

Going social for training, tuning and evaluation

Maarten de Rijke

Sources

- Richard Berendsen, Aleksandr Chuklin, Katja Hofmann, Anne Schuth, Manos Tsagkias, Wouter Weerkamp, Shimon Whiteson

Why do we need to evaluate results in IR?

- Ranking documents is a hard problem
 - Depends on hundreds of criteria
 - Relevance, time, document quality, user background, preferences, task setting, ...
- We need evaluation resources
 - For assessing the quality of the result items or pages
 - For training ranking functions
 - For tuning systems ranking functions

Evaluation strategies in IR

■ Offline

□ Cranfield style

- Fixed set of queries and documents judged by trained people
- Compare ranking systems by how good their ranked lists are

■ Online

□ A/B testing: randomly assign some users to the “control” group and the “treatment” group

- Compare ranking systems by analysing the clicks of the users in the “control” group against those in the “treatment” group

□ Interleaving: present combined lists made out of two rankings

- The system that contributed more clicks is better

Evaluation strategies in IR (2)

■ Supervised

- Have experts label data
- Used the labeled data for training, tuning, evaluation

■ Unsupervised

- Use unlabeled data for training, tuning, evaluation
- More precisely, use **naturally occurring labels** for training, tuning, evaluation
 - Explicit signals such as anchor text, descriptors, hashtags, ...
 - Implicit signals such as clicks, bookmarks, save's, ...

Relative advantages

■ Supervised/unsupervised

- Labeled data scarce
- Never enough training, tuning, testing data

■ Offline/online

- Online schemas are user-based and as a result assumed to give more realistic insights into real system quality
- Offline measurements are easier to reproduce

This lecture

- Use naturally occurring side-products ...
 - ... of user interactions
 - ... of edited or user generated content creation
- for training, tuning and testing purposes



This lecture

■ 1. Pseudo test collections

- Exploit naturally occurring labels for training and tuning purposes and then test on editorial data
- Naturally occurring data: labels assigned to data as part of other activities (anchor text, hashtags, ...)

■ 2. Click models

- Try to infer the quality of the search results based on logs of user actions
- Naturally occurring data: clicks

■ 3. Interleaving

- Try to infer the relative quality of rankers based on examining interactions with combined result lists
- Naturally occurring: clicks

1. Pseudo test collections

Setting the Scene

■ Pseudo test collections

- **For training** (e.g, a *learning to rank* system)
- For evaluation

■ Idea:

- use the rich annotations available in the digital library domain
- generating training material works even with relatively sparse data
 - cold start problem

A bit of background

- Automated test collections for know item search
 - (Azzopardi et al, 2007)
- Linked to documents are relevant to the anchor text
 - (Asadi et al, 2011)
- Documents listed under OpenDir category are relevant to that category
 - (Beitzel et al., 2003)
- **Next couple of slides:** scientific articles sharing keywords are relevant to these keywords
 - (Berendsen et al, 2012)
- Tweets labeled with a hashtag are relevant to (a query derived) that hashtag
 - ...

Outline of Part 1

- **Intro: training an LTR system**
- Idea
- Method
- Experiments
- Results
- Discussion

Learning to rank

■ Given:

- a query, document collection, and features

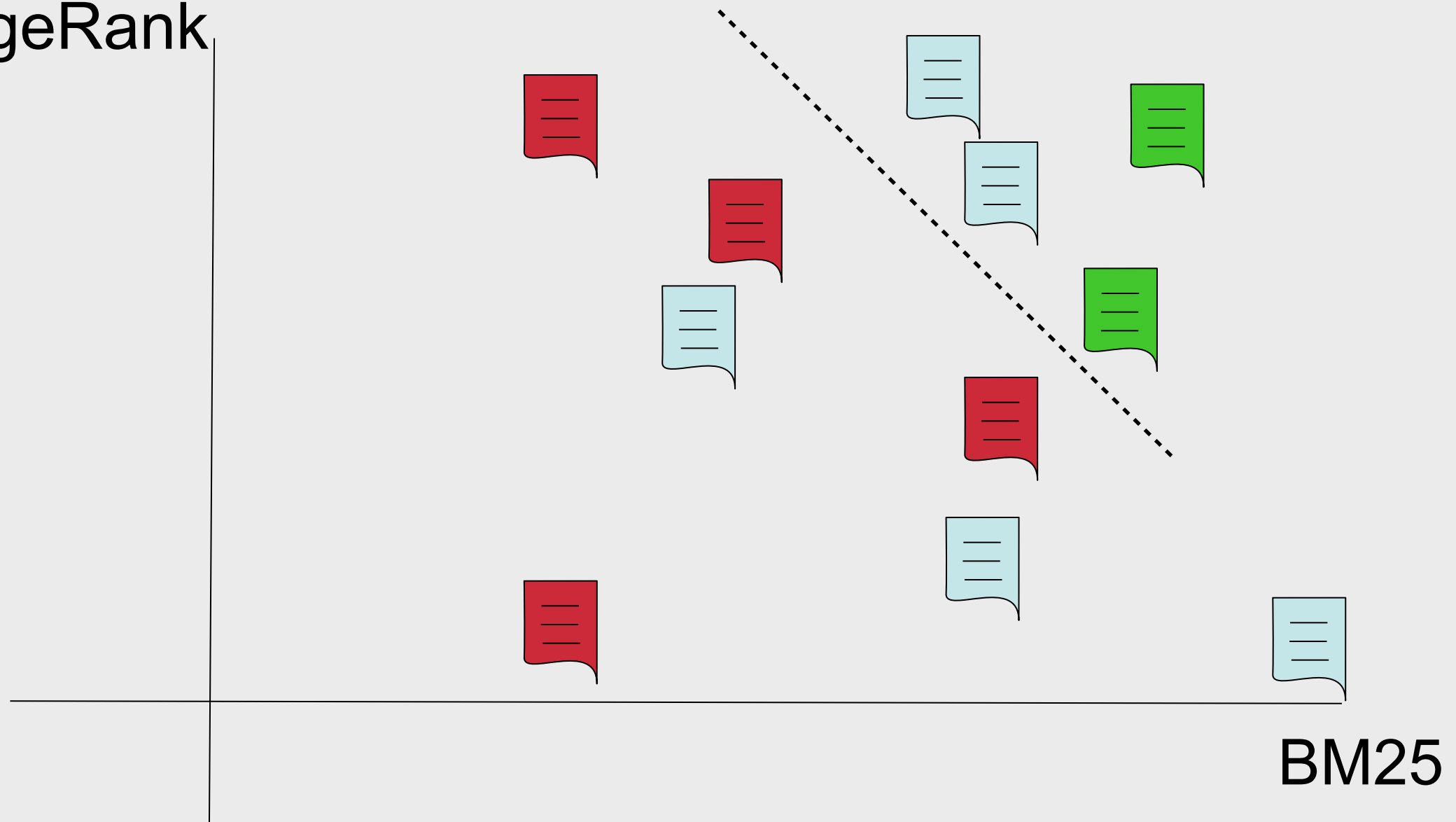
■ Produce an optimal ranking

■ Features:

