Interactive Modeling of Concept Drift and Errors in Relevance Feedback

Antti Kangasrääsiö
Yi Chen
Dorota Głowacka
Samuel Kaski
first.last@hiit.fi

Problem Setting

Exploratory search
- User has to learn while searching
- Iterative reformulation of search query

Search system SciNet (IUI 2013, 2015) allows interactive formulation of the search query through relevance feedback on a visualized user model.

However, the user model makes many implicit assumptions
- All user feedback is assumed equally accurate
- The user is assumed to make no mistakes in giving feedback
- No learning or change in search interests is assumed to occur

Our Contribution

Hypothesis

It might be useful for the user that she is notified if some of her feedback is identified to be inaccurate, and to be able to make suitable adjustments if needed

In this paper we present
- User model that estimates
  - User's current interest
  - Accuracy of feedback
- Timeline interface that
  - Visualizes the user's feedback history
  - Highlights inaccurate feedback
  - Allows user to adjust, delete and indicate that feedback is accurate

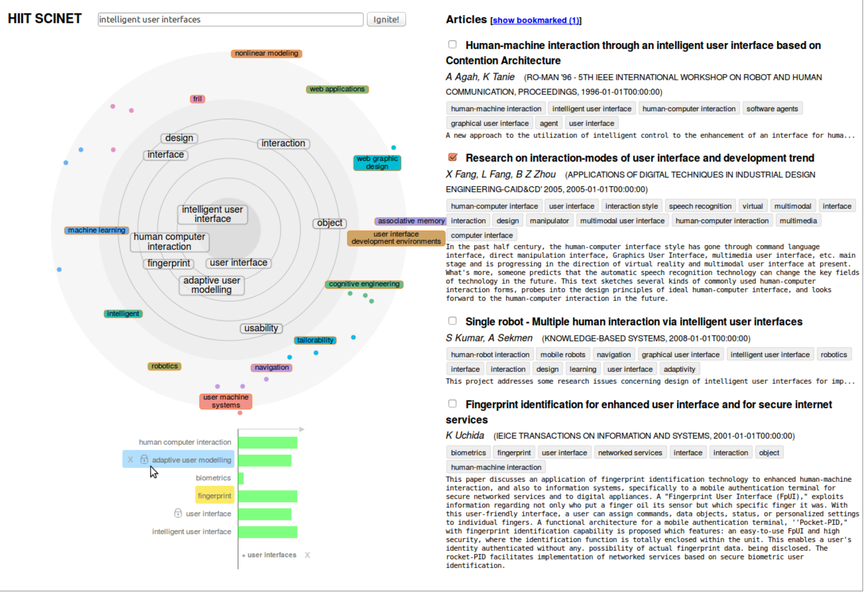
User Model

We assume a linear Gaussian observation model for relevance feedback, where the accuracy may be different for each observation. This gives us the following model (ARD model):

$$y_i \sim Normal(x_i\phi, \sigma^2/w_i),$$
$$\phi_j \sim Normal(\mu_\phi, \lambda_\phi),$$
$$\sigma^2 \sim InverseGamma(\alpha_{\sigma^2}, \beta_{\sigma^2}),$$
$$w_i \sim Gamma(\alpha_{w_i}, \beta_{w_i}), w_i^{fix} \sim Delta(1.0),$$

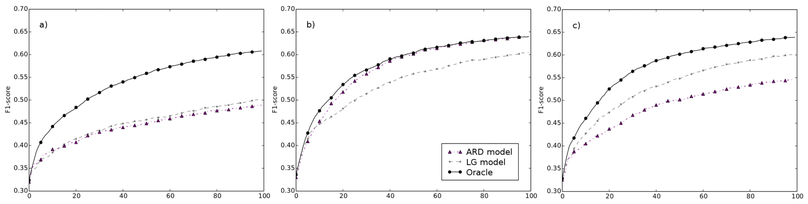
where the distribution of *w* depends on whether the user has marked that feedback as accurate. We approximate the posterior distribution of the model parameters using mean-field variational inference.

User Interface



Simulation Study

We conducted an experiment with a simulated user. The user was searching for newsgroup articles using noisy relevance feedback. We compared the performance of the new model to an oracle and to a baseline (LG model) that did not estimate accuracy of feedback.



