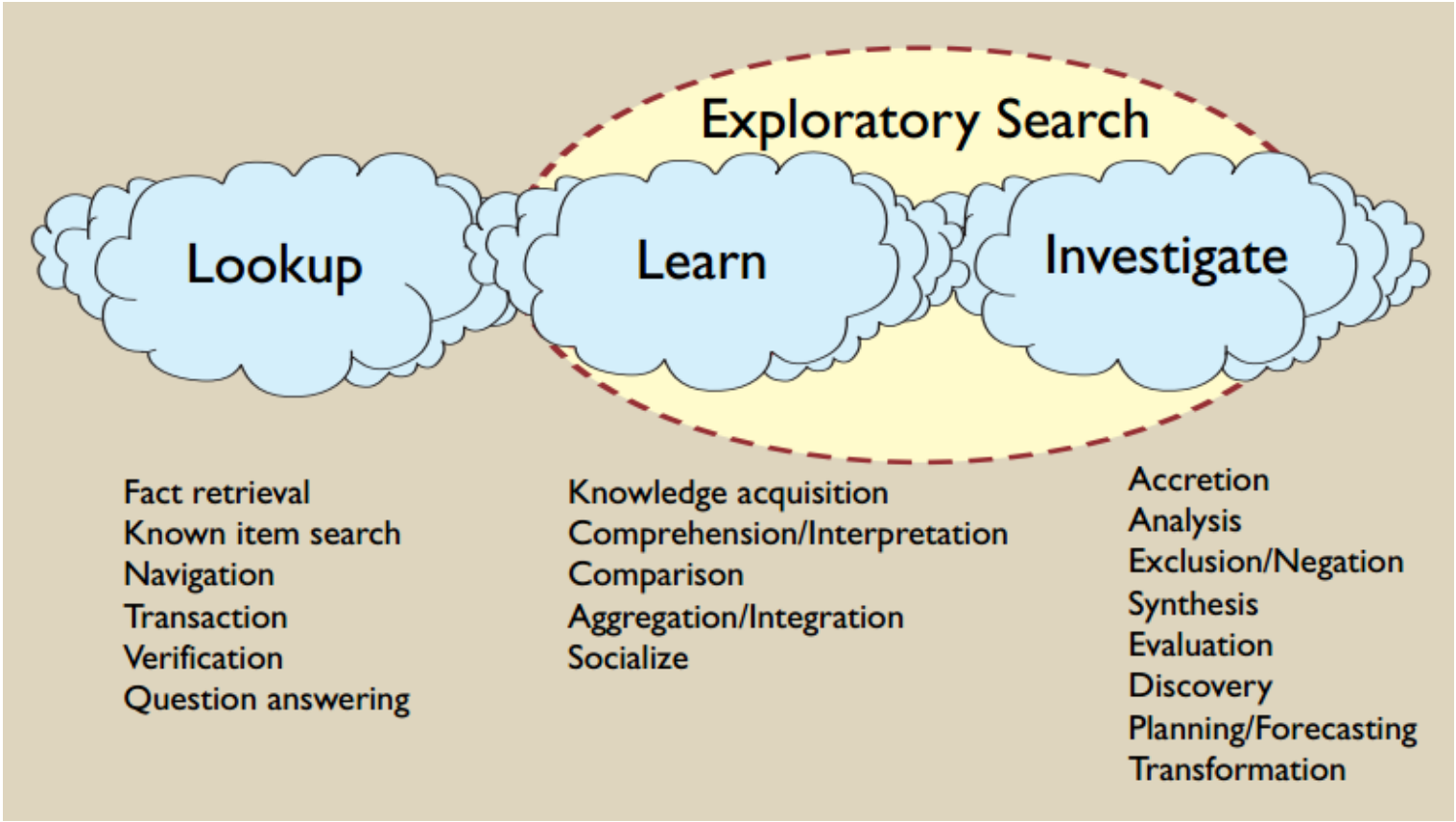
Towards Fine-grained Adaptation of Exploration/Exploitation in Information Retrieval