- Query-document features
- Document features
- Query features

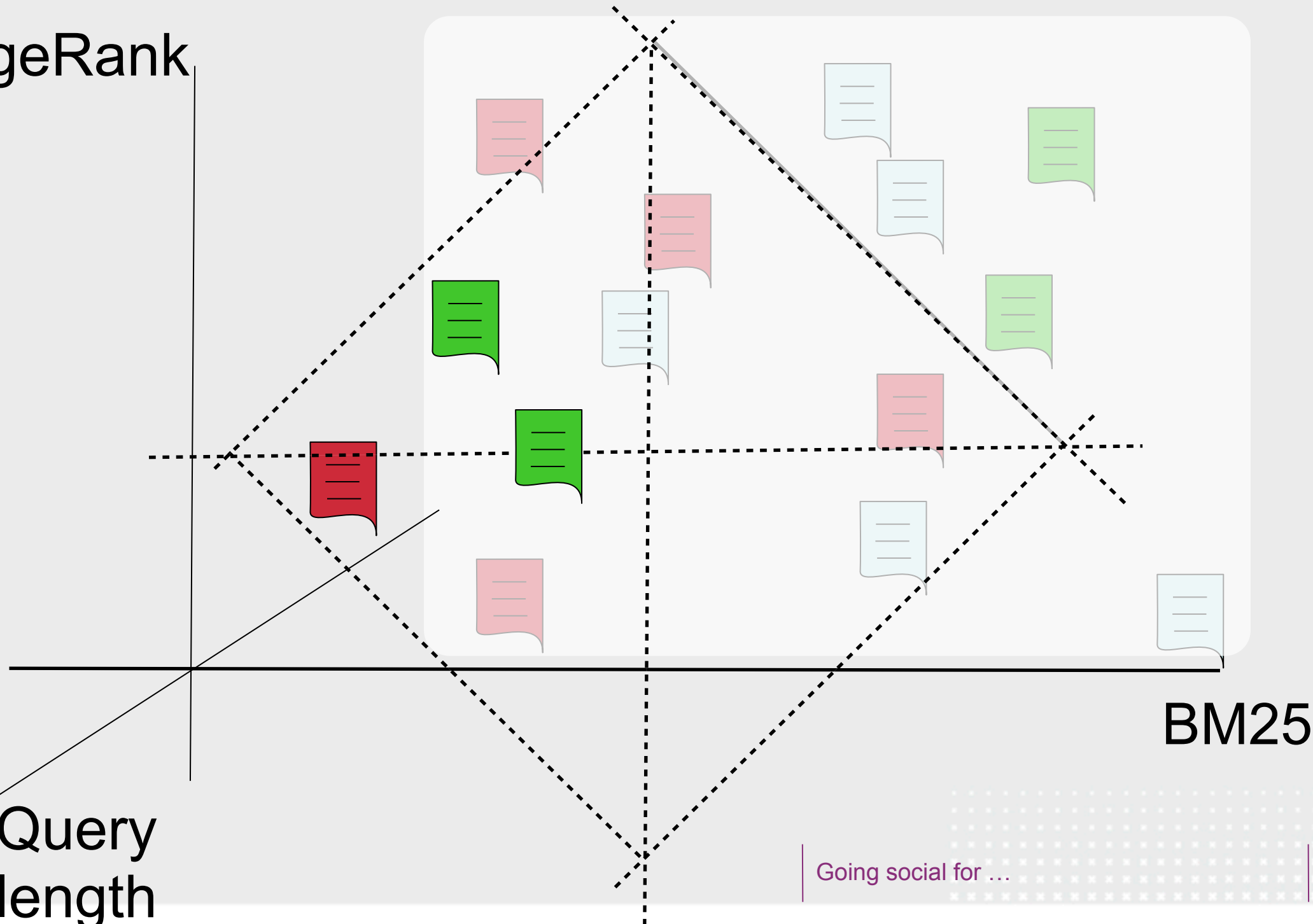
Learning to rank, training phase

PageRank



Learning to rank, a second query

PageRank



Learning to rank, summing up

- Class skewedness, imbalance:
 - Far fewer relevant documents than irrelevant documents
- Need to learn one function for all queries

Wishlist for generating pseudo topics/judgments

■ For training:

- End goal: good LTR performance on new, unseen queries
- Need many training queries
- Diverse queries, to explore the feature space sufficiently during training.
 - For example, no clear preference for just one feature

Idea: Use annotations

- Rich annotations in digital library collections
- Can we use these annotations to:
 - Group documents by topic, using them as pseudo judgments?
 - Generate a query?
- Test this with:
 - A collection of scientific articles

Scientific articles

■ Keywords

- ☐ from a thesaurus
- ☐ free terms

■ Author information, co-authorship

■ Citations

```
<DOCNO>19940100925</DOCNO>
<TITLE>Psychisch kranke Mitarbeiter in Betrieben : die Sichtweise der betrieblichen
Helfer</TITLE>
<TITLE-ENG>Mentally ill employees in companies : the viewpoint of company assistants</TITLE-
ENG>
<AUTHOR>Schubert, Andreas</AUTHOR>
<PUBLICATION-YEAR>1988</PUBLICATION-YEAR>
<LANGUAGE>DE</LANGUAGE>
<CONTROLLED-TERM>psychische Krankheit,Mitarbeiter,Betrieb,Helfer,soziales
Netzwerk,Bezugsperson,Integration</CONTROLLED-TERM>
<CLASSIFICATION>Industriesoziologie, Betriebssoziologie, Arbeitssoziologie, industrielle
Beziehungen,soziale Probleme,Sozialpolitik</CLASSIFICATION>
<TEXT>"Ausgehend von der äußerst problematischen Situation psychisch kranker und
behinderter Menschen auf dem allgemeinen Arbeitsmarkt wird die besondere Bedeutung
innerbetrieblicher Hilfen dargestellt. Dazu wird modellhaft die Situation eines Mitarbeiters mit
'seelischen Problemen' in einem Betrieb skizziert, um somit die potentiellen Bezugspersonen
und damit ein mögliches innerbetriebliches soziales Netzwerk zu kennzeichnen. Die
Fragestellung der dargestellten Untersuchung ist, inwieweit die per Gesetz zur Unterstützung
Behinderter und damit auch psychisch behinderter Mitarbeiter verpflichteten 'betrieblicher
Helfer', diese Funktion tatsächlich wahrnehmen, d.h. inwieweit das Hilfspotential dieser
Gruppe sich umsetzt in ein für den Betroffenen erfahrbares innerbetriebliches soziales
Netzwerk. Dazu werden die Ergebnisse einer schriftlichen Befragung von 144 betrieblichen
Helfern referiert. Als Fazit der Untersuchung muß von einem relativ geringen Kenntnisstand
betrieblicher Helfer bzgl. der Auswirkungen psychischer Krankheit ausgegangen werden, von
negativen Einschätzungen der Leistungs- und Integrationsmöglichkeiten psychisch
behinderter Mitarbeiter und von einer starken Tendenz dieser Gruppe, die Problematik und
damit die Betroffenen auszugrenzen oder, bei betriebsinternen Vorfällen, an betriebliche
Entscheidungsträger wie direkte Vorgesetzte, Personal- und Betriebsleitung 'abzuschieben'.
Da häufig weder interne noch externe Fachleute hinzugezogen werden, ist der Aufbau eines
innerbetrieblichen Netzwerkes als sehr schwierig einzuschätzen. Positive Beispiele belegen
allerdings die Integrationsmöglichkeiten für psychisch Behinderte auch in 'normalen'
Betrieben." (Autorenreferat)</TEXT>
<TEXT-ENG>"Because of the extremely problematical situation of psychologically disturbed
people so far as the job market is concerned this paper stresses the importance of help inside
the concerns. In order to show potential sources of help and thus a possible supportive
network inside a firm a model case of a worker with 'psychological problems' is sketched.
This investigation was aimed at discovering how far the legal obligation to assist handicapped
people inside industrial concerns, and thus also psychologically handicapped workers, is
actually fulfilled by the 'industrial helpers', i.e. how far the potential help offered by these
```

