

Untangling Result List Refinement and Ranking quality:

A Framework for Evaluation and Prediction

Jiyin He, Marc Bron, Arjen de Vries,
Leif Azzopardi, and Maarten de Rijke

Batch Evaluation

- **Cost effective** evaluation: prediction of search effectiveness based on a series of assumptions on how users use a search system
- Requirements:
 - A collection of documents
 - A set of test queries
 - Relevance judgements
 - **An evaluation metric**

Evaluation metrics and user interaction

information retrieval 🔍

15.800.000 RESULTS Narrow by language ▼ Narrow by region ▼

[Information retrieval - Wikipedia](#) [Translate this page](#)
[nl.wikipedia.org/wiki/Information_retrieval](#) ▼
Information retrieval (IR) houdt zich bezig met het zoeken naar informatie in documenten, naar documenten zelf, naar metadata die de documenten beschrijft, en ...
[Modellen](#) · [Evaluatie](#) · [Belangrijke ...](#) · [Literatuur](#)


[Information retrieval - Wikipedia, the free encyclopedia](#)
[en.wikipedia.org/wiki/Information_retrieval](#) ▼
Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be ...
[Overview](#) · [History](#) · [Model types](#) · [Performance and ...](#) · [Awards in the field](#)

[Information Retrieval Day](#) [Translate this page](#)
[www.informationretrievalday.nl](#) ▼
Op de Information Retrieval Day 2011 stonden acht ervaren sprekers stil bij de huidige veranderingen in het informatiemanagement en gaven zij een beeld van de ...

[Introduction to Information Retrieval - Stanford University](#)
[nlp.stanford.edu/IR-book/information-retrieval-book.html](#) ▼
Introduction to Information Retrieval. This is the companion website for the following book. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze ...

[INFORMATION RETRIEVAL - University of Glasgow :: ...](#)
[www.dcs.gla.ac.uk/Keith/Preface.html](#) ▼
PREFACE TO THE FIRST EDITION (London: Butterworths, 1975) The material of this book is aimed at advanced undergraduate information (or computer) science students ...

- Evaluation metrics (Carterette, 2011)

- 
- A user interaction model: How users interact with a ranked list
 - Associating user interactions with effort or gain
 - Current batch evaluation metrics: boils down to the ranking quality of the results

Beyond a ranked list

The image shows two side-by-side search results pages. The left page is from Amazon, displaying search results for 'information retrieval'. It features a sidebar with 'Facets' (Books, Kindle Store, Refine by) and a main list of books. The right page is from UvAweb, displaying search results for 'information retrieval'. It features a sidebar with 'Categories' (Filter on: Type, Period) and a main list of search subjects. Green circles highlight the facets/categories sections on both pages.

Amazon Search Results:

- Search: information retrieval
- 1-16 of 48,376 results for "information retrieval"
- Facets:**
 - Books >**
 - Computer Network Administration
 - Computers & Technology
 - Databases
 - Computer Programming
 - Online Internet Searching
 - + See more
 - Kindle Store >**
 - Computer Databases
 - Computers & Technology
 - Computer Programming
 - + See All 26 Departments
 - Refine by**
 - Eligible for Free Shipping
 - Free Shipping by Amazon
- Results:**
 - Introduction to Information Retrieval** by Raghuvaran and Hinrich Schütze (J...)
 - \$26.55 to rent Hardcover Prime
 - \$55.69 to buy
 - Only 18 left in stock - order soon.
 - \$38.10 Kindle Edition
 - Auto-delivered wirelessly
 - More Buying Choices - Hardcover
 - \$40.43 new (53 offers)
 - \$40.44 used (36 offers)
 - Information Retrieval: Imp...** by Stefan Buettcher, Charles L. A. Cla...
 - \$58.00 \$37.42 Hardcover Prime
 - Only 5 left in stock - order soon.
 - \$38.10 Kindle Edition
 - Auto-delivered wirelessly
 - More Buying Choices - Hardcover
 - \$37.04 new (30 offers)
 - \$25.94 used (21 offers)

UvAweb Search Results:

- Search the UvAweb
- Filter on: information retrieval
- Categories:**
 - Type
 - ☐ Article (3490)
 - ☐ Course (1957)
 - ☐ Event (1527)
 - ☐ News (1429)
 - ☐ Course catalogue (1426)
 - ☐ Programme (83)
 - ☐ Organisational unit (41)
 - ☐ Vacancy (36)
 - ☐ Discipline (3)
 - Period
 - ☐ Last 24 hours (24)
- Search subject**
 - Results 1 - 20 of 9996
 - Nieuwe inzichten en ontwikkelingen in zoekmachinetechnologie**
 - datastromen in de stad van de toekomst? Deze en andere vragen worden beantwoord 'European Conference on Information Retrieval' (ECIR '14).
 - www.uva.nl/nieuws-agenda/agenda/alle-evenementen/content/congressen/...
 - Information Retrieval for Information Services**
 - Burscher works on the project 'Information Retrieval for Information Services COMMIT research framework. His dissertation and research focus on the auton analysis for communication research. Currently, he works ...
 - ascor.uva.nl/research/phd-research-projects/...
 - Information Retrieval - Artificial Intelligence**
 - engines to text analysis. Information Retrieval has developed from a number of resea including Computer Science, Library Science, Artificial Intelligence, Data Mining, and Processing. While Information Retrieval builds on techniques from ...
 - gss.uva.nl/masters-programmes/content26/study-programme/...

