Improving controllability and predictability of an interactive user model driven search interface

Antti Kangasrääsiö [◇] **Dorota Głowacka** [♣] **Samuel Kaski** [◇♣]

[◇] Helsinki Institute for Information Technology HIIT
Department of Information and Computer Science
Aalto University
[♣] Helsinki Institute for Information Technology HIIT
Department of Computer Science
University of Helsinki
first.last@hiit.fi

Abstract

In this extended abstract, we discuss user controllability and system predictability in the context of exploratory search. When a user is directing a search engine by interactively refining a probabilistic user model using uncertain relevance feedback, usability problems regarding controllability and predictability may arise. By interpreting user's actions as setting a goal for an optimization problem instead of passive relevance feedback, and by allowing the user to see the predicted effects of an action before committing to it, these problems can be reduced. A preliminary user study with an exploratory search interface indicates small improvements in user acceptance, perceived usefulness and task performance.

1 Introduction

In exploratory search, the user searches for information in a domain that she is not initially familiar with. Because of this, search interfaces assisting the user in exploratory search are faced with a difficult problem: how to help the user direct the exploratory search with uncertain feedback. If the user were an expert and the feedback certain, it could be interpreted in a purely exploitative manner. However, as in exploratory search, this is not the case, there needs to be a suitable amount of exploration mixed in to help the user with the search task.

Probabilistic user models can be used to handle the exploration/exploitation trade-off, but there can be usability problems if they are implemented in a naïve way, where the user feedback is interpreted simply as datapoints to fit a model. As the user is not a passive function that is sampled by the system, but an active entity that is trying to steer the system through iteratively improving the model, there needs to be a layer of interpretation between the user and the model. This layer is responsible for translating the user feedback into requirements for the state of the user model and predicting the effects different user actions may have on the model.

Next, we highlight some related research, discuss our solutions to the usability problems in more detail and describe the results of a short user study.

2 Related Work

Traditionally, research in information retrieval systems focuses on improving the predictive accuracy of recommendations mostly by developing new algorithms. Recent studies [8, 11] have shown that visual features and enhanced interaction can greatly improve the user engagement in the search process and consequently the performance of the retrieval system. Some of the interface characteristics studied are transparency, explainability, predictability and controllability, of which user controllability is the least explored aspect of information retrieval systems.

Some of the initial solutions include result clustering [3], relevance feedback [6], query suggestion [1], and faceted search [13]. However, the proposed techniques are rarely used in practice due to the high cognitive load of going through a large list of suggestions or providing feedback for a large number of items [6].

There have been also numerous attempts to engage the user into the feedback loop through interactive visualizations combined with learning algorithms to support users to comprehend the search results [2], and visualization and summaries of results [7]. These solutions give users more control, however, they do not adapt to the moment-by-moment information-needs of the user [10]. Recently, reinforcement learning (RL) techniques have been applied to facilitate exploratory search [4, 5, 9]. The RL-based systems prevent the user from getting trapped in a local context and facilitate exposure to a large area of the information space, however, they do not allow the user to predict the effects of their actions.

3 Proposed Approaches

3.1 User Feedback as a Goal for an Optimization Problem

To allow the user feedback to have the intended effects, the user feedback values should be interpreted as goals for optimization problems regarding the next state of the system, instead of just additional datapoints. This way the user has an automatic assistant that steers the system towards the desired target indicated by the user. In this modelling approach the user is assumed to be an active entity trying to steer the system, instead of a passive entity that is sampled by the system, as is usual in reinforcement learning approaches. The general idea for this approach is illustrated in Algorithm 1.

```
**Input**: user model $\mathbf{M}_\mathrm{t}$, user feedback $\mathbf{f}$, model update rule $u(\cdot)$, optimality criterion to be
        minimized $o(\cdot)$
**Output**: new user model $\mathbf{M}_\mathrm{t+1}$, optimal feedback $\mathbf{f'}$
$\mathbf{f'} \leftarrow \underset{\mathbf{f'}}{\mathrm{argmin}}\, o(u(\mathbf{M}_\mathrm{t}, \mathbf{f'}), \mathbf{f})$
$\mathbf{M}_\mathrm{t+1} \leftarrow u(\mathbf{M}_\mathrm{t}, \mathbf{f'})$
```

Algorithm 1: Optimal way to update user model based on user feedback. $\mathbf{M}_\mathrm{t}$ is (the data structure representing) the user model at time t. **f** is the feedback, or a series of feedbacks, given by the user. **f'** is the optimal feedback, or series of feedbacks, leading to the optimal model. The length of **f** and **f'** may differ, i.e. even though the user only gives a single feedback value ($\mathbf{f} = \mathrm{f}_1$), the algorithm may use multiple values ($\mathbf{f'} = [\mathrm{f'}_1, .., \mathrm{f'}_n]$) to take the model to the optimum state. Model update rule $u(\cdot)$ takes a model and one or more feedback values and returns an updated model. Optimality criterion $o(\cdot)$ takes a model and a user-given feedback value and returns the model error from the optimum.