Alan Medlar, Joel Pyykko, Dorota Glowacka

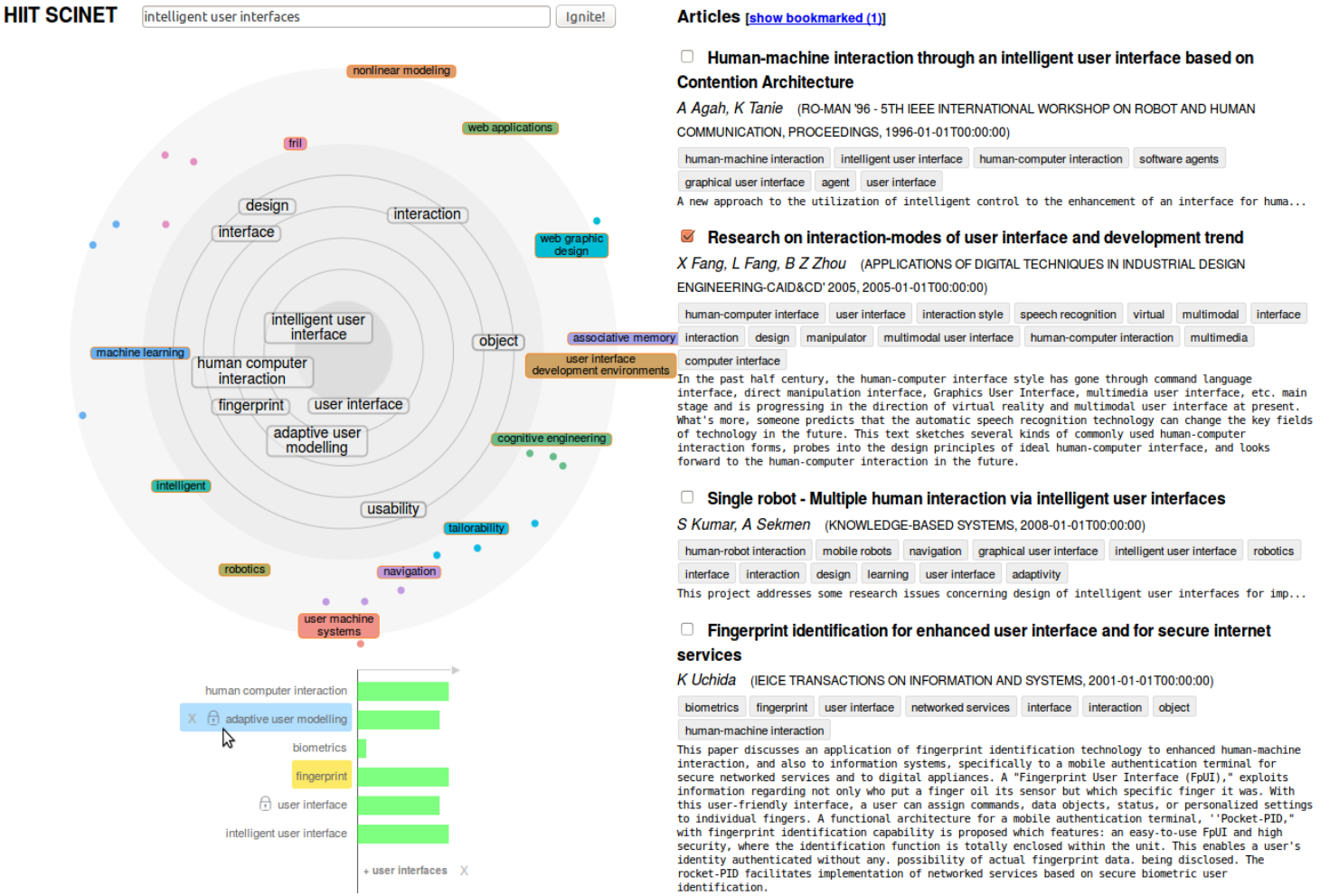
University of Helsinki

Exploratory Search



G. Marchionini. Exploratory search: from finding to understanding. Comm. ACM, 49(4):41–46, 2006

Complex Interface, Single Modes (IUI 2013, 2015, 2016)