Facets

Categories

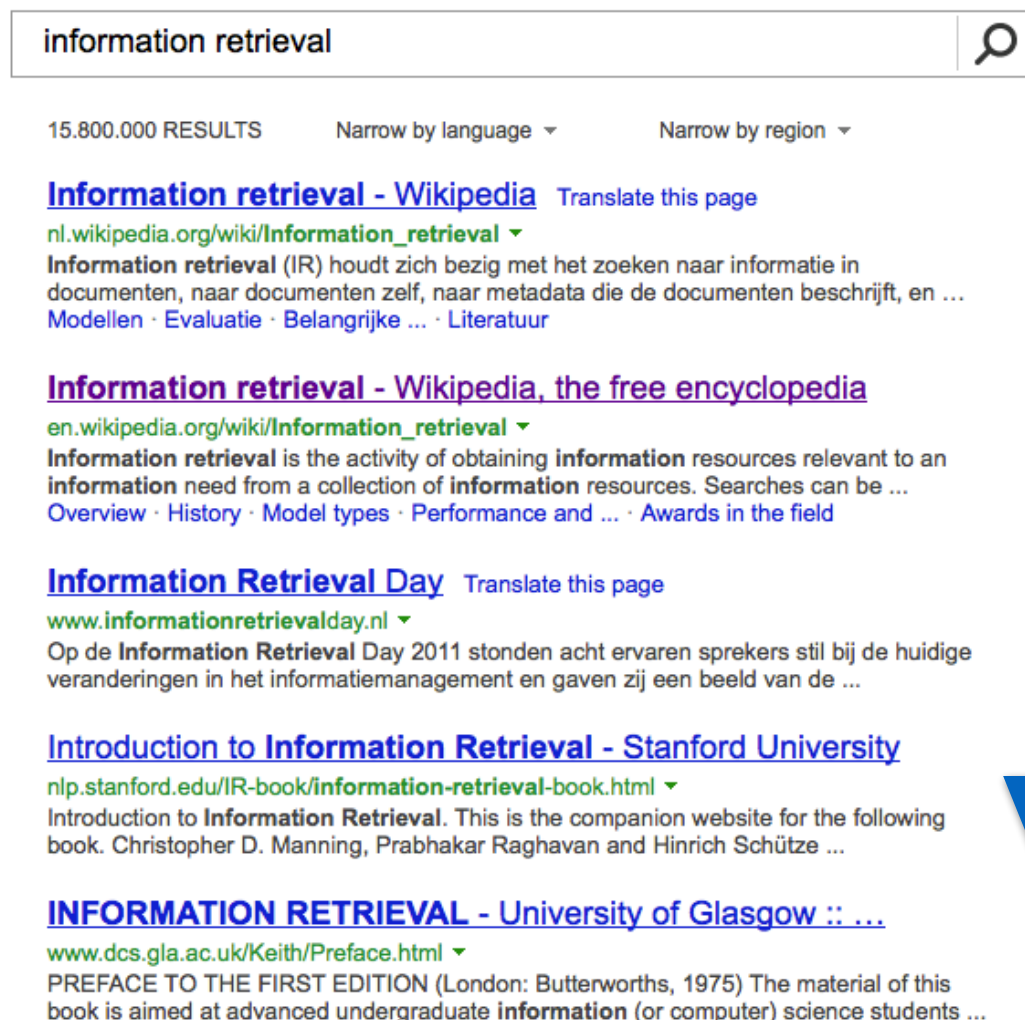
result list refinement (RLR) elements

Q: how do we evaluate and compare systems under varying conditions of ranking quality, interface elements, as well as different user search behaviour?

Our solution

- An effort/gain-based user interaction model
 - How users interact with a **ranked list** and the **RLR elements**
 - Associating user interactions with effort and gain
- Applications
 - **Prediction**: system performance w.r.t a particular application and user group
 - Model parameters derived from user studies
 - **Simulation**: whole system evaluation under varying conditions: ranking quality, interface elements, user types
 - Model parameters based on hypothesised values

Modelling user interaction: with a ranked list



A screenshot of a search engine results page for the query "information retrieval". The search bar at the top shows the query and a magnifying glass icon. Below the search bar, it indicates "15.800.000 RESULTS" and provides filters for "Narrow by language" and "Narrow by region". The results list includes several entries with titles, URLs, and brief descriptions. A large blue arrow points from the search results towards the flowchart on the right.

information retrieval

15.800.000 RESULTS Narrow by language ▾ Narrow by region ▾

[Information retrieval - Wikipedia](#) [Translate this page](#)
nl.wikipedia.org/wiki/Information_retrieval ▾
Information retrieval (IR) houdt zich bezig met het zoeken naar informatie in documenten, naar documenten zelf, naar metadata die de documenten beschrijft, en ...
[Modellen](#) · [Evaluatie](#) · [Belangrijke ...](#) · [Literatuur](#)

[Information retrieval - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Information_retrieval ▾
Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be ...
[Overview](#) · [History](#) · [Model types](#) · [Performance and ...](#) · [Awards in the field](#)

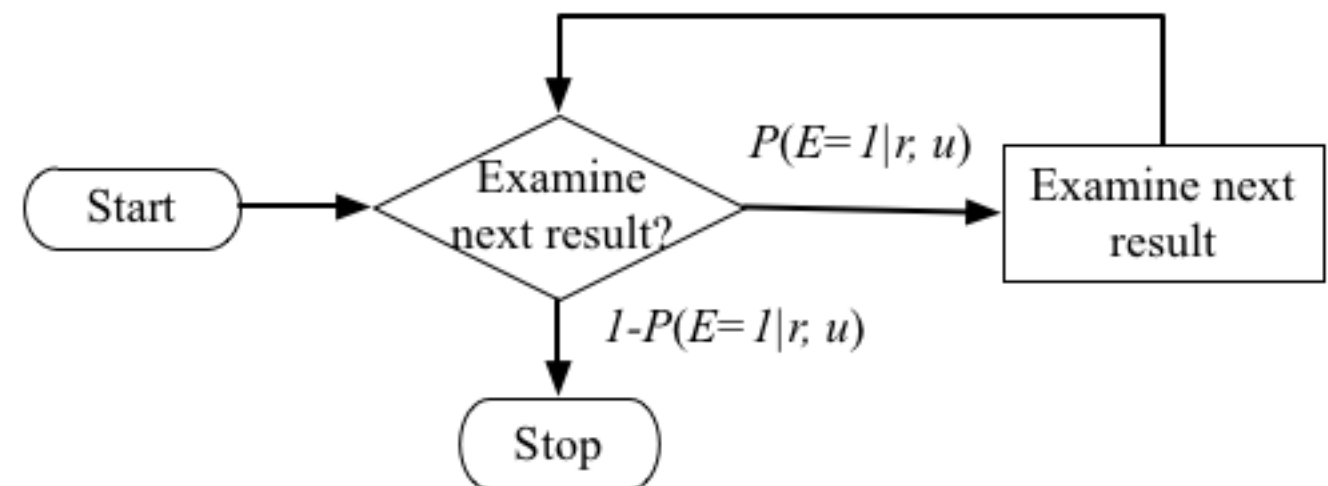
[Information Retrieval Day](#) [Translate this page](#)
www.informationretrievalday.nl ▾
Op de Information Retrieval Day 2011 stonden acht ervaren sprekers stil bij de huidige veranderingen in het informatiemanagement en gaven zij een beeld van de ...

[Introduction to Information Retrieval - Stanford University](#)
nlp.stanford.edu/IR-book/information-retrieval-book.html ▾
Introduction to Information Retrieval. This is the companion website for the following book. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze ...

[INFORMATION RETRIEVAL - University of Glasgow :: ...](#)
www.dcs.gla.ac.uk/Keith/Preface.html ▾
PREFACE TO THE FIRST EDITION (London: Butterworths, 1975) The material of this book is aimed at advanced undergraduate information (or computer) science students ...