Method

- Select (a tuple of) annotations
 - Use associated documents R_q as pseudo judgments
- Keep annotation tuples with $100 \leq k \leq 1000$ documents
- Generate query from tuple and associated documents

Selecting annotations

- keywords from ontology (CONTROLLED)
- CONTROLLED²
- CONTROLLED x CLASSIFICATION x METHOD
(CT x CL x MT)

Annotations tuples with enough documents

■ Intuitions

- More than 100, because:
 - reliability of LLR when generating a query
 - more training examples for a query probably useful.
- Less than 1000, because:
 - topics should not be overly general

Generating a query from annotation tuple with docs

■ Two methods

- Use all words in annotation tuple
- LLR:
 - Compare word frequencies in R_q with the rest of the corpus:
 - Rank words by how significantly over-represented they are in R_q
 - Using log-likelihood ratio,
 - Cut off at ten words

Research questions

■ For training

- How do our generation methods compare wrt. LTR performance, measured in MAP?
 - to each other?
 - to training on editorial topics?

■ For evaluation

- Do our pseudo test collections rank systems (by their MAP score) similar to editorial test collections; measured in Kendall's tau?

Datasets

■ CLEF Domain Specific Track

- 2007 (28 topics, titles only)
- 2008 (25 topics, titles only)
- English GIRT/CSA corpora,
 - ~170K titles
 - ~40K with abstract

Petras, 2008

Learning to rank: features

■ Query-document features

- Indri runs, Terrier runs

■ Document features

- document length, age of publication
- author features derived from author collaboration graph

■ Query features:

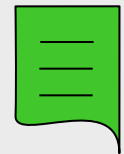
- Query clarity

Learning to rank, training process

- For each pseudo-query
 - select positive and negative training examples
- train an off-the-shelf LTR system, pairwise, linear model, SVM-based [Shalev-Shwartz, 2007]

Learning to rank: training examples

For each query q :



documents in R_q (associated with annotation tuple)



documents from

- bottom of a LM ranked list,
- which are not in R_q ,
- $2 * |R_q|$ (address skewedness)
- following (Asadi et al, 2011)

Results, MAP

Train on	2007	2008
editorial 2007	0,22	0,30
editorial 2008	0,23	0,32
CONTROLLED	0,20	0,27
CONTROLLED, LLR	0,12	0,19
CONTROLLED^2	0,21	0,29
CONTROLLED^2, LLR	0,12	0,20
CT x CL x M	0,13	0,16
CT x CL x M, LLR	0,19	0,26

Pairwise Fisher random test, $\alpha=0.001$

Going social for ...

Results, MAP

- For 2007, we are close, but significant differences
- For 2008, we are close, and in some cases no significant difference
- LLR is worse than using words in annotation tuples, except for CT x CL x M, where it is better

Intuitions about why this should work

- People annotated documents to make them more findable to topics related to the annotation
- Annotations produced by people involved in science
 - These people are also the users of the system: similar notion of relevance.
- While generated training material may contain noise,
 - There are many pseudo queries
 - They are diverse

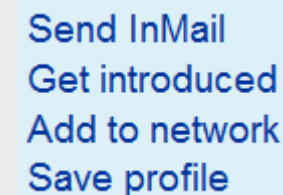
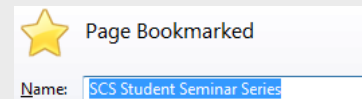
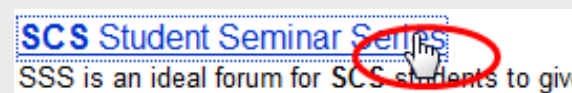
Similarly for microblog search

- Now use hashtags as a labels
- Infer relative importance of pseudo query terms, pseudo relevant documents, ... from social signals
 - Retweets, follower/followee relation, ...

2. Click models

Types of user feedback

- Clickthrough
- Browser action
- Dwell time
- Explicit judgment
- Other page elements



Interpreting clicks

[CIKM 2008 | Home](#)

Napa Valley Marriott Hotel & Spa: Napa Valley, California October 26-30, 2008

[cikm2008.org](#) · [Cached page](#)

[Papers](#) [Program Committee](#)
[Themes](#) [News](#)
[Important Dates](#) [Napa Valley](#)
[Banquet](#) [Posters](#)

Show more results from cikm2008.org

[Conference on Information and Knowledge Management \(CIKM\)](#)

Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...

[www.cikm.org](#) · [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM'02\)](#)

SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.

[www.cikm.org/2002](#) · [Cached page](#)

[ACM CIKM 2007 - Lisbon, Portugal](#)

News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.

[www.fc.ul.pt/cikm2007](#) · [Cached page](#)

[CIKM 2009 | Home](#)

CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...

[www.comp.polyu.edu.hk/conference/cikm2009](#) · [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)

CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...

[cikmconference.org](#) · [Cached page](#)

[CIKM 2004](#)

Identify challenging problems facing the development of future knowledge and information systems, and shape future directions of research by soliciting and reviewing high quality ...

[ir.iit.edu/cikm2004](#) · [Cached page](#)

[CIKM](#)

International Conference on Information and Knowledge Management (CIKM) CIKM Home Page

ACM DL: CIKM 17. CIKM 2008: Napa Valley, California, USA. James G. Shanahan, Sihem

Amer-Yahia ...

[www.informatik.uni-trier.de/~ley/db/conf/cikm/index.html](#) · [Cached page](#)

- Clicks are good...
 - Are these two clicks equally “good”?
- Non-clicks may have excuses:
 - Not relevant
 - Not examined

Click position-bias

- Higher positions receive more **user attention** (eye fixation) and clicks than lower positions.
- This is true even in the extreme setting where the order of positions is **reversed**.
- “Clicks are informative but biased”.
 - Joachims et al, 2007

Clicks as relative judgments

■ “Clicked > Skipped above”

[CIKM 2008 | Home](#)

Napa Valley Marriott Hotel & Spa: Napa Valley, California October 26-30, 2008

[cikm2008.org](#) - [Cached page](#)

[Papers](#) [Program Committee](#)
[Themes](#) [News](#)
[Important Dates](#) [Napa Valley](#)
[Banquet](#) [Posters](#)
[Show more results from cikm2008.org](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)

Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...

[www.cikm.org](#) - [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM'02\)](#)

SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.

[www.cikm.org/2002](#) - [Cached page](#)

[ACM CIKM 2007 - Lisbon, Portugal](#)

News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.

[www.fc.ul.pt/cikm2007](#) - [Cached page](#)

[CIKM 2009 | Home](#)

CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...

[www.comp.polyu.edu.hk/conference/cikm2009](#) - [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)

CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...

[cikmconference.org](#) - [Cached page](#)

[CIKM 2004](#)

Identify challenging problems facing the development of future knowledge and information systems, and shape future directions of research by soliciting and reviewing high quality ...

[ir.iit.edu/cikm2004](#) - [Cached page](#)

[CIKM](#)

International Conference on Information and Knowledge Management (CIKM) CIKM Home Page
ACM DL: CIKM 17. CIKM 2008: Napa Valley, California, USA. James G. Shanahan, Sihem Amer-Yahia ...