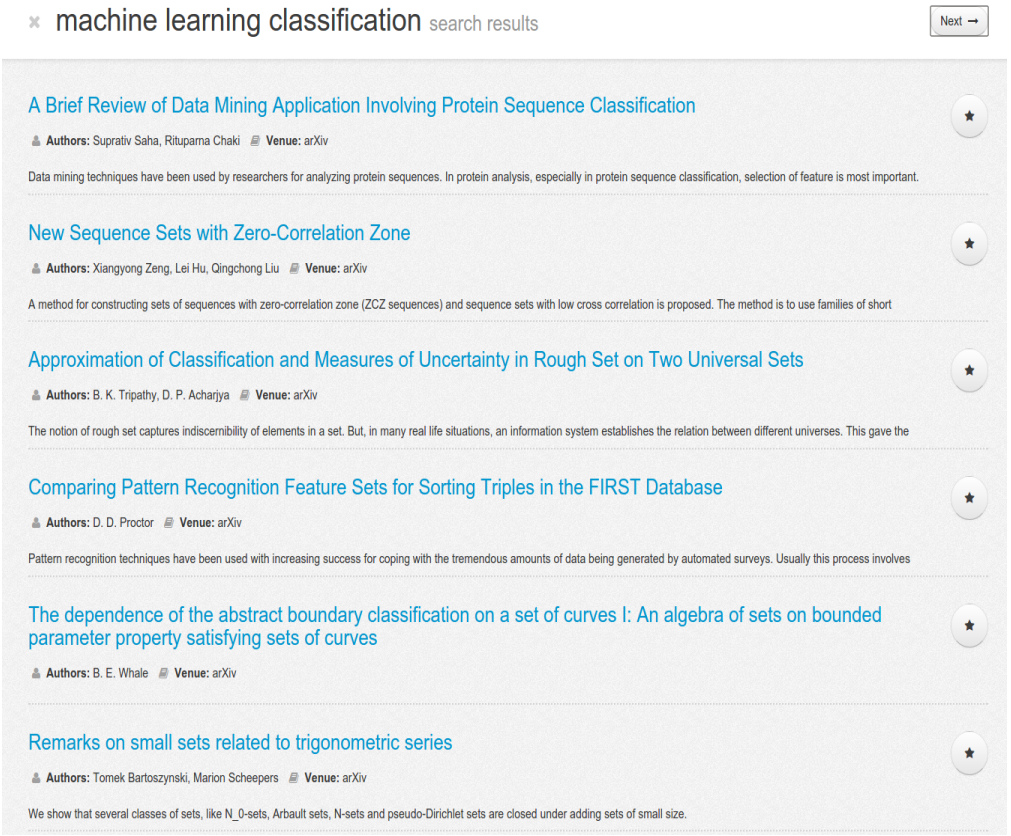


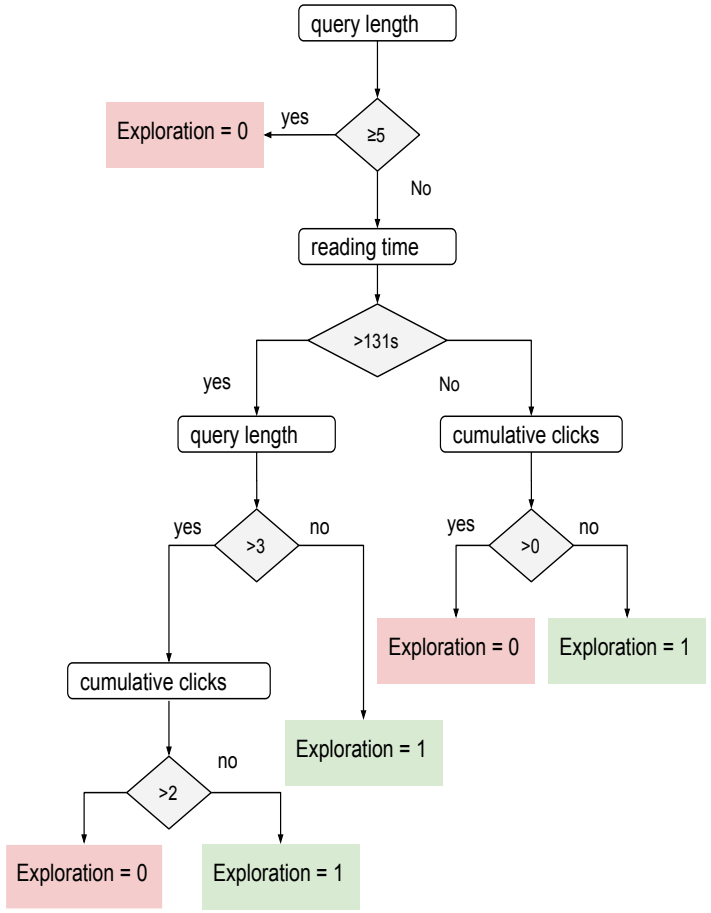
Glowacka et al. *IUI 2013*, Kangasraasio et al. *IUI 2015*, Kangasraasio et al. *IUI 2016*

Problem Definition

- Specialised interfaces to support exploratory search are often not ideal for lookup tasks.
- Users need to know in advance that they will conduct an exploratory search.
- Users need to switch between systems based on task type.
- Research shows that users prefer the simple interfaces used to support lookup tasks.

Simple Interface, Multiple Modes (IUI 2016)





Document Exploration

In each iteration *t*, LinRel calculates:

$$a_i = x_i \cdot \left(X_t^T X_t + \mu I \right)^{-1} X_t^T$$

for each document *i* in dataset and selects for presentation top *n* documents that maximize:

$$\arg\max_X \{ a_i \cdot y_t + \frac{c}{2} \| a_i \| \}$$

for some constant $c > 0$

Adaptive Document Ranking

- To help the user to explore the document space, we use the LinRel algorithm (Auer 2002)
- The algorithm contains parameter *alpha* which controls how diverse the presented documents are: the higher the value of *alpha*, the more diverse the results.
- In our classifier-based system, if a search is classified as exploratory, the value of *alpha* = 1.
- For lookup tasks, alpha = 0.