E.g., following assumptions of user behaviour as in RBP

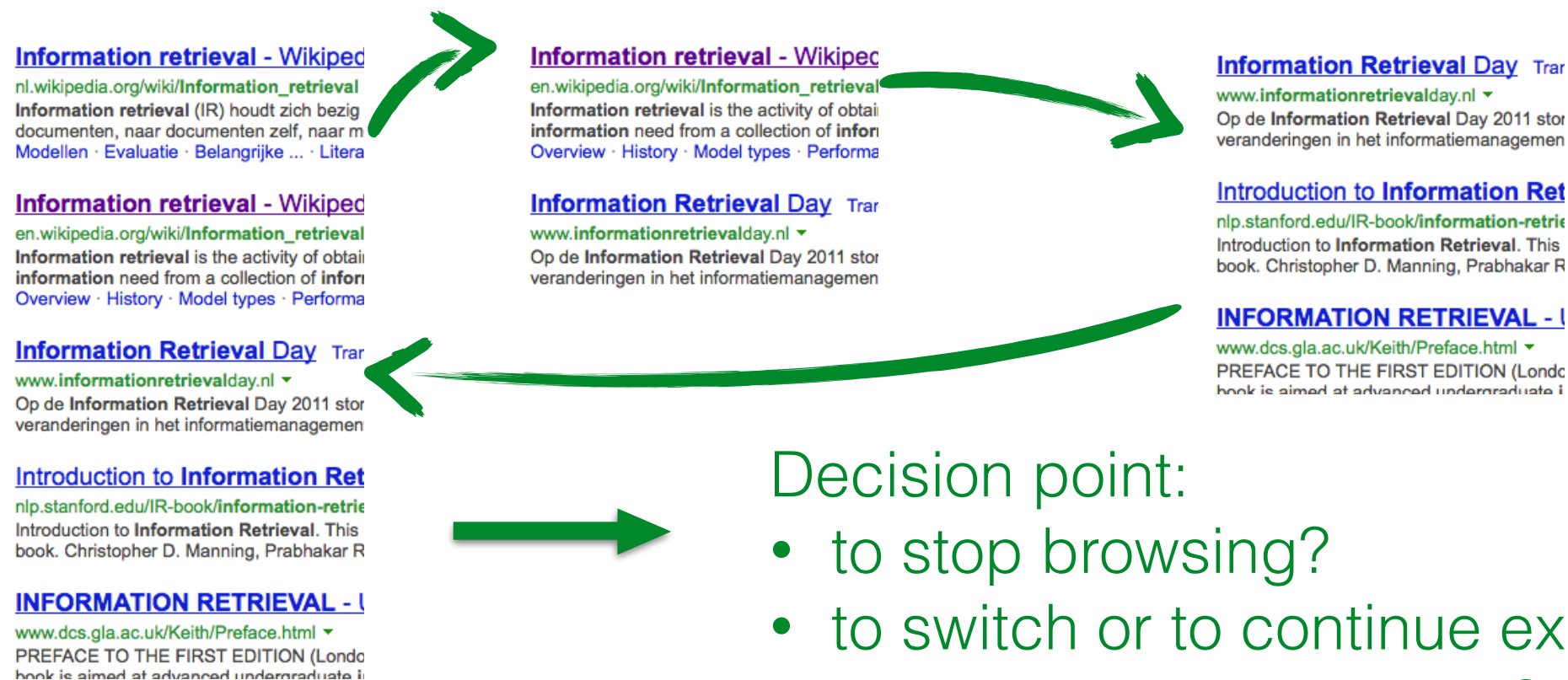
Parameter:
continuation



➡ Decision point: when to stop?

Modelling user interaction: with result list refinement

- Result refinement = switching between different filtered versions of the ranked list (sublists)



Combinatory number of possible user paths:

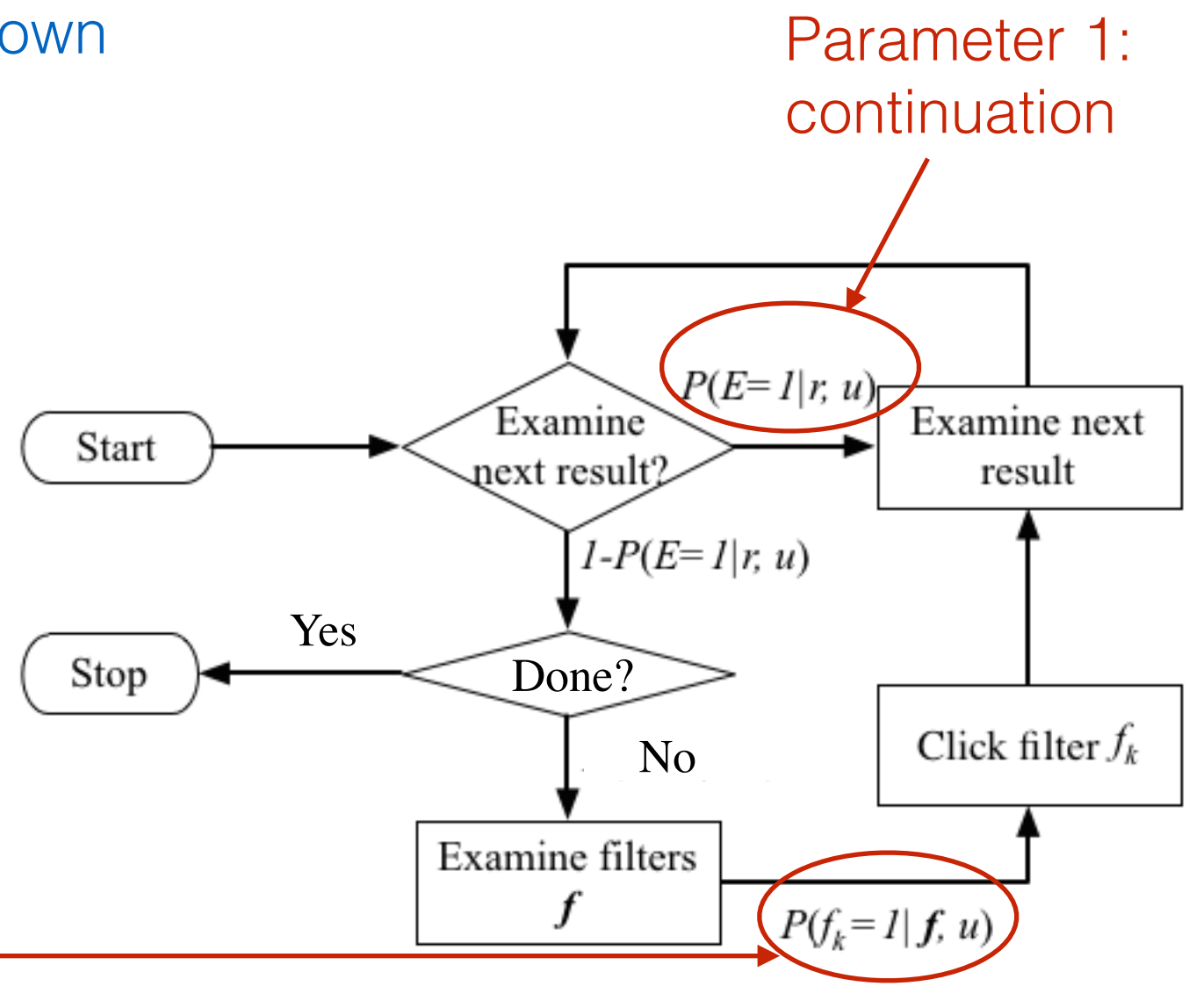
- A Monte-Carlo solution

Modelling user interaction: with result list refinement



- Action path constraints
 - In each sublist, users browse **top-down**
 - common assumption; reducing possible paths from $n!$ to constant
 - Users **skip and only skip** documents already seen
 - preventing inflated relevance and infinite switching
 - **Deterministic quitting point**
 - gain based: quit when certain amount of effort is spent
 - effort based: quit when certain amount of gain is achieved

Modelling user interaction: with result list refinement

- Action path constraints
 - In each list, users browse **top-down**
 - Users **skip and only skip** documents already seen
 - **Deterministic quitting point**
 - gain based: quit when certain amount of effort is spent
 - effort based: quit when certain amount of gain is achieved