Left graph
- User did not make corrections to past feedback

Center graph
- Search engine made suggestions on what feedback to correct
- User corrected feedback values is case of true positive suggestions
- User marked feedback accurate in case of false positive suggestions

Right graph
- As above, but user did not mark feedback accurate in case of false positives

User Study

Baseline: timeline was hidden and user model did not estimate accuracy of feedbacks

16 participants who performed 2 search tasks, one with each search system

Main quantitative results
- New system had better and more diverse results (ResQue 1,3,16)
- New system made it easier to find useful results (ResQue 2,14,15)
- New system made it easier to notice mistakes in feedback and make corrections (ResQue 9,10,11,13)
- Search engine behavior was easier to understand with the new system (ResQue 4,5,6,12; SUS 6)
- Overall usability was similar between the systems (SUS: new 72, bl 68, p=0.7)
- Overall recommendation performance was better with the new system (ResQue: new 55, bl 50, p=0.04)
- Users made less keyword queries with the new system but also interacted more overall with it
- Expert evaluation of task performance was similar with both of the systems, as were the quality of shown keywords and articles (evaluation was made blindly against the given search task description)

N	B	p	Question
3.8	**3.8**	0.9	1: I think that I would like to use this system frequently
2.6	**2.3**	1.0	2: I found the system unnecessarily complex
3.9	**3.9**	1.0	3: I thought the system was easy to use
2.0	**2.0**	1.0	4: I think that I would need the support of a technical person to be able to use this system
3.6	**3.6**	0.8	5: I found the various functions in this system were well integrated
2.2	2.9	0.2	6: I thought there was too much inconsistency in this system
4.3	**4.4**	0.4	7: I would imagine that most people would learn to use this system very quickly
2.1	**2.0**	0.9	8: I found the system very cumbersome to use
3.9	**4.0**	0.6	9: I felt very confident using the system
1.8	1.9	0.8	10: I needed to learn a lot of things before I could get going with this system

Table 1: SUS score question averages for the new interface (N) and the baseline (B) system with p-values. Questions were scored on a 5-point Likert scale from 1 (disagree) to 5 (agree). The better value in each row is in boldface; higher is better for odd numbered questions and lower is better for even numbered questions.

N	B	p	Question
4.1	3.9	0.3	1: The items recommended to me matched what I was searching for
4.6	4.1	0.02	2: The recommender system helped me discover new items
4.0	3.5	0.05	3: The items recommended to me are diverse
3.8	3.4	0.08	4: The layout of the recommender interface is adequate
3.6	3.2	0.2	5: The recommender explains why the items are recommended to me
3.8	3.4	0.2	6: The information provided for the recommended items is sufficient
3.6	3.4	1.0	7: I found it easy to tell the system what I want / don't want to find
4.1	**4.3**	0.5	8: I became familiar with the recommender system very quickly
4.2	3.8	0.2	9: I found it easy to modify my search query in the recommender
3.6	3.1	0.06	10: I found it easy to notice if some of my query modifications were not correct any more
3.9	3.6	0.3	11: I found it easy to find suitable ways to modify my query
3.9	3.5	0.05	12: I understood why the items were recommended to me
3.5	2.9	0.09	13: I found it easy to notice if I had made a mistake in modifying my query
3.9	3.6	0.05	14: Using the recommender to find what I like is easy
3.5	3.2	0.3	15: I found it easy to re-find items I had been recommended before
4.3	4.0	0.2	16: The recommender gave me good suggestions
4.0	3.8	0.5	17: Overall, I am satisfied with the recommender
4.3	4.0	0.3	18: The recommender can be trusted
4.1	3.9	0.5	19: I would use this recommender again, given the opportunity

Table 2: ResQue score question averages for improved (I) and baseline (B) system with p-values. Questions were scored on a 5-point Likert scale from 1 (disagree) to 5 (agree). The better value in each is in boldface; higher is better.

Interview main results
- Majority of users preferred new interface to baseline
- Interface helped track and compare keywords the user had interacted with
- Users felt subjectively more in control of the system
- Some users reported they did not use "marking feedback accurate" feature at all
- Some users felt that there were sometimes too many highlighted keywords on the timeline

Conclusion

Based on the simulation and user experiment, we find that the new system
- Improves the ability of the users to notice and correct mistakes in their feedback
- Shows some indication of improving the quality of the results, at least subjectively evaluated by the users
- Allows the users to more easily direct their search through new feedback options

Future research
- Better ways to infer what user feedback is still useful in modeling the current interest?
- Better ways to ask the user for clarification if conflicting feedback is found?