This study

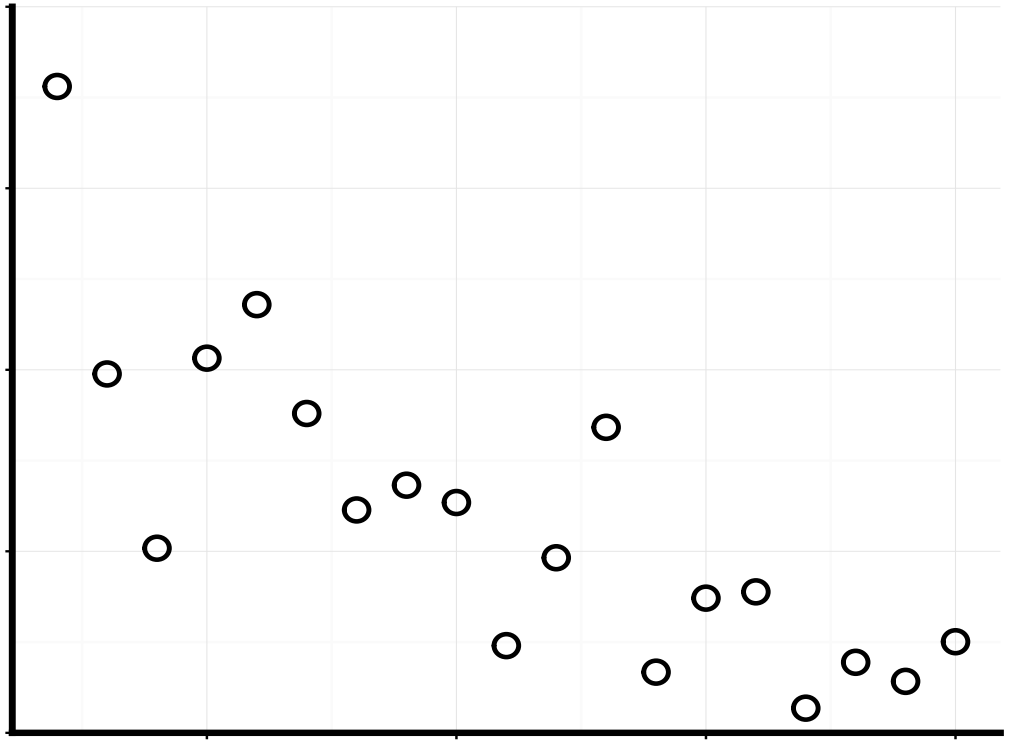
- Previous work used classifier to select between lookup and exploratory tasks
- Single exploration rate per task type, but are search types really discrete categories?
- Can we build a regression model instead of a classifier?

Approach

- Ordinary regression fits a model based on linear relationships between response variable and exploratory values
- Problem: we don't know apriori optimal exploration rate for given user and can't observe their behaviour under these conditions.
- Solution: instead of using specific exploration rates as the response variable, we created censored intervals based on user feedback

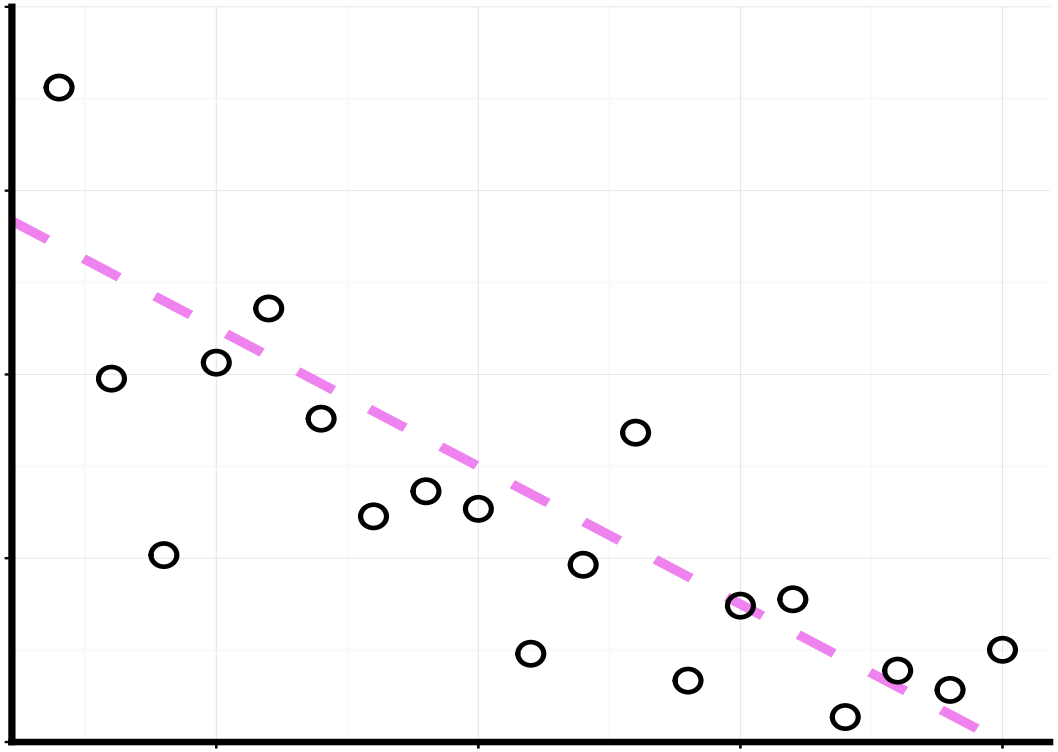
Approach

Ordinary regression fits a model based on linear relationships between response variable and exploratory values