User actions, efforts, and gain

- From user action paths to user efforts and gain
 - Each action is associated with an effort
 - Each action may or may not result in a gain, i.e., finding relevant document
- User actions
 - Examine result, refine a list, pagination
- Simple assumption about effort and gain
 - Equal unit effort for all actions  Total effort = # actions
 - Equal unit gain for all relevant documents  Total gain = # relevant docs found

Validation of prediction

- RQs
 - Does the predicted effort correlate to user effort derived from usage data?
 - Can we accurately predict when a RLR interface is beneficial, compared to a basic interface?
- 3 Steps
 - Obtaining usage data from user study
 - Measuring (real) user effort
 - Predicting user performance by calibrated user interaction model
- Data
 - TREC 2013 Federated Search track
 - 50 topics with retrieved web pages and snippets, all judged
 - Results from 108 verticals, each associated with one or more categories

Obtaining usage data: study design

- User task (He et al., 2014)
 - Finding 10 relevant documents
 - Manageable effort, potential for considerable effort save
 - within 50 clicks
 - Preventing randomly clicking all results
 - Snippet based relevance judgement with user feedback
 - Reduced user variability in relevance judgement
- Experiment design
 - Between subject
 - Randomised topic and interface assignment

Obtaining usage data: interfaces

Basic interface

Please find 10 pages that are relevant to:

tuning fork ④

You want to buy a tuning fork.

Show explanations and examples of relevant/irrelevant results ↗

Current topic:
tuning fork

Clicks left: 19 ②

Target to reach 10

I give up! ③

Number of results found: 618

★ [Tuning Fork](#) ⑤

<http://www.desmondreedmusic.com/tuning-fork/>

Tuning Fork (clueweb12-1716wb-63-27206) <http://www.desmondreedmusic.com/tuning-fork/>

Tuning Fork Tuning Fork A440 Tuning Fork A 440 HZ With A Pouch \$3.99 Vintage tuning fork C 256 Phillip Harris Co \$1.69 One C Tuning Fork \$5.28 Wittner 4 1 8 Tuning Fork A 440 FAST SHIPPING \$4.95 Tuning Fork Key of C NEW \$5.49 Vintage...

★ [How Tuning Forks Work: Non-musical Uses for Tuning Forks](#)

<http://www.howstuffworks.com/tuning-fork.htm>

Tuning forks have been around for centuries and are the only sure-fire way to tell if an instrument is in tune. Learn how tuning forks work...Some doctors might use tuning forks to test hearing loss. ©iStockphoto/Thinkstock While keeping orchestras and concert bands in check, tuning forks...

★ [Tuning Fork Therapy: Planetary Tuning Forks - Page 13](#)

<http://books.google.nl/books?id=zPj738XZAKYC&pg=PA13&dq=tuning+fork&hl=en&sa=X&e...>

Ways to Use your Tuning Forks 1. Place the handle of a vibrating tuning fork on a specific chakra. 2. Place the handle of vibrating tuning fork on specific reflex point. 3. Place the handle of a vibrating tuning fork on a muscle. 4. Place the handle ...

⑥ 1 More »

RLR interface

Please find 10 pages that are relevant to:

tuning fork ④

You want to buy a tuning fork.

Show explanations and examples of relevant/irrelevant results ↗

Current topic:
tuning fork

Clicks left: 19 ②

Target to reach 10

I give up! ③

Number of results found: 618

★ [Tuning Fork](#) ⑤

<http://www.desmondreedmusic.com/tuning-fork/>

Tuning Fork (clueweb12-1716wb-63-27206) <http://www.desmondreedmusic.com/tuning-fork/>

Tuning Fork Tuning Fork A440 Tuning Fork A 440 HZ With A Pouch \$3.99 Vintage tuning fork C 256 Phillip Harris Co \$1.69 One C Tuning Fork \$5.28 Wittner 4 1 8 Tuning Fork A 440 FAST SHIPPING \$4.95 Tuning Fork Key of C NEW \$5.49 Vintage...

★ [How Tuning Forks Work: Non-musical Uses for Tuning Forks](#)

<http://www.howstuffworks.com/tuning-fork.htm>

Tuning forks have been around for centuries and are the only sure-fire way to tell if an instrument is in tune. Learn how tuning forks work...Some doctors might use tuning forks to test hearing loss. ©iStockphoto/Thinkstock While keeping orchestras and concert bands in check, tuning forks...

★ [Tuning Fork Therapy: Planetary Tuning Forks - Page 13](#)

<http://books.google.nl/books?id=zPj738XZAKYC&pg=PA13&dq=tuning+fork&hl=en&sa=X&e...>

Ways to Use your Tuning Forks 1. Place the handle of a vibrating tuning fork on a specific chakra. 2. Place the handle of vibrating tuning fork on specific reflex point. 3. Place the handle of a vibrating tuning fork on a muscle. 4. Place the handle ...

⑥ 1 More »