[www.informatik.uni-trier.de/~ley/db/conf/cikm/index.html](#) - [Cached page](#)

■ Preference pairs:

- #5>#2, #5>#3, #5>#4

■ Use RankSVM to optimize retrieval function

■ Limitation

- Confidence of judgments

Click models

- Given a set of web search click logs:
 - **Predict clicks:** output the probability of click vectors given a new order of URLs
 - Estimate how good a URL is with regard to the information of the user/query

Uses of click models

■ Optimize the retrieval function

- Use predictions from a click model as an additional signal to be used in the ranking function
- Feature for a learning to rank system

■ Online advertising

- User model for sponsored search auctions
- Click through rate prediction

■ Search engine evaluation

■ Behavioral analysis

Summary Statistics	Navigational	Informational
Expected First Clicked Position	1.34	2.46
Expected Last Clicked Position	1.47	3.52
Examination Depth	1.48	3.61

Designing click models

■ Influentials

- UBM model (Dupret and Piwowarski; User Browsing Model)
- Cascade model (Craswell et al.)
- DBN model (Chapelle and Zhang; Dynamic Bayesian Network)

■ General description

- When a user submits a query q , she gets back 10 results u_1, \dots, u_{10}
- Random variables used to describe user behavior
 - Examination E_k whether the users looked at the k-th document (hidden)
 - Click C_k whether the user clicked on the k-th document (observed)

Designing click models

■ In UBM

$$\square P(E_k = 1 | C_1, \dots, C_{k-1}) = \gamma_{kd}$$

where γ_{kd} is a function of two integer parameters: the current position and the distance to the rank of the previous click

$$\square E_k = 0 \Rightarrow C_k = 0$$

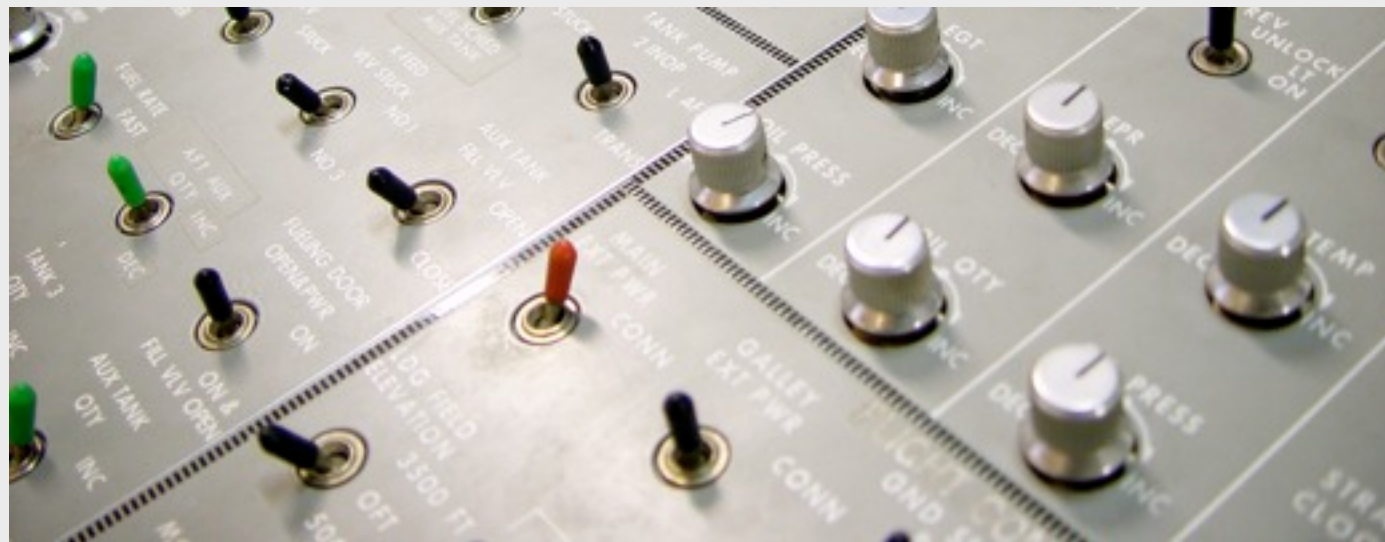
$$\square P(C_k = 1 | E_k = 1) = a_{u_k}$$

a variable for the attractiveness of the document u_k

■ Several click models have so far been mapped to offline metrics: from online to offline

3. Interleaving

Ranking for Search



<http://www.flickr.com/photos/sameli/540933604/>

Task

Combine hundreds of ranking features to get the best ranking for each search task / user

Approach

Today


Offline – use manual annotations for manual tuning or supervised learning,
problems: resources, fidelity, scale

Tomorrow?

Online – learn directly from natural user interactions with the search system

Learning from Natural User Interactions with an IR system

(e.g., clicks on search results)

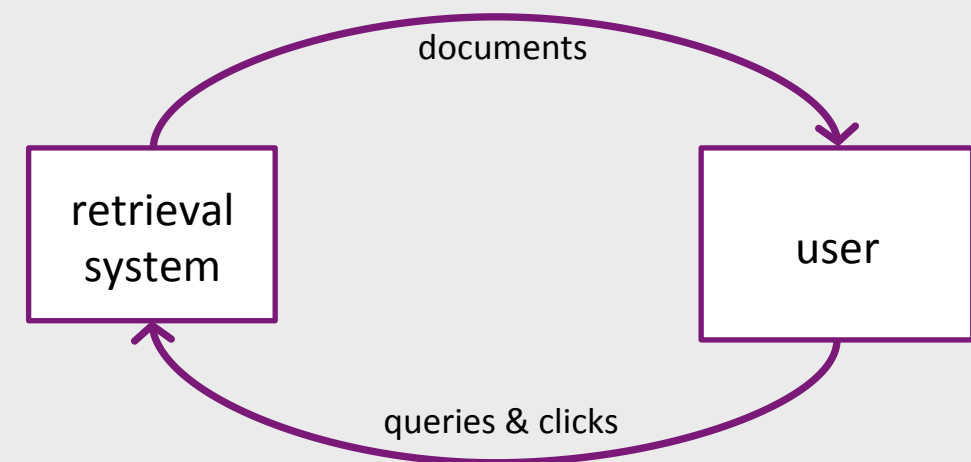
- | | | |
|--|--|---|
| <ul style="list-style-type: none"> ■ Easy to collect while system operates ■ Reflect natural user behavior and preferences ■ Enable online learning |  | <ul style="list-style-type: none"> ■ Noisy ■ Provide only relative preference indications |
|--|--|---|

How can IR systems learn reliably and efficiently from noisy, relative feedback?

Approach – Overview

Reinforcement Learning Approach

- Learn by trying out actions (document lists), and observing feedback
- Follow a listwise learning approach (compare two ranking functions per round)
- Assume independent queries (contextual bandit problem)



- 1 observe query
 - 2 generate result list
 - 3 infer feedback from clicks
 - 4 update retrieval function
- How to reliably infer feedback?
- How to learn efficiently online?