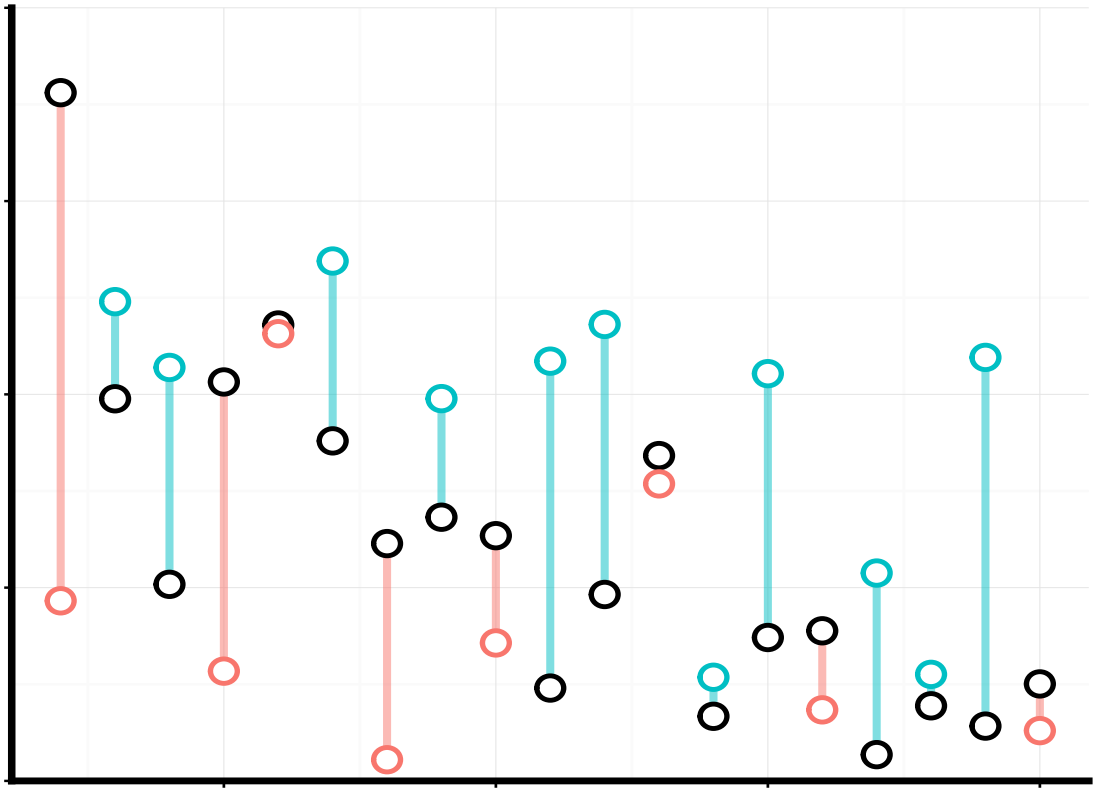
Approach

Ordinary regression fits a model based on linear relationships between response variable and exploratory values



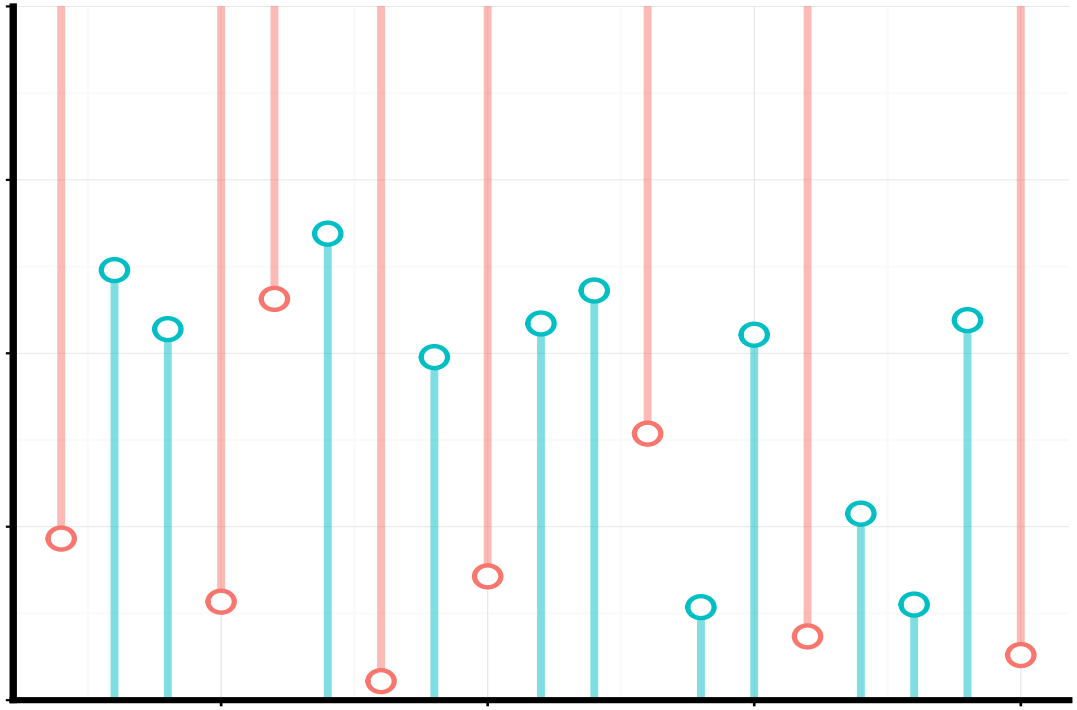
Problem

We don’t know apriori optimal exploration rate for given user and can’t observe their behaviour under these conditions.