Solving this problem requires making two design choices: what optimality criterion to use, and what algorithm to use for calculating the optimal feedback values. In our implemented solution, the optimality criterion is that the parts of the model that the user has explicitly given feedback on should be changed so that the resulting model agrees with the feedback. For example, if the user indicates that a certain keyword has relevance X to her search intent, then the optimal value for the relevance of that keyword in the resulting model is X. In our implemented solution, **f'** is a finite list of relevance feedback values for the keyword chosen by the user.

3.2 Predicting Effects of Feedback Actions

When the user is giving feedback, it is not obvious that she will be able to predict the effects these actions will have on the system, given that the behavior of the system relies on a complex underlying model. If making the model simpler is not possible without sacrificing performance, one way to solve this problem is to enable the user to see a prediction of the effects of any action available to her before she commits to it. This way the user will be less surprised by the effects of the action and is able to choose the action based on the expected consequences.

However, there are some practical problems with accurately predicting the effects of different actions. For example, the system may be so complex that accurately simulating and visualizing the effects for any possible action is infeasible in a real-time fashion. Further problems may arise if the system has randomized elements in it, for example, in order to support exploration, because in this situation the amount of possible future states may be practically infinite.

It is thus more practical to use approximate prediction. This prediction should allow the user to get a prediction of the probable effects, while still being feasible to construct in a timely fashion. Our approach for constructing this approximate prediction is to sample the possible future states of the system, fit a simpler function approximation to these points and use it for visualizing the possible effects.

Constructing this approximation requires making two main choices: which points to sample and what function family to use for constructing the approximation based on these samples. In our implemented solution we sample the extreme relevance values for the keyword the user wishes to give feedback on and use a linear interpolation between them.

4 System Architecture

We implemented a version of our proposed improvements to the SciNet scientific article search engine [4]. This system allows the user to direct the search by interacting with a visualization of the user search intent model, composed of keywords and their estimated relevances. The model is visualized to the user using the Intent Radar, where top 10 relevant keywords are shown on a circular layout so that the closer to the center a keyword is, the more relevant it is. Additional keyword suggestions are also visualized on the edge of the radar view. An example of the visualized user model is shown in Figure 1, along with a scenario illustrating one of the problems with the baseline system. An example of the improved system is shown in Figure 2.

The user gives feedback on the user model by moving one keyword at a time to a new location in the radar. When the user lets go of the keyword, feedback for this keyword is calculated based on the distance from the center, the model is updated, new articles are retrieved and this new state is then visualized to the user.

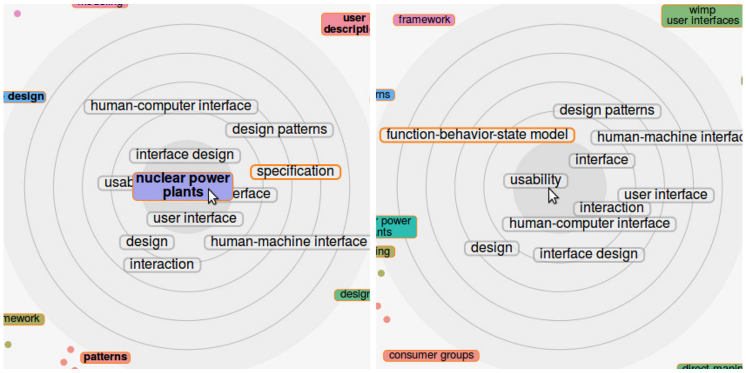

Figure 1: In the baseline system, sometimes giving maximal relevance feedback to a keyword by dragging it to the center of the Intent Radar (left) may result in a new model that does not contain this keyword within the top 10 keywords (right). The central keywords represent the top 10 keywords.

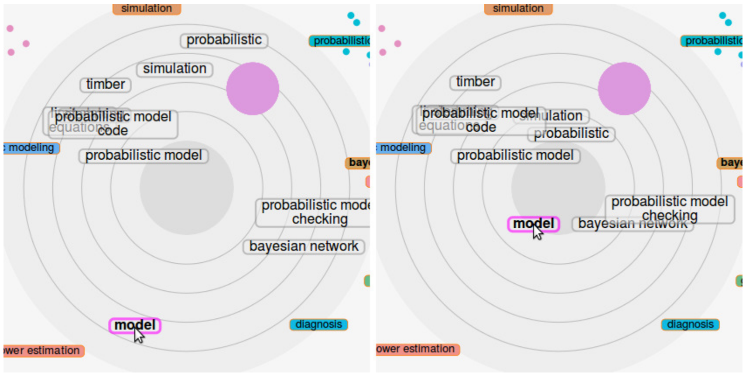

Figure 2: In the improved system, while the user is dragging a keyword over the Intent Radar, the locations of (other) central keywords move simultaneously according to the prediction. The purple dot indicates the original location of the keyword, and the user can move the keyword back there to cancel the prediction for the current keyword if she wants to try out another one.