Outline for 3. Interleaved comparison

3.1 Inferring Feedback

Baseline & Motivation

Probabilistic Interleave

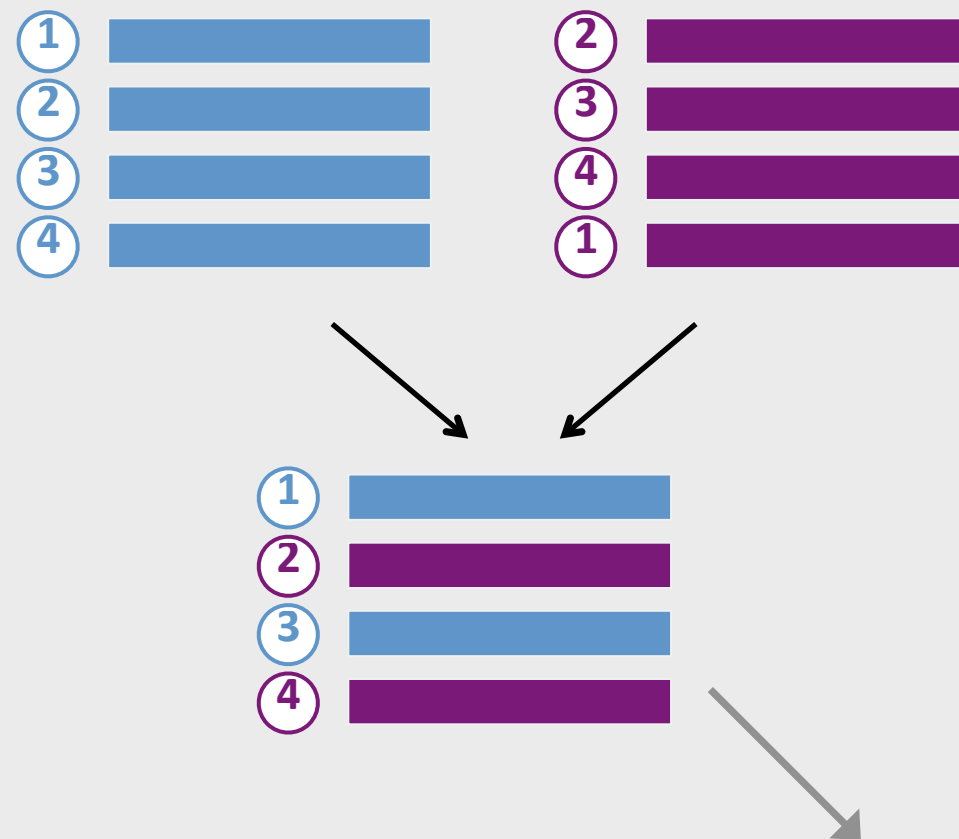
Increasing Efficiency: Marginalization

Increasing Efficiency: Data Reuse

3.2 Online Learning with Data Reuse

3.3 Summary & Outlook

Baseline: Team Draft



Key idea: Keep track of **assignments**
(which list contributed which document)

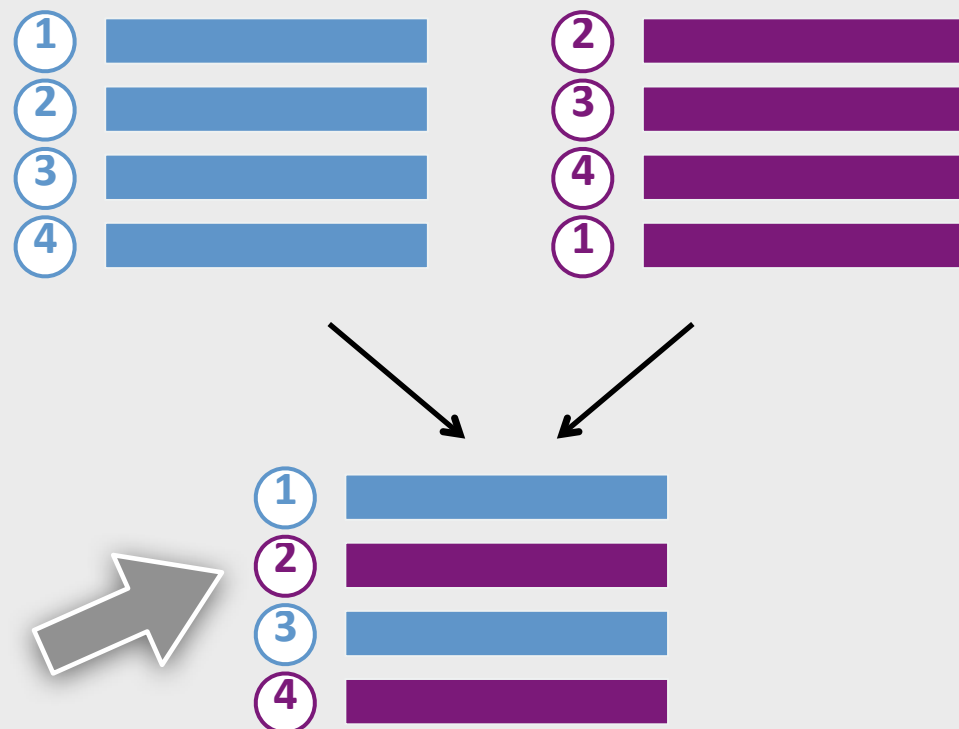
■ Goal: Compare two result lists using click data

■ Procedure:

- 1) **Generate interleaved result list**
(randomize per pair of ranks)
- 2) Observe user clicks
- 3) Credit clicks to original rankers to infer outcome

$$o \in \{-1, 0, +1\}$$

Baseline: Team Draft



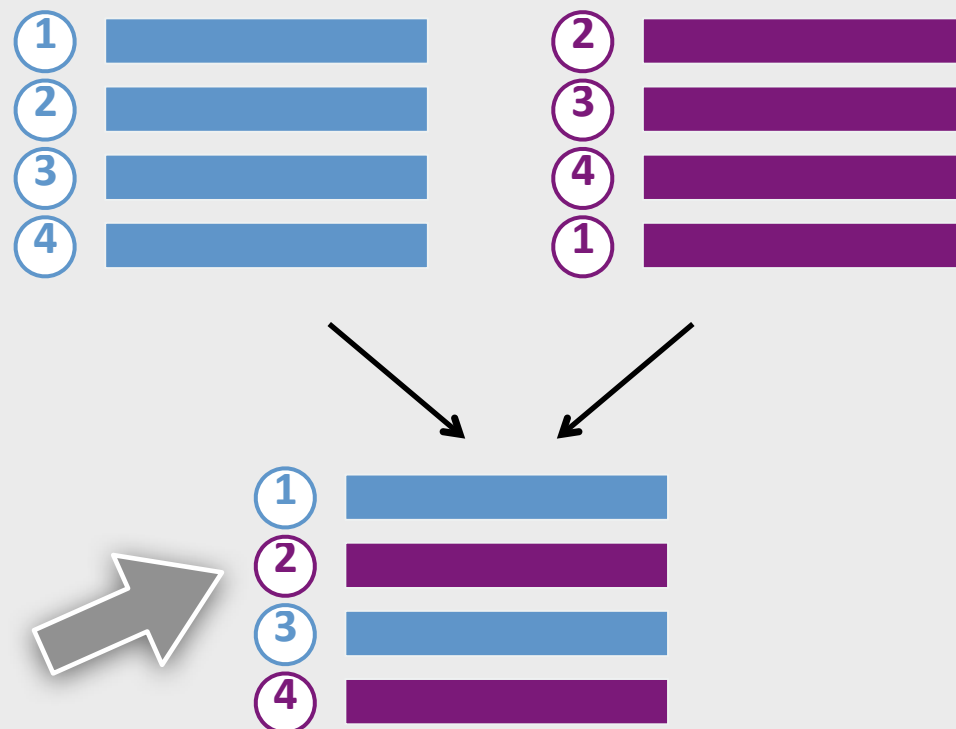
■ Goal: Compare two result lists using click data

■ Procedure:

- 1) Generate interleaved result list (randomize per pair of ranks)
- 2) **Observe user clicks**
- 3) Credit clicks to original rankers to infer outcome

$$o \in \{-1, 0, +1\}$$

Baseline: Team Draft



Does Team Draft accurately measure preferences between result lists?