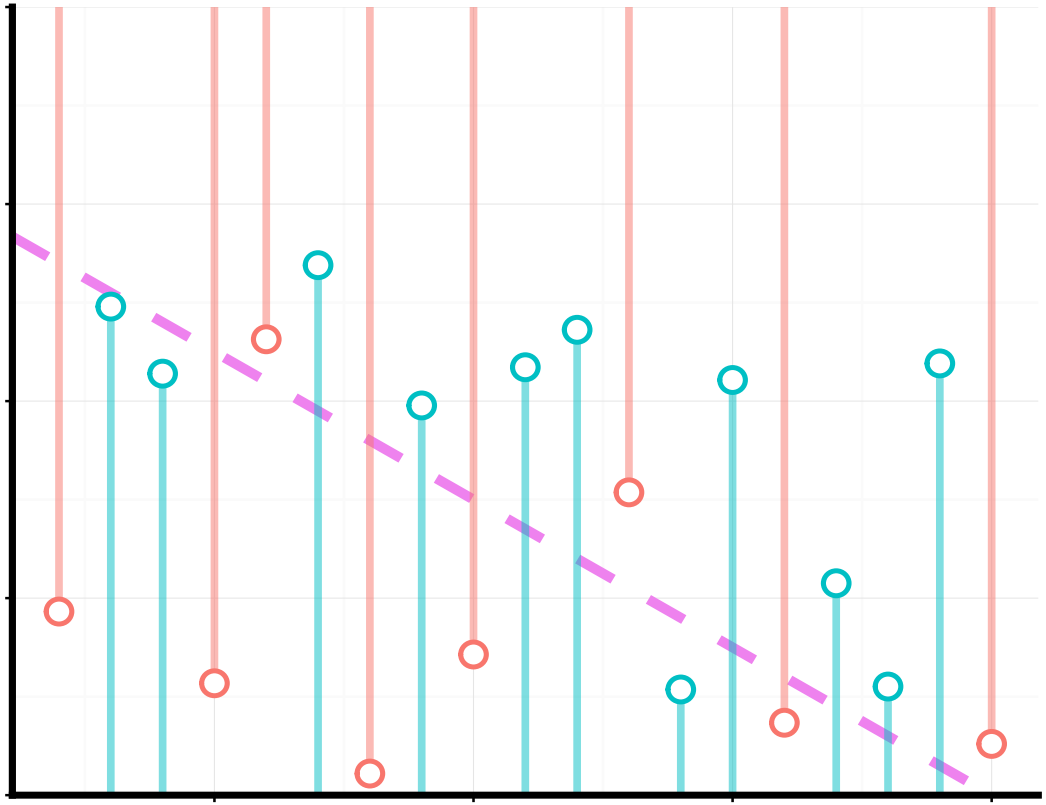
Solution

Instead of using specific exploration rates as the response variable, we created censored intervals based on user feedback.



Solution

Instead of using specific exploration rates as the response variable, we created censored intervals based on user feedback.



User Study Design

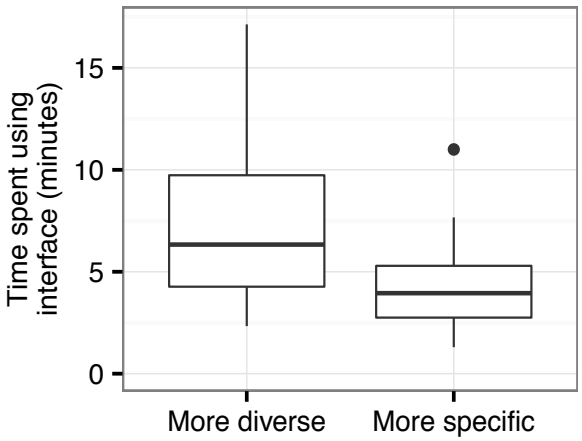
- Random exploration rate
- Participants asked to rate knowledge of topic
- Collect simple metrics: clicks, reading time, etc.
- Participants: 20 MSc students from a CS dept. performing two searches
- Data: 1.1 million arXiv documents
- After each search, participants completed a short questionnaire

