Layered Evaluation of Adaptive Search

Peter Brusilovsky
University of Pittsburgh
School of Information Sciences
Pittsburgh, PA, 15260
peterb@pitt.edu

Rosta Farzan
University of Pittsburgh
Intelligent Systems Program
Pittsburgh, PA, 15260
rosta@cs.pitt.edu

Jae-wook Ahn
University of Pittsburgh
School of Information Sciences
Pittsburgh, PA, 15260
jaa38@pitt.edu

ABSTRACT
The goal of this paper is to discuss how adaptive search systems which embed exploratory options should be evaluated. We argue that a state-of-the art evaluation of adaptive search systems should follow a "layered evaluation" approach. To support and explain this argument we describe how the layered approach was applied to the evaluation of the adaptive search component of Knowledge Sea II, a system that is powered by a social navigation support mechanism.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness)

General Terms
Measurement, Design, Human Factors

Keywords
Social search, adaptive systems, exploratory search systems, layered evaluation

1. INTRODUCTION
The growing need for effective organization and maintenance of the increasing number of Web-based educational resources motivated us to construct a personalized information access system, Knowledge Sea II (KSII). KSII provides various types of information access methods, including two-level visualizations (a knowledge map plus a similarity-based visualization), hypertext browsing, recommendation, and social search. Personalization for all these access methods is provided by social navigation (SN) support [1], [5]. SN is a relatively well-known personalization approach for browsing-based and recommendation-based information access; however, its use for search personalization has been almost unexplored.

The adaptive search component of KSII combines a traditional vector search engine with SN support, allowing every user to benefit from the collective wisdom of the whole community. To stress it we will refer to it as "social search." The results of the search are adapted to the user by taking into account both the past interactions of the individual user and the user's group. The SN support of KSII includes various information access methods that allow the user to do exploratory searching. She can start the exploration by browsing or by entering the map, then use her newly acquired knowledge about the domain's terminology to choose better query terms. She can also modify her initial query after consulting SN information provided by the system. The main goal of this paper is to discuss how adaptive search systems with this exploratory nature should be evaluated, using KSII search as a model. We argue that state-of-the art evaluation of adaptive search systems should follow a "layered evaluation" approach that is an active focus of research in the area of user-adaptive systems [2]. The core idea behind layered evaluation is that specific sub-components or layers of any user-adaptive system should be understood and evaluated independently. Layered evaluation can overcome shortcomings of the conventional methodologies, which try to test the adaptation process as a whole and can miss success or failure of critical sub-components. In our approach to layered evaluation, we divided the adaptation process into two parts: decision-making and interface adaptation and then evaluated each of them. In this paper, the nature of our adaptive social search system is presented (section 2) and our layered evaluation framework is discussed (section 3). The paper concludes in section 4 with a summary and brief discussion of the future direction of our research.

SIGIR'06, August 6–11, 2006, Seattle, WA, USA.

2. SOCIAL SEARCH IN KSII
Social navigation (SN) in KSII incorporates several information access methods, including social search. SN support is offered through by visually marking links with icons and color codes. Figure 1 shows an example of search results that have been annotated with SN cues. Standard information about each document in the list is given—such as rank (7), document source (Univ. of Leicester), title (Pointers), and a similarity score (0.4057) —while traffic- and annotation-based SN cues are on the right. The foreground and background colors of the human icon depict user and group traffic, associated with time spent reading this document [4]. The darker the color is, the higher the traffic. The background color of the annotation represents annotation density. The foreground icons represent the type and overall status of the annotation. For example, a "thumbs-up" icon represents positive individual annotation while the warm temperature shown on the "thermometer" represents positive group annotation. For example, the document "Pointers" shown on Figure 1 is ranked 8th in terms of its similarity score to the user query but is very popular among the community of the users. Thus the user might want to examine the contents of this

document, despite its relatively low score, to learn how to improve her query terms for the next stage of her exploration.