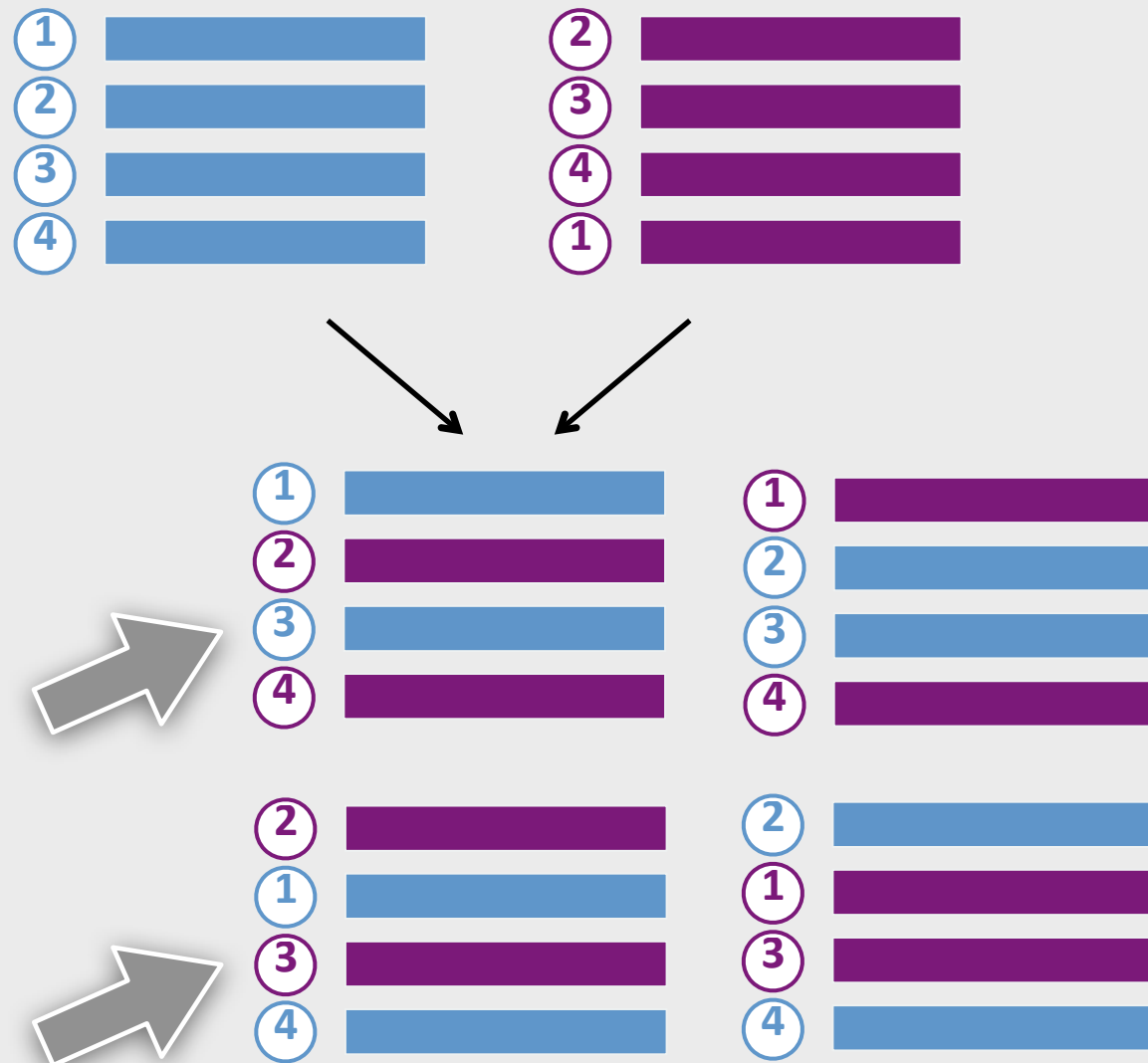
■ Goal: Compare two result lists using click data

■ Procedure:

- 1) Generate interleaved result list (randomize per pair of ranks)
- 2) Observe user clicks
- 3) **Credit clicks to original rankers to infer outcome**

$$o \in \{-1, 0, +1\}$$

Example



- Assume document 3 is relevant (and clicks perfectly reliable)
- Which list should win the comparison?

What Should Interleaved Comparison Methods Measure?

- We propose **fidelity**: in expectation, a preference is detected if and only if a ranker ranks more relevant documents higher
- **Assumption**: positive correlation between clicks and relevance
- **Goal**: develop an interleaved comparison method that exhibits fidelity

Outline for 3. Interleaved comparison

3.1 Inferring Feedback

Baseline & Motivation

Probabilistic Interleave

Increasing Efficiency: Marginalization

Increasing Efficiency: Data Reuse

3.2 Online Learning with Data Reuse

3.3 Summary & Outlook

Our Solution: Probabilistic Interleaving

- Introduce a **probabilistic interleave** method that constructs result list from distributions over documents
- Derive **sound comparison estimators** that exhibit fidelity from the probabilistic interleave model
- Make **estimators more efficient** using (1) marginalization and (2) data reuse and show that fidelity and soundness are maintained

Probabilistic Interleaving

Step 1

Define **probability distributions** (softmax functions) over documents,

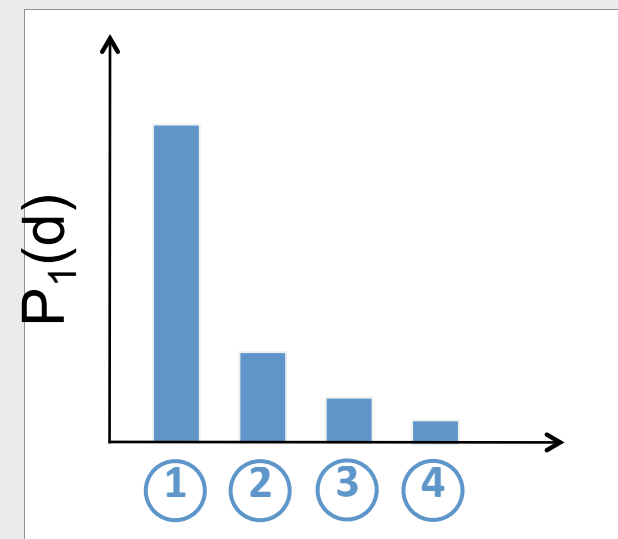
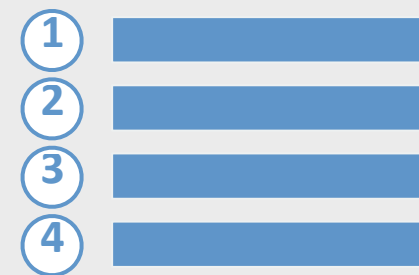
e.g.:

$$P_i(d) = \frac{1}{\sum_{d'} \frac{1}{r_i(d')^\tau}} \frac{1}{r_i(d)^\tau}$$

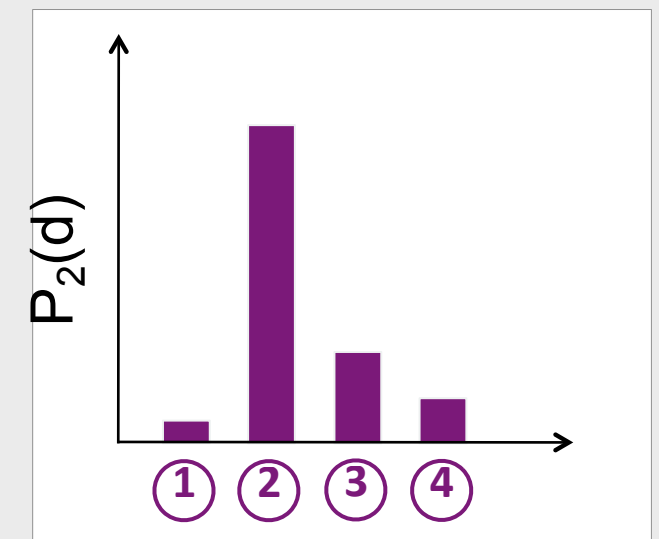
$r_i(d)$ - rank of document d on list l_i