➤ All categories 618

➤ Academic 65

➤ Audio 25 ①

➤ Blogs 10

➤ Books 32

➤ Encyclopedia 31

➤ Entertainment 6

➤ Games 4

➤ General 37

➤ Health 50

➤ Kids 37

➤ Local 4

➤ News 46

➤ Photo/Pictures 72

➤ Q&A 44

➤ Recipes 1

➤ Shopping 56

➤ Social 12

➤ Software 21

➤ Sports 26

➤ Tech 36

➤ Travel 3

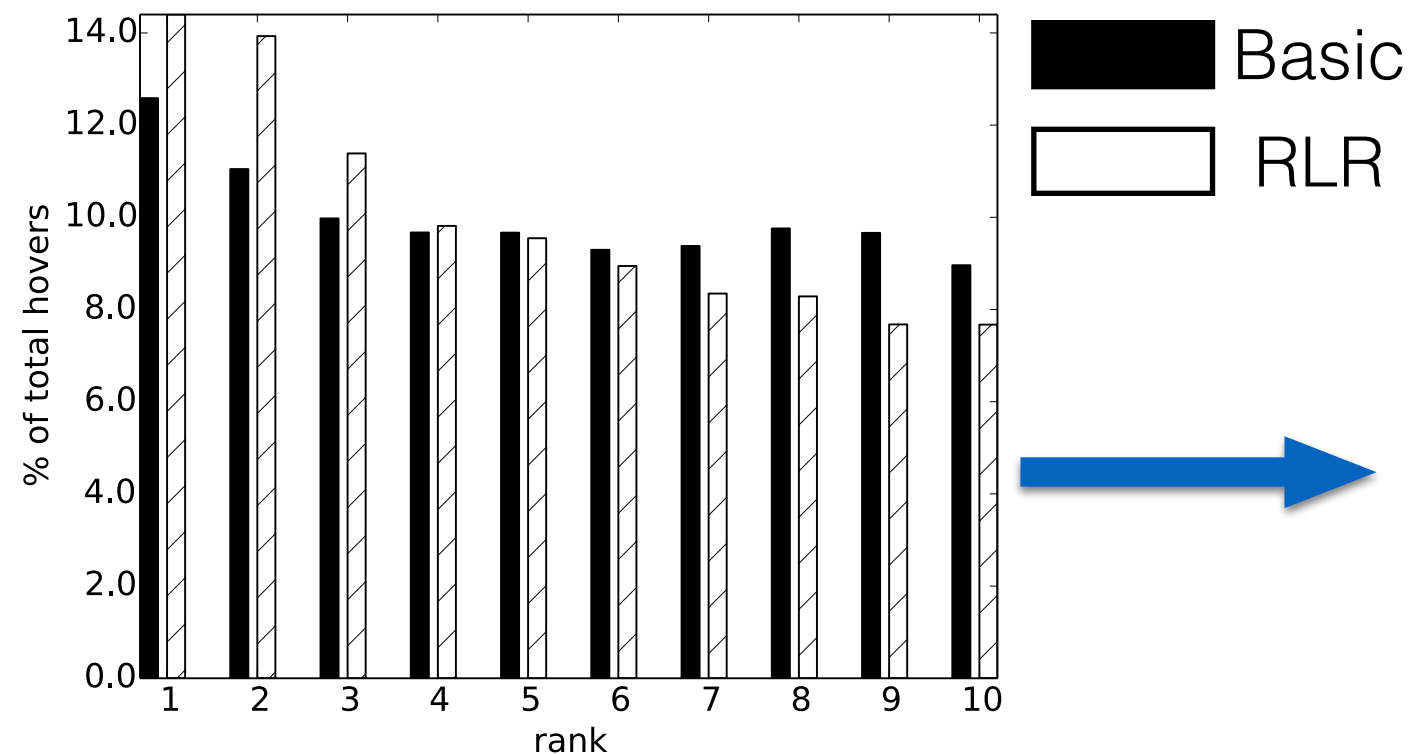
➤ Video 44

Obtained usage data

	Basic	RLR
Completed task instances	145 (Median p. task: 2)	255 (Median p. task: 3)
#Participants	49	48
#Uncompleted task instances	35	28

Measuring (real) user effort

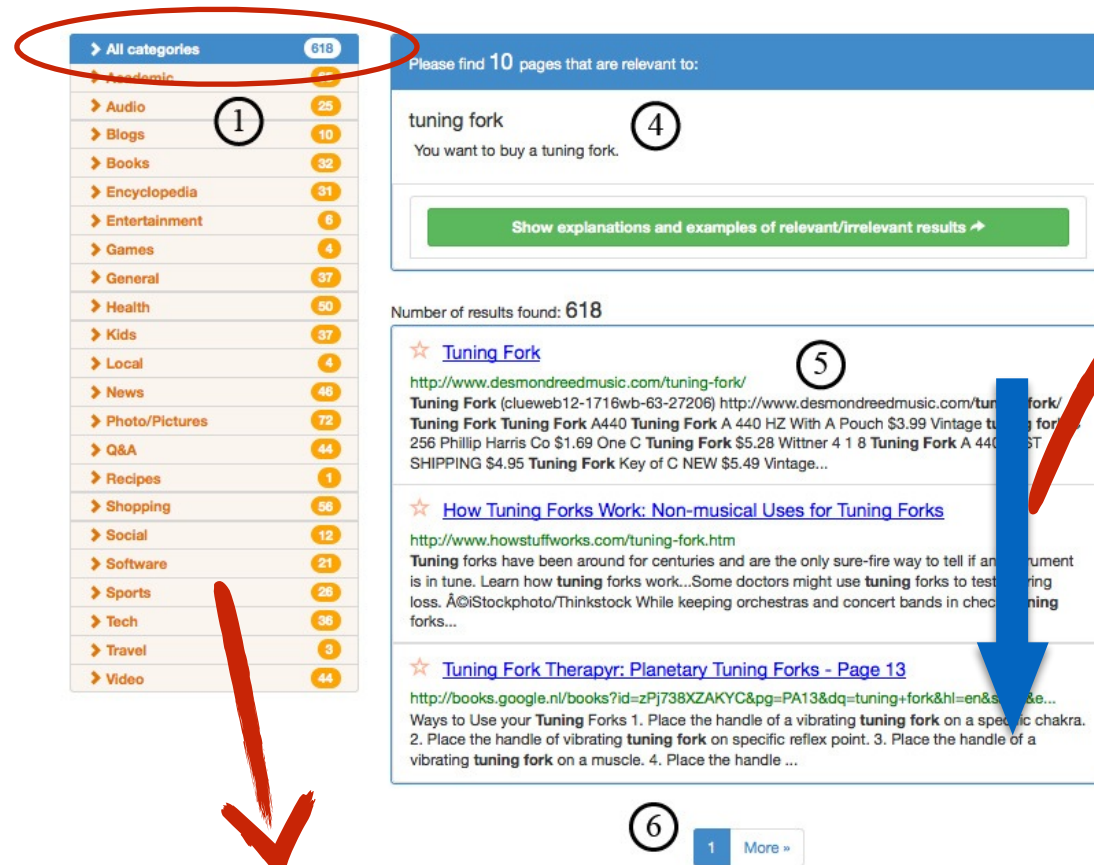
- Examine result: mouse hover over a result snippet
 - # results visited on a SERP =
 - all results in a page before a “pagination” action +
 - up to the last clicked result on the last visited page



Mild position bias: as a result of snippet-based result examination

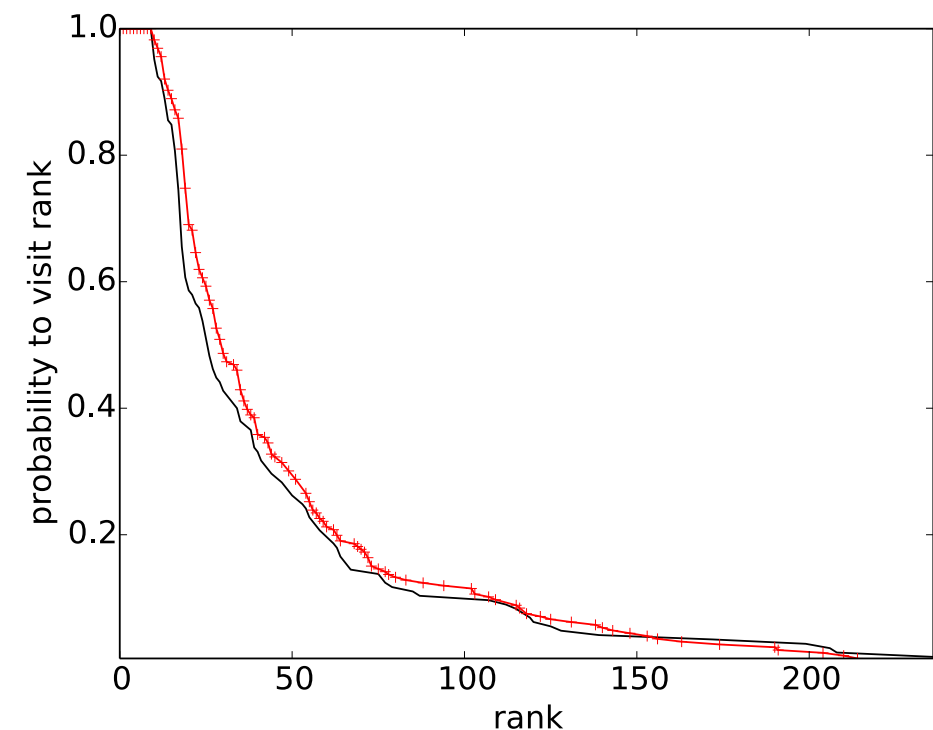
Predicting user effort with calibrated interaction model