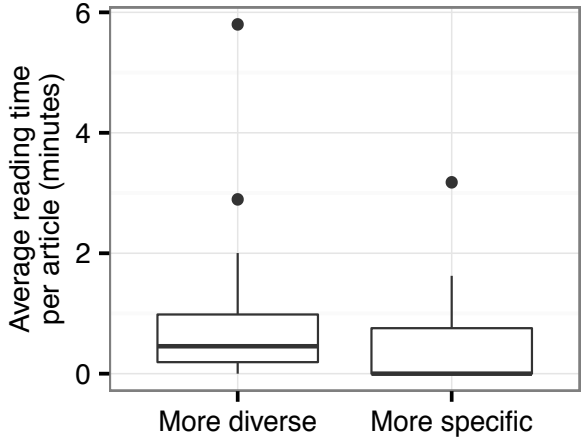
"The search results recommended by the system contained documents closely related to the initial search query as well as articles related to other topics with varying degrees of relevance to the initial query. Based on the search session that you have just completed, would you prefer the search results to contain: a) more articles closely related to the initial search query; b) more articles related to other topics with varying degrees of relevance to the initial search query"

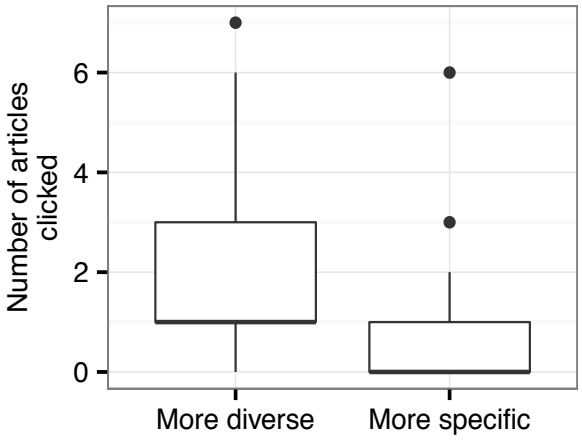
Interval Regression

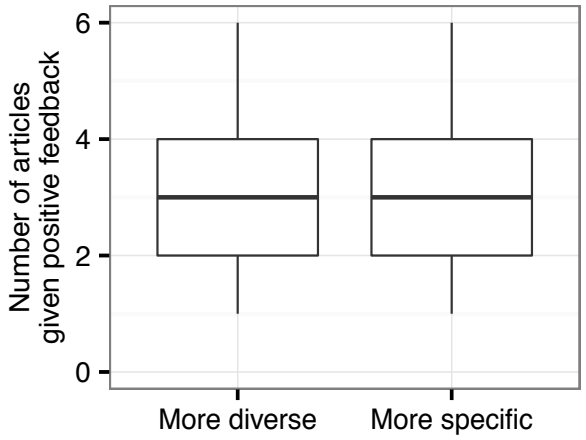
- Left-censored interval [0, *c*] for *more specific* documents and right-censored interval [*c*, *+inf*] for more *diverse documents*
- Three significant predictor variables: time spent with interface, number of clicked articles, self-reported knowledge
- Interval regression performed with Survival R package.