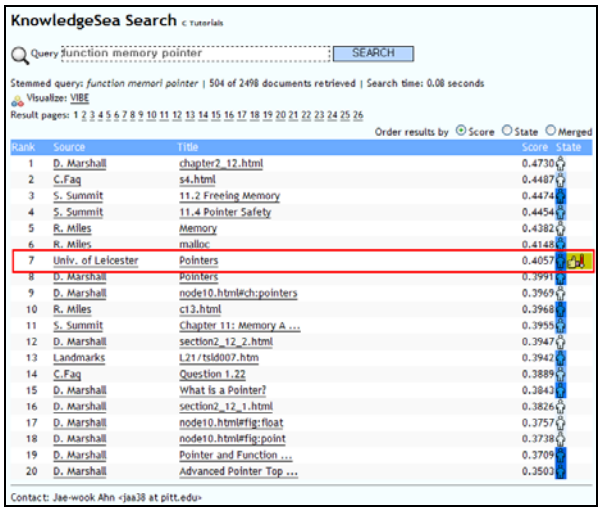

Figure 1 Social search with social navigation cues

3. LAYERED EVALUATION OF SOCIAL SEARCH

The need for the layered evaluation framework arose from the insight that conventional evaluation methods cannot pinpoint the effectiveness of critical layers of the adaptation process, which perform different tasks contributing to the final results. Current practices frequently attempt to evaluate adaptation as a whole by comparing the whole adaptive application to a baseline, an equivalent, non-adaptive application. However, even if the results turned out to be better than the baseline's, we cannot hastily conclude that all of its components perform well. Vice versa, if the adaptive system as a whole is lower than the baseline's, there is still the possibility that one of its layers was actually successful [3]. To address this problem, several authors have introduced layered evaluation frameworks. Brusilovsky et al defined user-modeling and adaptation evaluation layers in [2]. Weibelzahl introduced a 4-layer approach: the reliability and validity of input data, interface, adaptation decision, and the interaction [6].

To evaluate social search in KSII we adopted a 2-layer approach which considers the decision-making and adaptation layers separately. Based on the interaction history of the user's social group, the decision-making layer decides which pages should be useful and to what extent. The adaptation layer decides how to express to each user this calculation of the social importance of a specific page. In the current version of KSII, this layer generates icon-based annotations, as shown in Figure 1. However, this is only one possible way to express the social importance of documents.

3.1 Decision Making Layer

The goal of the decision-making layer is to predict how useful each document is to a user of a specific group. KSII uses two independent decision-making layers, based on traffic and annotation. Since the latter is rather straightforward, we focused on evaluating the traffic-based one. To argue that the traffic-based prediction works we needed to demonstrate that documents predicted as useful (those shown with darker blue backgrounds by the adaptation layer) are really useful. Our gold standard for rating the importance of pages is that students find them good and important. Therefore, we focused on pages with student annotation.

For evaluation, we computed the normalized access rate for pages with and without annotation. As can be seen in Figure 2, "good and important" pages are accessed twice as often. Thus page traffic average is a good indicator of page quality. These pages will have a generally darker background, according to our traffic-based SN support algorithm.

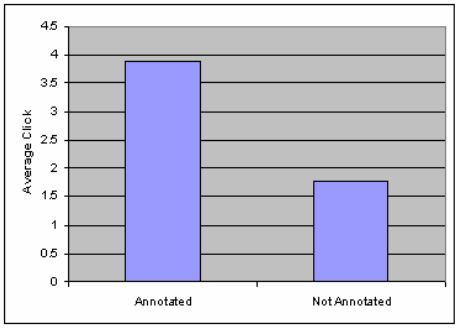

Figure 2 - Average click number over pages with and without annotations

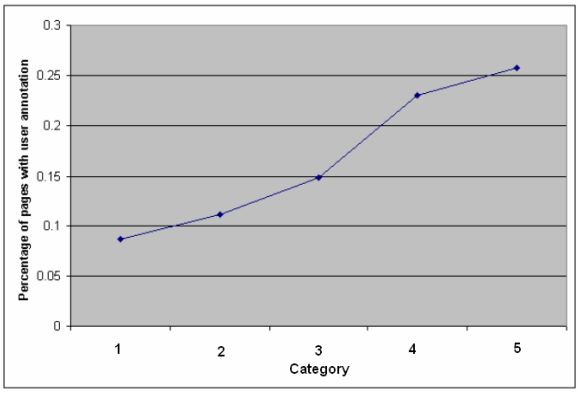

Figure 3 - Percentage of pages with user annotation for different levels of usage

To enhance the evaluation, we categorized accessed documents in five categories, based on the time spent on each page. The following table shows the details of this classification.

Category	Average Time Spent	Darkness of Background	Level of Recommendation
1	< 65 sec	No background	None
2	< 97 sec	Light blue	Slightly
3	< 152 sec	Blue	Recommended
4	< 217 sec	Dark blue	Considerably
5	> 282 sec	Very dark blue	Highly

For each category we computed the percentage of pages that were annotated by the students. To exclude the dependency of annotation and visit, we excluded annotations made by users of the target semesters while including annotations made by users of past and future semesters. As shown in figure 3, the pages with

darker backgrounds (higher usage) have a higher percentage of annotation. This data shows that important pages are being predicted as useful by our SN adaptation which means the important pages are augmented with darker background.