Default selection



Parameter 1: continuation

Probability a result is visited @ rank K

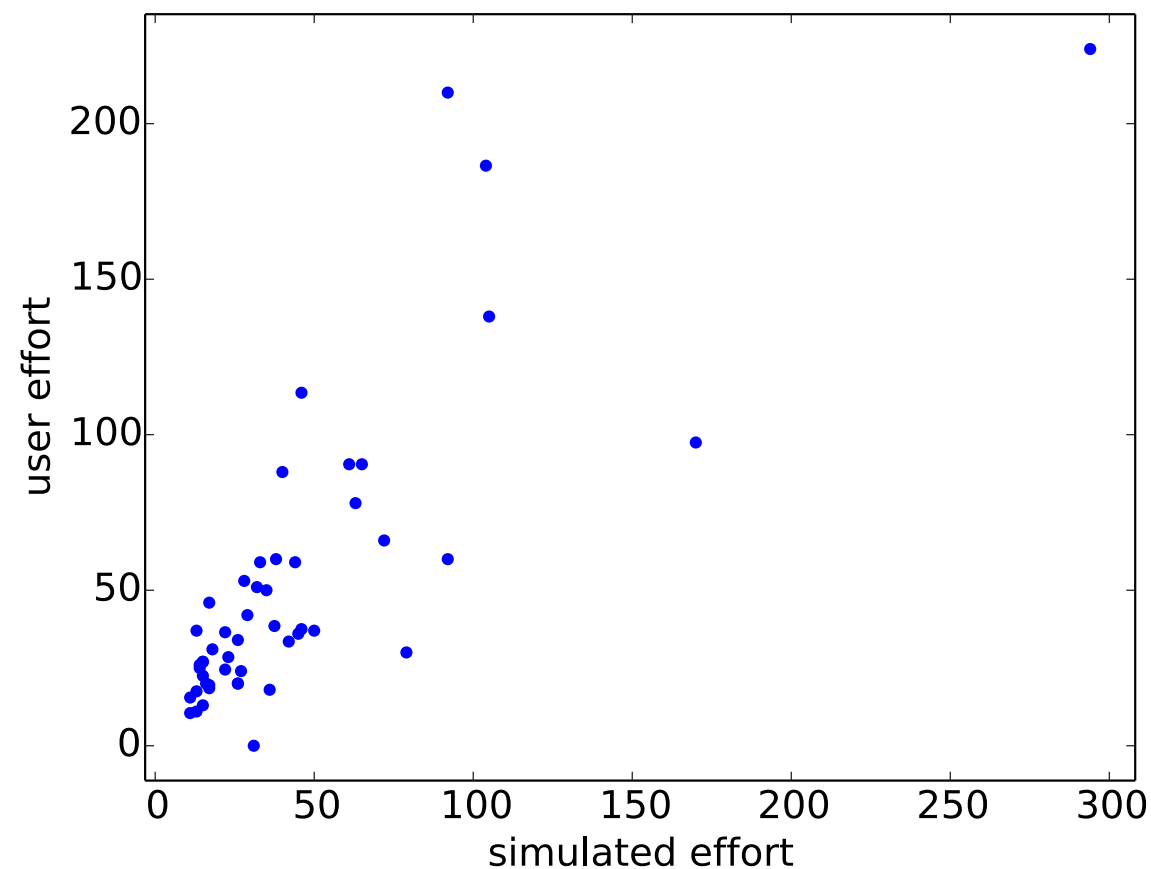


Parameter 2: List selection

- Per topic, the relative frequency that a filter is chosen
- Default selection: “All categories”

Q1: Does the predicted effort correlate to user effort?

- Predicted effort: an approximation of real user effort
 - Correlation as a measure for the accuracy of approximation

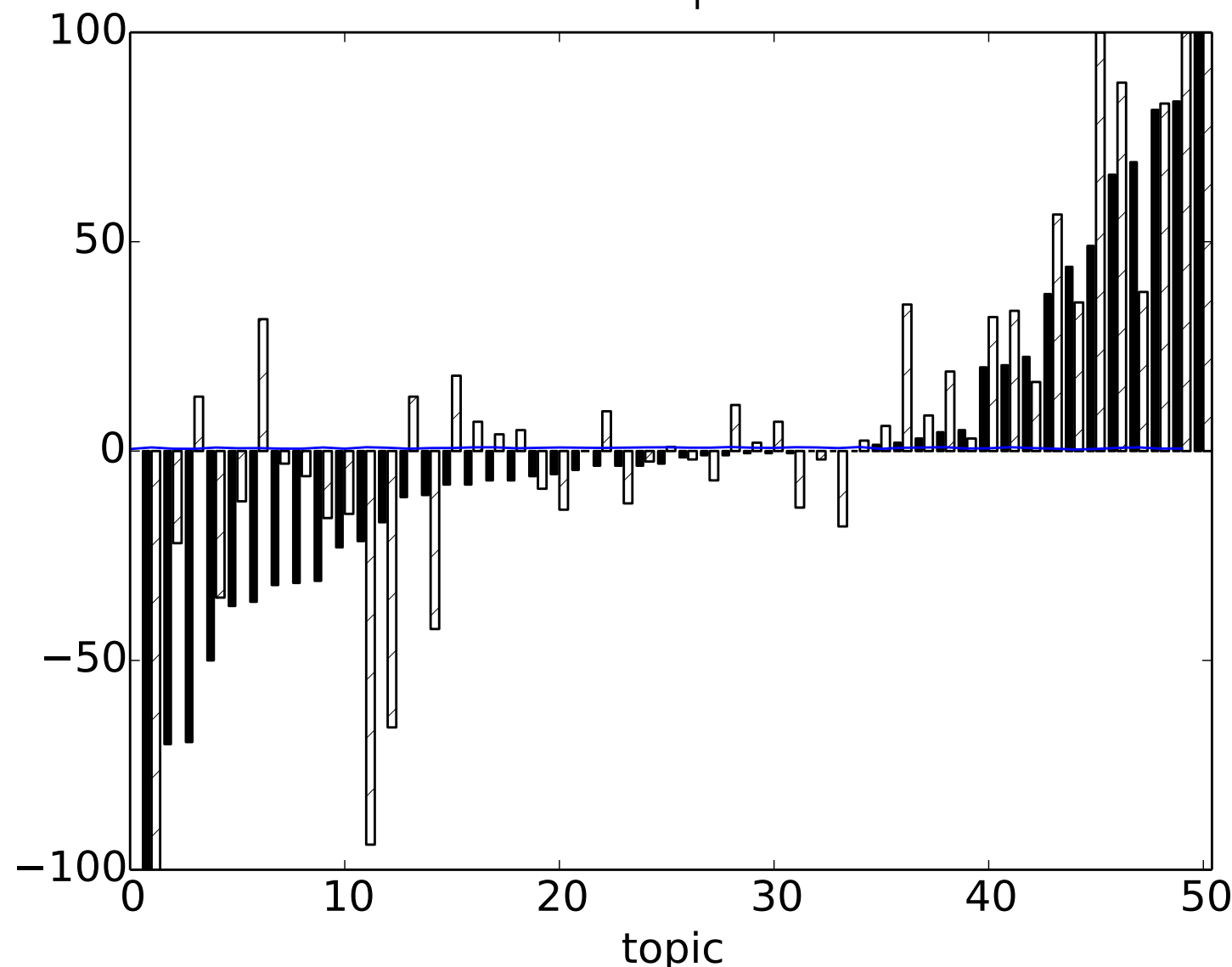


Pearson correlation between the predicted effort and user effort:
0.79 (p-value < 0.01)

Q2: Can we accurately predict when a RLR interface is beneficial?