τ - decay parameter

Z - normalizing constant



Example probability distribution over documents for l_1



Example probability distribution over documents for l_2

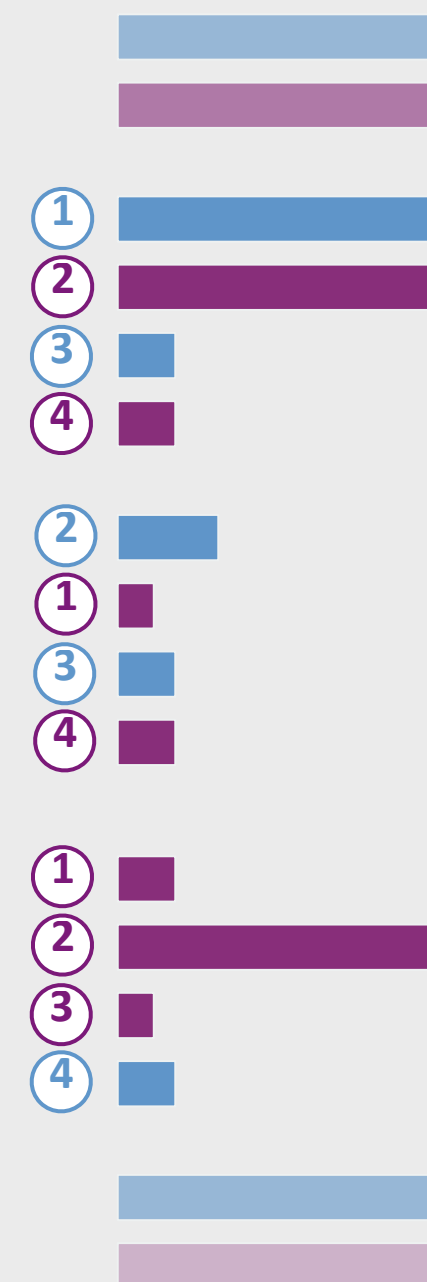
Probabilistic Interleaving

Step 2

Interleave **per rank**:

- 1) Draw softmax function (1, 2) to contribute the next document
- 2) Draw a document from the selected distribution (without replacement)

- ➡ Any permutation of candidate documents is possible (has non-zero probability)
- ➡ Any assignment is possible (i.e., each ranker can contribute each document)



Naïve Estimator

- Comparison outcomes can be computed as under team draft
- The mean of sample outcomes is an unbiased estimator of the expected comparison outcome:

$$E[O] = \sum_{o \in O} \approx \frac{1}{n} \sum_{i=1}^n o_i$$

O - comparison outcome

n - sample size

Result: interleaved comparison method that exhibits fidelity ... what about efficiency?

Outline for 3. Interleaved comparison

3.1 Inferring Feedback

Baseline & Motivation

Probabilistic Interleave

Increasing Efficiency: Marginalization

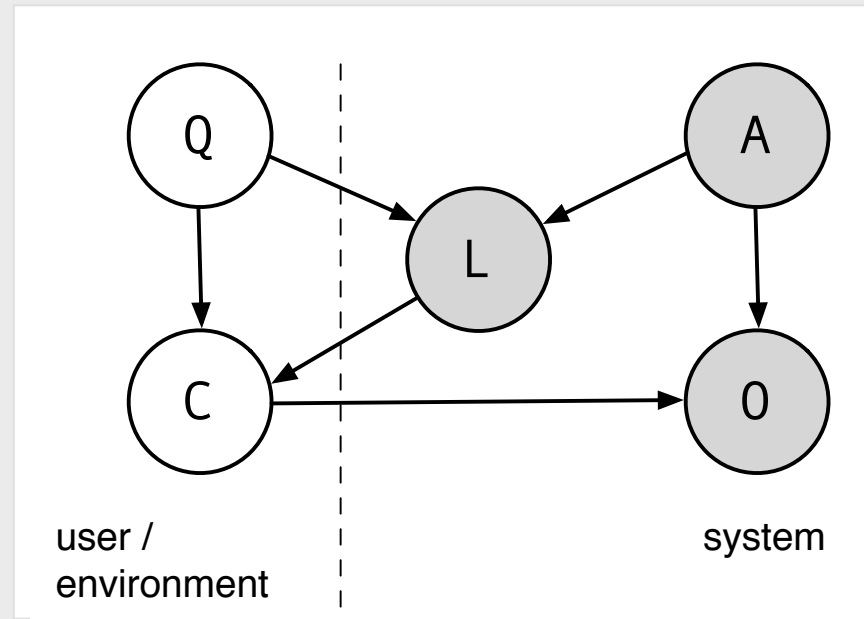
Increasing Efficiency: Data Reuse

3.2 Online Learning with Data Reuse

3.2 Summary & Outlook

Marginalized Estimator

- Based on a model of the probabilistic interleave process



A - assignment
C - clicks
L - list
O - outcome
Q - query

- Estimate comparison outcomes by marginalizing over all possible assignments:

$$E[O] \approx \frac{1}{n} \sum_{i=1}^n \sum_{\mathbf{a} \in \mathbf{A}} \sum_{o \in O} o P(o | \mathbf{a}, \mathbf{c}_i) P(\mathbf{a} | \mathbf{l}_i, q_i)$$

Evaluation – Simulation Framework

Use simulation framework to simulate user interactions under varying conditions from annotated data sets and click models [1]

- Use fully annotated **learning to rank data set** (MSLR) to simulate user interaction and assess comparison methods
- Probabilistic **click model**: assume Dependent Click Model [2] and define click and stop probabilities based on relevance grades provided with the data sets

[1] K. Hofmann, S. Whiteson, and M. de Rijke. *Balancing exploration and exploitation in learning to rank online*. ECIR'11, 2011.