5 Evaluation

To study the effects of these improvements, we conducted a short user study on 12 university students and staff members. Each user performed two search tasks, one using the baseline system and another with the improved system. One task was a focused exploratory task and the other a broad exploratory task. Both tasks were about fact retrieval regarding topics the users were not very familiar with. Familiarity in the topic was rated on 1 to 5 Likert scale and all the users reported familiarity less than 5, with the average rating of 2.0. In the broad task, the questions had multiple correct answers, whereas in the focused task the question scopes were more narrow. The study was balanced with respect to the combination of the type of interface, task and order. After both tasks a short semi-structured interview was conducted.

The answers to the task questions were rated in a double-blind manner by an expert in both fields. The grading was done on a 1 to 5 Likert scale per question, where 5 corresponded to an excellent answer and 1 to a completely wrong answer. One third of the answers were rated separately by another expert. The average inter-rater reliability based on Spearman rho was 0.75, which can be considered adequate. Also the keywords and articles viewed by the users were rated for relevancy by an expert. P-values for the results were calculated using the two-sided Wilcoxon rank-sum algorithm. The interviews were analyzed using content analysis [12].

6 Results

Based on the results we deemed two users as outliers because on one task they received lowest possible points on task performance and over 80 % of the articles they viewed were rated irrelevant. It was likely that these tasks were too difficult for these users. These two users were excluded from the analysis.

The improved system resulted in better performance in the focused exploratory task (3.1 for improved, 2.2 for baseline) and worse performance in the broad exploratory task (3.0 for improved, 3.8 for baseline), but these differences were not statistically significant ($p = 0.2$ and $p = 0.1$ respectively). However, the difference of task performances for the baseline between the two tasks was statistically significant ($p = 0.01$). This indicates that the baseline is more efficient in broad than focused exploratory tasks. For the improved system this difference did not exist ($p = 0.6$).

In the interviews, 7 out of 10 users reported that they felt that the visualized prediction helped them in the task. The main stated reason for this was that it helped them predict the effects of their actions, but it was also felt useful for illustrating which visualized keywords were related to each other, as dragging a keyword over the Intent Radar would often cause similar keywords to move in a similar fashion.

Based on the interviews, the majority of the users stated that they preferred the improved interface over the baseline. Five users preferred the improved interface overall, 2 users had mixed preferences depending on the situation, 1 user preferred the baseline overall and 2 users indicated no explicit preference.

7 Discussion and Future Work

In this extended abstract, we described a problem with the usability of systems relying on probabilistic user models that are refined by the user in an iterative manner. We have proposed a solution for improving the usability of this kind of systems and shown some initial results indicating small improvements in user acceptance, perceived usefulness and task performance. Although the results did not indicate statistically significant improvements on the baseline system, with a larger user study it would be possible to find out whether the effects are real.

Acknowledgments

This work has been partly supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170) and TEKES (Re:Know). The research leading to this results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 611570. Certain data included herein are derived from the Web of Science prepared by THOMSON REUTERS, Inc., Philadelphia, Pennsylvania, USA: Copyright THOMSON REUTERS, 2011. All rights reserved. Data is also included from the Digital Libraries of the ACM, IEEE, and Springer.

References

[1] P. Anick. Using terminological feedback for web search refinement: A log-based study. In *Proc. of SIGIR*, pages 88–95. ACM, 2003.

[2] D. H. Chau, A. Kittur, J. I. Hong, and C. Faloutsos. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proc. of CHI*, pages 167–176. ACM, 2011.

[3] D. R. Cutting, D. R Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proc. of SIGIR*, pages 318–329. ACM, 1992.

[4] D. Głowacka, T. Ruotsalo, K. Konyushkova, K. Athukorala, S. Kaski, and G. Jacucci. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *Proc. of IUI*, pages 117–128. ACM, 2013.

[5] M. Karimzadehgan and C. Zhai. Exploration exploitation tradeoff in interactive relevance feedback. In *Proc. of CIKM*, pages 1397–1400. ACM, 2010.

[6] D. Kelly and X. Fu. Elicitation of term relevance feedback: An investigation of term source and context. In *Proc. of SIGIR*, pages 453–460. ACM, 2006.

[7] B. Kules, M. Wilson, M. C. Schraefel, and B. Shneiderman. From keyword search to exploration: How result visualization aids discovery on the web. Technical report, 2008.

[8] D. Parra, P. Brusilovsky, and C. Trattner. See what you want to see: Visual user-driven approach for hybrid recommendation. In *Proc. of IUI*, pages 235–240. ACM, 2014.

[9] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proc. of ICML*, pages 784–791. ACM, 2008.

[10] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proc. of CHI*, pages 415–422. ACM, 2004.

[11] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proc. of IUI*, pages 351–362. ACM, 2013.

[12] R. Weber. *Basic content analysis*. Sage, 1990.

[13] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proc. of CHI*, pages 401–408. ACM, 2003.