■ Basic user effort - RLR user effort (difference between user effort on two interfaces)

□ Basic user effort - RLR predicted effort (difference between actual user effort on basic interface and predicted user effort on RLR interface)



- Accuracy of prediction

	P	R	F1
Basic better	0.85	0.55	0.66
RLR better	0.52	1	0.68

Validation of prediction: conclusions

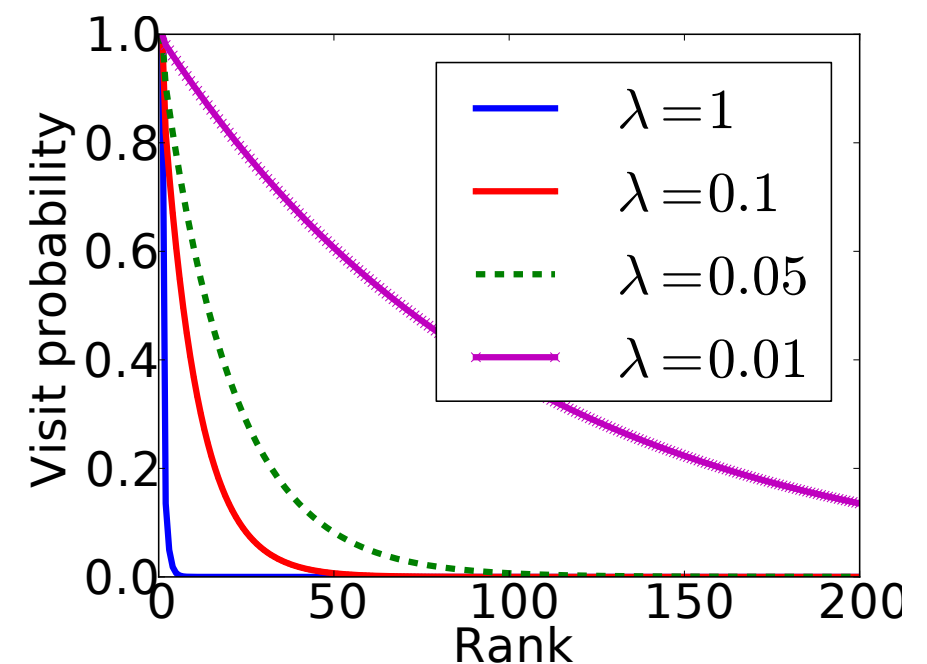
- Our RLR user interaction model is able to accurately predict user effort
- Different interfaces are suitable for different queries (i.e., of different ranking quality)
- Model allows prediction of which interface is most suitable

Whole system evaluation: hypothesised users

- RQs
 - When does an RLR interface help to save user effort compared to a basic interface?
- Study whole system performance under varying conditions:
 - Ranking quality
 - Sublist characteristics
 - User behaviour

Hypothesised user parameter setting

- Intuition: some users are more patient than others
- Parameter 1: continuation
 - at each rank r , draw a decision as a bernoulli trial
 - Bernoulli parameterised by a exponential decay function to approximate the empirical distribution of rank biased visit



1: impatient users
0.01: patient users

$$P(E = 1|r, u) = e^{-\lambda r}$$

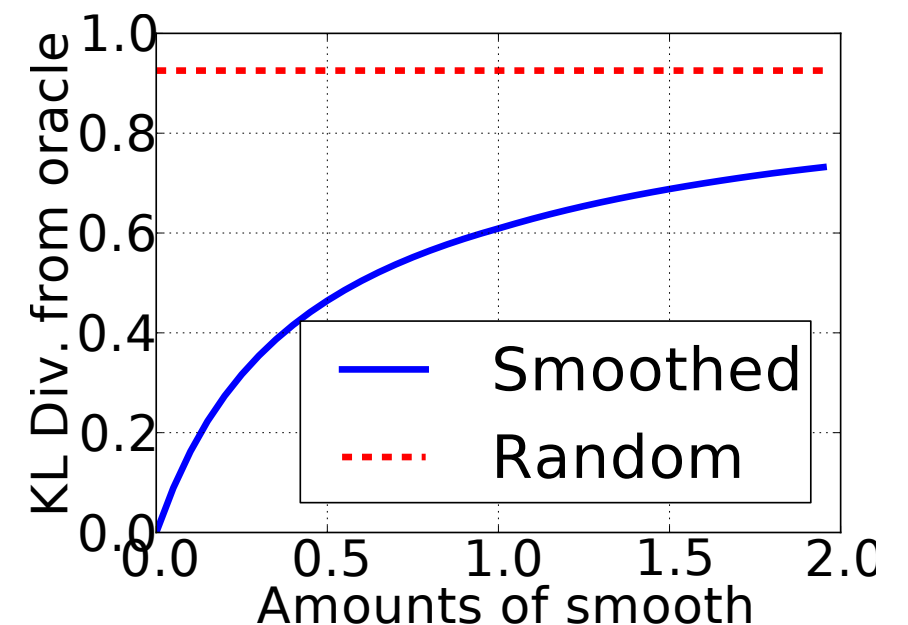
Hypothesised user parameter setting

- Intuition: some users make better selection of sublists than others
- Parameter 2: list selection
 - draw a decision vector from a categorical distribution

$$f_1, \dots, f_k \sim \text{Cat}(K, \vec{c})$$

- setting user prior knowledge of the candidate lists with its conjugate prior

$$\vec{c} \sim \text{Dir}(K, \vec{\alpha})$$



Uniform: no idea what to select

NDCG: informed selection

Factors influencing RLR effectiveness

- Query difficulty for the basic interface (D_q)
 - Efforts needed to accomplish a task with basic interface
- Sublist relevance (R_q)
 - Averaged NDCG score over sublists of a query
- Sublist entropy (H_q)
 - Entropy of relevant documents distributed among sublists
- User accuracy (U)
 - Controlled by the amount of smooth added to the prior of list selection
 - Level 1 (oracle based on NDCG);
 - 15% (level 2), 50% (level 3), 67% (level 4) less accurate compared to level 1
- User task
 - to find 1, 10, or “all” relevant documents

Method

- Fit a generalised linear model (logistic regression)
- **DV**: whether a RLR interface outperforms (i.e., save efforts) a basic interface
- **IVs**: factors outlined above
- **Model selection**: forward and backward selection with Bayesian information criterion (BIC)
- Explain the relation between DV and IVs and their interactions

