and features of their primary search engine [15]. A way to design for this implication is to replicate the features and appearance expected of a search engine (e.g., by incorporating the above facilities, and adopting a design similar to Google's).

The participants had very high expectations of Prospector, based on the quality of search results returned by Google, and expected Prospector to outperform Google. Some users explicitly anticipated that the result they needed be listed first or second. These expectations are apparently hard to meet, especially as users will want to see an added value fast and it may take some time for a personalized search engine to deliver top-quality results. When evaluating a personalized system, one has to make sure that participants have a sound mental model of the system so that they can form reasonable expectations. They must understand whether and how much time and effort are required for the system to 'get to know' the user. This issue is not only limited to evaluation of course: for a system to have a fair chance of user acceptance, it must ensure that its users' expectations are realistic from the very beginning.

The evaluation has suggested some circumstances in which personalized search might be more rewarding for users. These are the searches which our participants described as 'personal', or searches without a clear-cut answer. Typical for these searches are, as Marchionini terms it, relatively low answer specificity, high volume and high timeliness [19]: the answer to the search is not easily recognized as being correct (e.g., a suitable hotel in Amsterdam), it has to be found in a large body of information and, finally, the user has to invest some time in finding the right answer. Dou et al. [17] found navigational queries with low click entropy (i.e., most users chose the same result) to be less ambiguous and not suitable for personalization. Teevan et al. [18] proposed methods to detect such queries and predict the usefulness of personalization on the basis of query properties, result quality and search history.

This evaluation reinforces the notion that the application of a dual approach is instrumental in fully understanding a personalized system ([3]): If we had relied on the system-centered approach only, the results would have had a too positive skew, while the results derived from the user-centered approach alone would have put into question the system's effectiveness. In other words, by applying a double focus, one can acquire a more complete view of system usefulness (albeit, potentially with contradicting evidence). Furthermore, in certain cases such an approach makes it possible to cross-validate and ground or disprove findings (e.g., with the user model viewing behavior in this study). Furthermore, the dual focus provides us with the option to not only determine Prospector's effectiveness and chances of acceptance in a real-world setting, it also resulted in redesign input that enables us to further improve the system (e.g., by incorporating spelling suggestions).

In our longitudinal evaluation we have experienced a reduction in user activity as time progressed. This might point to need, for people running similar studies, to actively encourage participants to continue using the system under investigation (e.g., through reminders, or by providing some form of incentive). Last but not least, the application of a longitudinal evaluation setup has yielded insights which, we believe, may not have been attainable otherwise: we have been able to determine whether Prospector "works", but also what users see as the main drawbacks of the system.

Acknowledgements. The work reported in this paper has been supported in part by: (a) the EU-funded "Adaptive Learning Spaces" (ALS) project (229714-CP-1-2006-1-NL-MPP); and (b) the "Adaptive Support for Collaborative E-Learning" (ASCOLLA) project, supported by the Austrian Science Fund (FWF; project number P20260-N15).

References

1. McGrenere, J., Baecker, R.M., Booth, K.S.: A field evaluation of an adaptable two-interface design for feature-rich software. ACM transactions on computer-human interaction 14 (2007) article 3
2. Van Velsen, L., Van der Geest, T., Klaassen, R., Steehouder, M.: User-centered evaluation of adaptive and adaptable systems: a literature review. The knowledge engineering review 23 (2008) 261-281
3. Díaz, A., García, A., Gervás, P.: User-centred versus system-centred evaluation of a personalization system. Information Processing and management 4 (2008) 1293-1307
4. Brusilovsky, P., Millán, E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web. Springer, Heidelberg (2007) 3-53
5. Scholtz, J., Morse, E., Potts Steves, M.: Evaluation metrics and methodologies for user-centered evaluation of intelligent systems. Interacting with computers 18 (2006) 1186-1214
6. Paramythis, A., König, F., Schwendtner, C., Van Velsen, L.: Using thematic ontologies for user- and group-based adaptive personalization in web searching. Adaptive multimedia retrieval, Berlin, Germany (2008)
7. Van Velsen, L., Paramythis, A., Van der Geest, T.: User-centered formative evaluation of a personalized internet meta-search engine. (In review)
8. König, F., Van Velsen, L., Paramythis, A.: Finding my needle in the haystack: effective personalized re-ranking of search results in Prospector. (in review)
9. Renda, M.E., Umberto, S.: Web metasearch: rank vs. score based rank aggregation methods. The ACM symposium on applied computing, Melbourne, Florida (2003)
10. Shaw, J., Fox, E.: Combination of Multiple Searches. Text REtrieval Conference, Gaithersburg, MD (1993)
11. Kay, J.: Stereotypes, student models and scrutability. In: Gauthier, G., Frasson, C., VanLehn, K. (eds.): ITS 2000. Springer-Verlag, Berlin (2000) 19-30
12. Liaw, S.S., Huang, H.M.: An investigation of user attitudes toward search engines as an information retrieval tool. Computers in human behavior 19 (2003) 751-765
13. Paramythis, A., Weibelzahl, S.: A decomposition model for the layered evaluation of interactive adaptive sysems. In: Ardissono, L., Brna, P., Mitrovic, A. (eds.): User Modeling 2005. Springer Verlag, Heidelberg (2005) 438-442
14. Jameson, A.: Adaptive interfaces and agents. In: Jacko, J.A., Sears, A. (eds.): Human-computer interaction handbook (2nd ed.). Erlbaum, Mahwah, NJ (2007) 433-458
15. Keane, M.T., O'Brien, M., Smyth, B.: Are people biased in their use of search engines? Communications of the ACM 51 (2008) 49-52
16. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. Microsoft Research, Redmond, WA (1998)
17. Dou, Z., Song, R., Wen, J.: A large-scale evaluation and analysis of personalized search strategies. WWW, Banff, Canada (2007)
18. Teevan, J., Dumais, S.T., Liebling, D.J.: To personalize or not to personalize: modeling queries with variation in user intent. SIGIR'08, Singapore (2008)
19. Marchionini, G.: Information seeking in electronic environments. Cambridge university press, New York (1995)