3.2 Evaluation of social search with social navigation cues

Once we established the positive correlation between quality and SN, it was important to evaluate the effect of SN cues. The goal of the cues is to attract user attention to socially important documents and to encourage them to examine them. In our context, we needed to evaluate how much the SN cues affect students' decision to choose links within search results. Moreover, since KSII social search separates the visualization of query relevance (document position in the search list) from visualization of social importance (intensity of background color in SN cues), we were interested in comparing the influence of positive SN cues to the influence of being a top ranking in the list.

To evaluate this layer, we decided to compare the *effective* and *random relative access rates* for links with high rankings (top of the list) and links with traffic-based cues. The *random relative access rate* tells which fraction of clicks would have been made if the user randomly selected specific links in the search results list. Basically, it shows how often the links with this property appear in the search results list. The *effective relative access rate* reports the actual proportion of target quality links, compared to total accessed links. If the effective relative access rate is higher than random, it means that the links with this quality successfully encourage users to access them.

The first question to answer is: "Do students prefer links with better rankings?" (considering the first three documents in the search results list to be *top ranked*). Since every results page shows 20 links, the random relative access rate for the top three ranked documents is 3/20 = 0.15. Effectively, students accessed 53 documents from different search results lists, with 16 being top ranked. Therefore the effective relative access rate was 16/53 = 0.3, which is twice the random (0.15). This is evidence that the students do take the document rank into account, preferring links on the top of the list.

The second question to answer is: "Do students prefer links with traffic-based SN cues?" To answer this question, we attempted to separately evaluate links with any visible past traffic (number of past clicks >1) from links with higher traffic (past clicks >2). The reason is that the links with two past click were annotated with a very light blue color, which, we afraid, some users might ignore. The links with 3 and more past clicks were annotated with reasonably dark blue color and were hard to ignore.

Computing the random relative access rate for links with group traffic was a complicated procedure. For each of the 53 cases of link access we had to re-create the group traffic accumulated at the time of access to understand how many social-cued links the user saw when making the selection. For each case, we calculated this rate by dividing the number of visible links with the target level of traffic by the total number of links. Then, we averaged the probabilities over all 53 cases and found that for pages with visible traffic the random relative access rate is equal to 0.08. Out of 53 cases, students choose 17 documents from the visible traffic category. Therefore the effective relative access rate for links with visible traffic is 17/53=0.32, which is four times higher than the random access rate (0.08). A similar ratio (0.05 to 0.19) was obtained for links with high traffic. This result shows that students do prefer links with visible group traffic. Moreover, the ratio of effective access rate to random is twice as high for pages with visible traffic than for pages with top rankings. This provides evidence that pages marked by visible group traffic do influence students. Moreover, the presence of "group traffic" gives the page an even higher chance to be visited.

4. CONCLUSIONS

In this study, we demonstrated how a 2-layered evaluation framework could be used for evaluating an adaptive search interface which enables exploratory searching by users. We divided the evaluation process into decision-making and adaptation layers, in order to better understand the effectiveness of each sub-component process. We were able to show a correlation between the predicted and effective social utility of a page (i.e., pages automatically predicted as important for the group by the decision-making component were actually rated as important by students). We also provided evidence that the specific interface adaptation approach used in KSII to attract the user's attention to socially important pages does influence user behavior in the expected direction. The proposed evaluation framework should be able to evolve by adopting more layers, such as user-to-system interaction and input data validation. In future research, we are planning to use the same layered framework to evaluate other kinds of adaptive information access methods, including information visualization.

5. REFERENCES

[1] Brusilovsky, P., Chavan, G., and Farzan, R (2004). Social adaptive navigation support for open corpus electronic textbooks. In Proc. of Adaptive Hypermedia and Adaptive Web-Based Systems.

[2] Brusilovsky, P., Karagiannidis, C., and Sampson, D. Layered evaluation of adaptive learning systems. *Int. J Cont. Engineering Education and Lifelong Learning, 14, 4*, 2004, 402-420.

[3] Brusilovsky, P., Karagiannidis, C., and Sampson, D. (2001). The Benefits of Layered Evaluation of Adaptive Applications and Services. In *Proceedings of Workshop on Empirical Evaluation of Adaptive Systems*.

[4] Farzan, R. and Brusilovsky, P. (2005) Social navigation support in E-Learning: What are real footprints? In Proceedings of IJCAI'05 Workshop on Intelligent Techniques for Web Personalization.

[5] Farzan, R. and Brusilovsky, P. (2005) Social navigation support through annotation-based group modeling. In Proc of User Modeling.

[6] Weibelzahl, S., and Weber, G (2001). Advantages, Opportunities, and Limits of Empirical Evaluations: Evaluating Adaptive Systems. *Kunstliche Intelligenz.*

[7] Lowrence, S., and Giles, C.L. Context and Page Analysis for Improved Web Search (1998). *IEEE Internet Computing.*