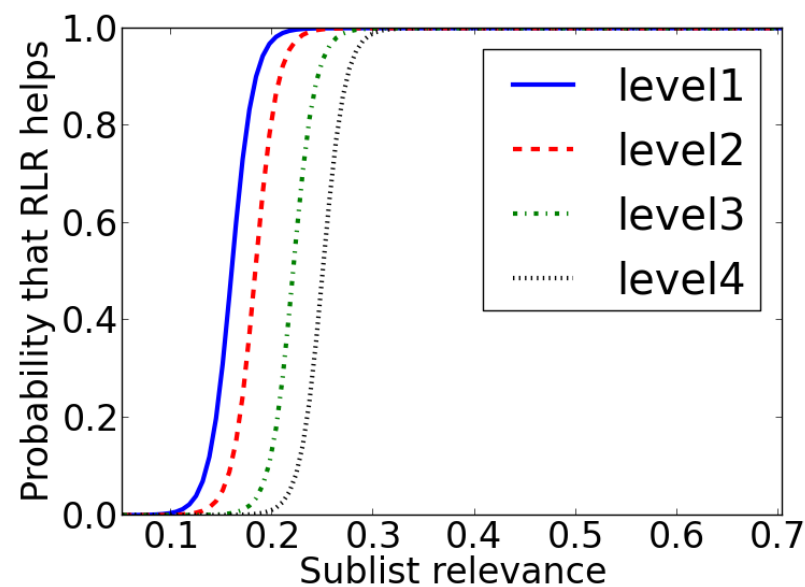
Main effects

Coefficients	Find-1	Find-10	Find-all
intercept	-7.340	-10.437	-0.534
Dq	0.106	-0.069	0.002
U-level2	3.223	-2.131	-5.106
U-level3	1.559	-5.528	-8.014
U-level4	-2.319	-8.194	-8.014
Hq	-1.044	3.635	-1.649
Rq	-	-49.792	114.940
Dq : U-level2	-1.655	-	-
Dq : U-level3	-2.004	-	-
Dq : U-level4	-2.068	-	-
Dq : Hq	1.310	-0.097	-
Dq : Rq	-	3.263	0.091
Hq : Rq	-	13.968	-57.277
Dq : Hq : Rq	-	-0.842	-

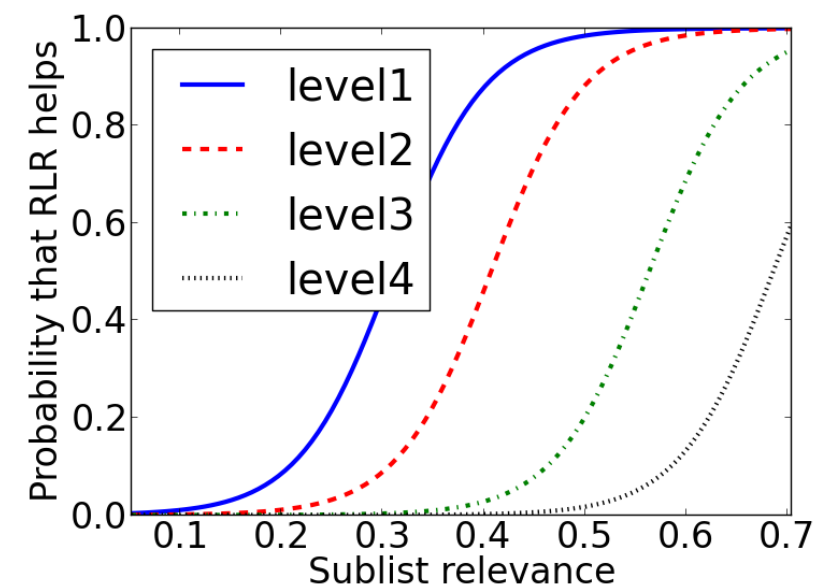
- Find - 1: none of the main effects are significant
- Find - 10 /all: users need to know which sublists to pick
- Find - all: having sublists with relevant documents ranked high is useful.

Interaction effects

Dq:Rq for Find -10



Dq:high; Hq:median

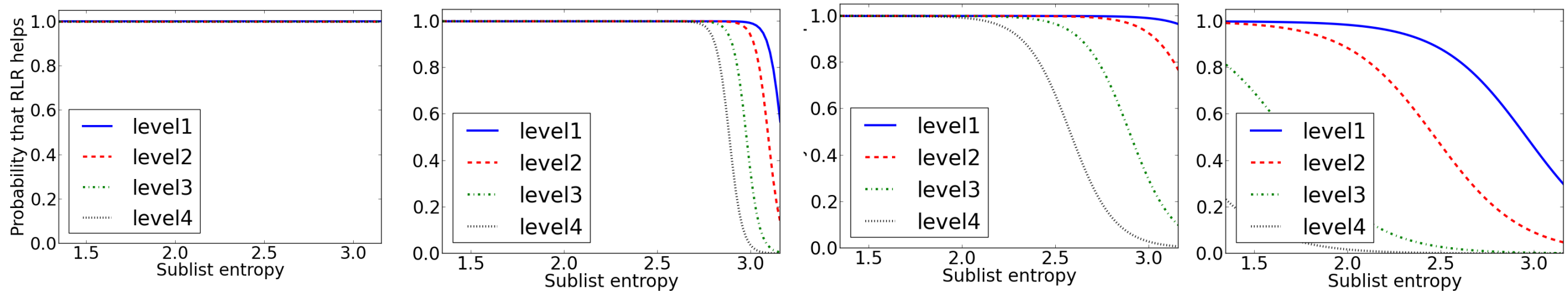


Dq:low; Hq:median

- When query is difficult for basic interface, sublists and users do not need to be very accurate for RLR to be more effective
- When query is easy for basic interface, higher quality of sublists and user accuracy are necessary

Interaction effects

Dq:Rq:Hq for Find -10



(a) Dq:high; Rq:high (b) Dq:high; Rq: low (c) Dq:low; Rq: high (d) Dq:low; Rq: low

- When query is difficult for basic interface, RLR is likely to be beneficial especially when few sublists contain most of the relevant documents
- When query is easy for basic interface, very specific conditions with respect to user accuracy, sublist relevance, and sublist entropy need to be met for RLR to be beneficial.

Relation to traditional metrics

	user effort	predicted effort	Δ user effort
nDCG@10	-0.21	-0.19	0.02
nDCG@all	-0.42*	-0.34	0.00
NRBP	-0.41*	-0.33	0.08
AP	-0.63**	-0.54**	0.02
binary nDCG@10	-0.54**	-0.44**	0.02
binary nDCG@all	-0.72**	-0.59**	0.04
Our model	0.79**	—	-0.49**

Pearson's linear correlation; p-value < 0.01 (*); <0.001 (**)

- Query difficulty alone is not sufficient to predict whether a RLR interface will be beneficial

Whole system evaluation: conclusions

- When ranking quality is low, sub-lists and user sublist selection do not have to be of high quality for RLR to be more effective than basic;
- When ranking quality is high, only under specific conditions RLR may be beneficial, i.e., quality sub-lists, recall oriented task, accurate users;
- Implication for HCIR experiments: are your queries, user tasks, and ranking algorithms appropriate to study the properties of the interface?

Conclusions

- A user interaction model for evaluating search systems with result refinement elements
- By instantiate the model with parameter values derived from real usage data, we have validated the predictive power of our model
- By simulating users with hypothesised parameter values, we can investigate whole system performance under varying conditions concerning ranking quality, interface differences, user types, and task types