Experimental Results









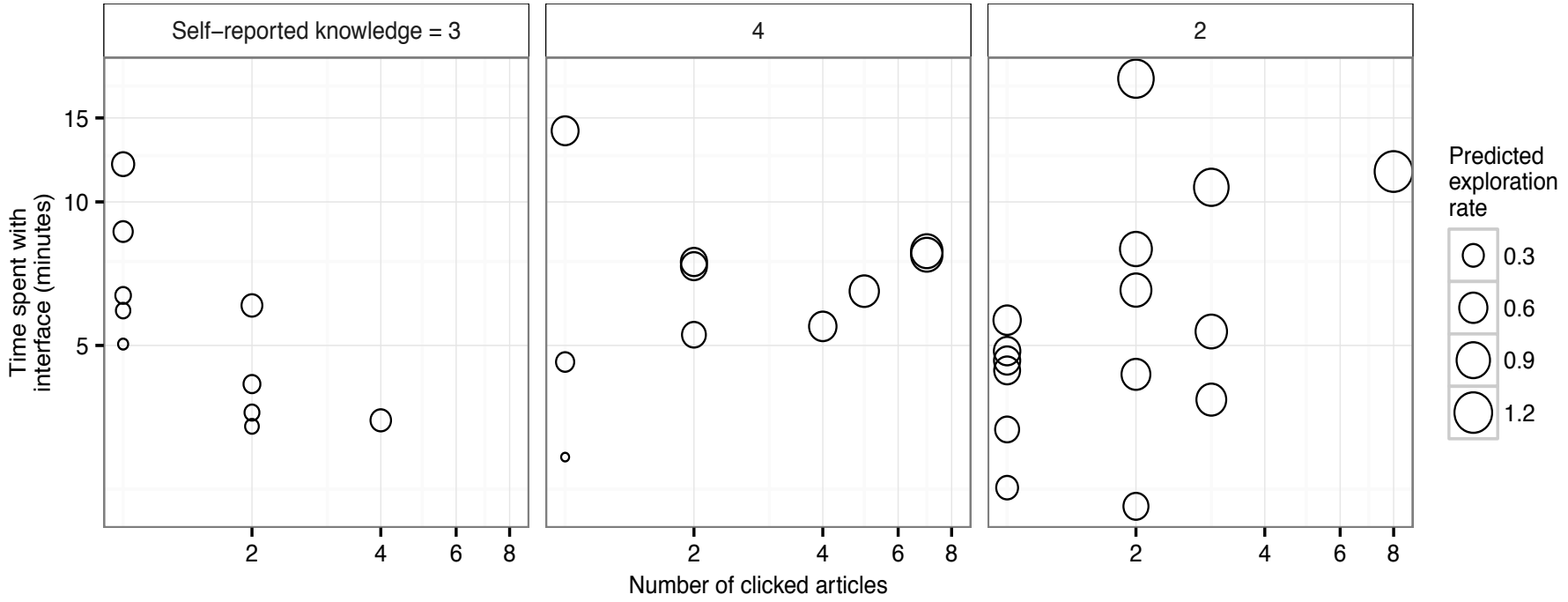
Regression Model

$$c = 0.29\ln(x_1) + 0.22\ln(x_2) - 0.44x_3 - 0.29x_4 + 0.06$$

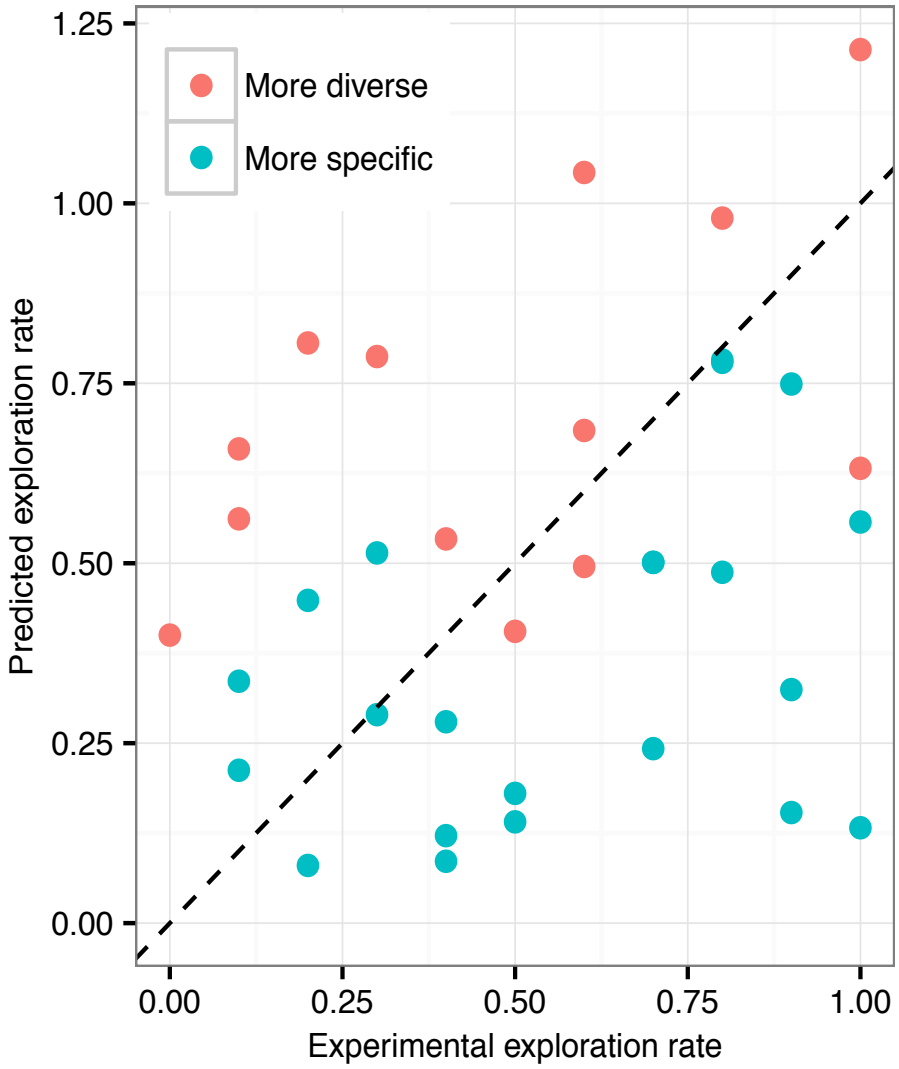
X1 = time spent with interface
X2 = number of documents clicked
X3 = dummy variable for self-reported knowledge level 3
X4 = dummy variable for self-reported knowledge level 4

Note: level knowledge 2 is used as the base-line

Graphical Representation of Model



Model Predictions and User Feedback



Summary

- We presented an approach for automatically adjust exploration rate in IR systems based on user and a specific search task.
- The model is based on easy to collect interaction data, such as clicks and reading time.
- Initial results show model's predictions highly consistent with users' perception of appropriate level of diversity of displayed documents.