[2] F. Guo, C. Liu, Y. M. Wang. *Efficient multiple-click models in web search*. WSDM '09, 2009

Accuracy

- Run interleaved comparisons for a large number of ranker pairs

(all 18k+ pairs based on individual features provided with the data set, over 1,000 queries)

- Measure accuracy against ground-truth NDCG difference

Run	Accuracy
team draft	0.898
probabilistic interleave	0.914
balanced interleave	0.881
document constraint	0.857

Accuracy

- Run interleaved comparisons for a large number of ranker pairs

(all 18k+ pairs based on individual features provided over 1,000 queries)

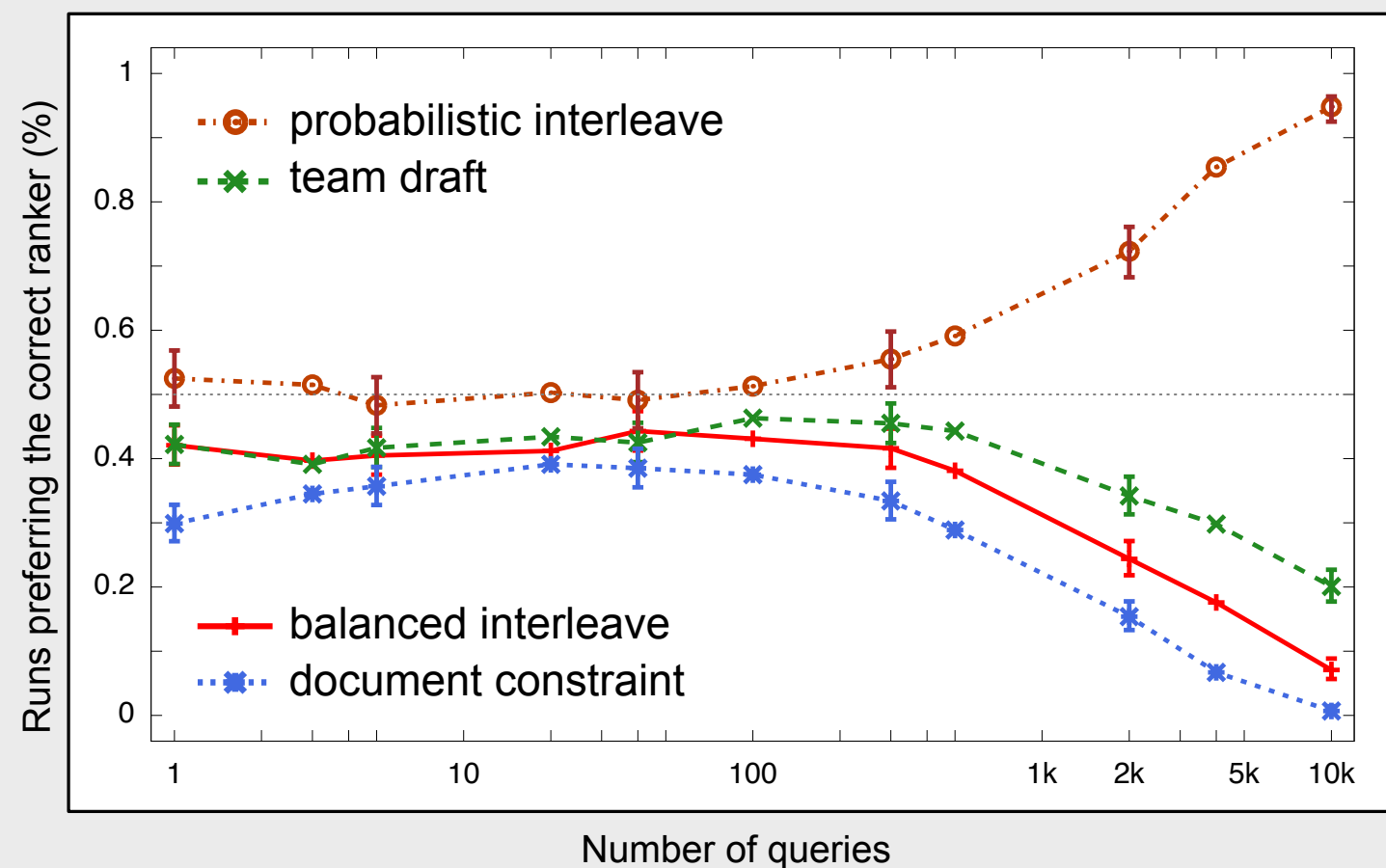
- Measure accuracy against ground-truth NDCG

Probabilistic
interleave compares
rankers more
accurately than
previous methods

Run	Accuracy
team draft	0.898
probabilistic interleave	0.914
balanced interleave	0.881
document constraint	0.857

Robustness to noise

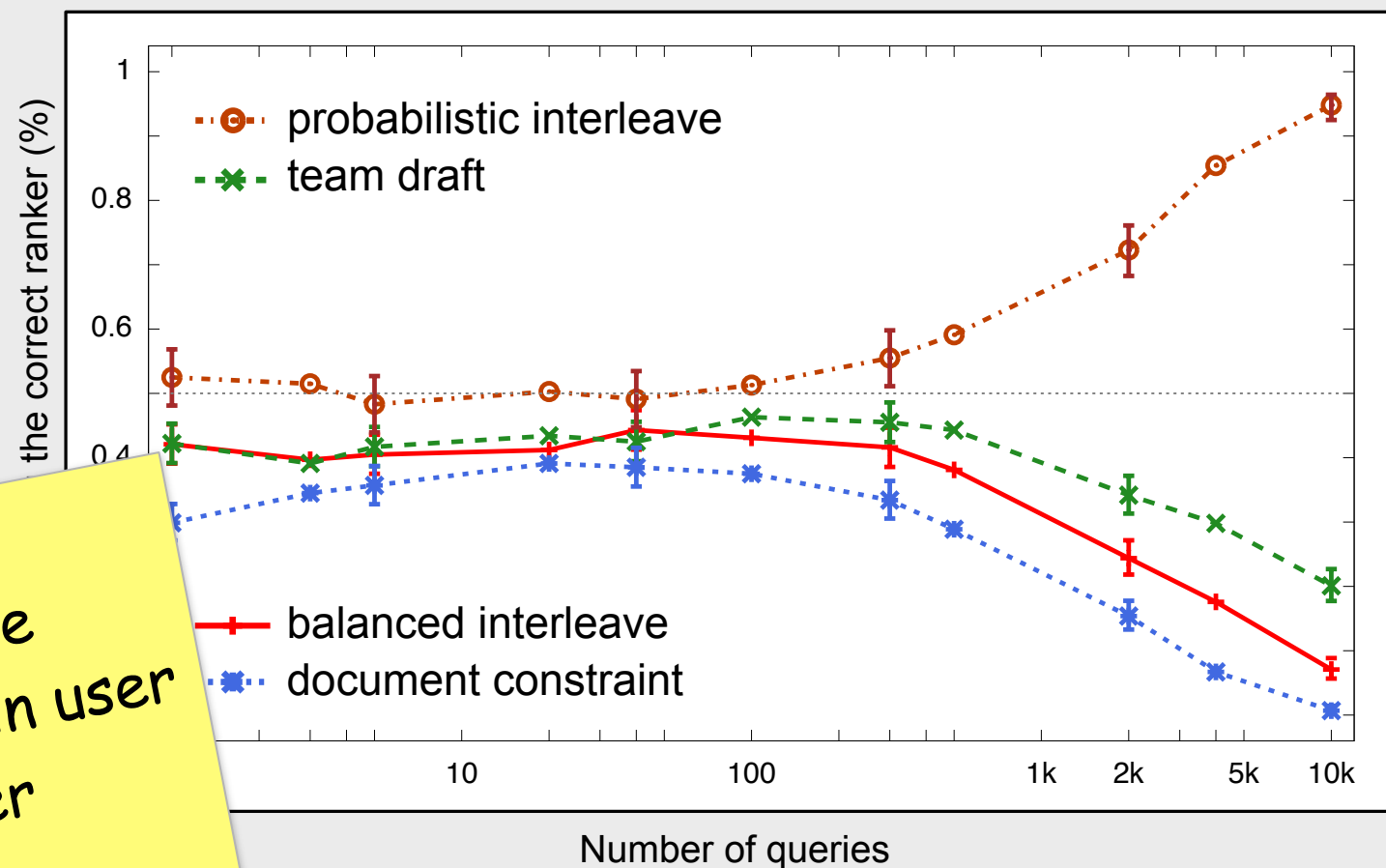
- Compare “difficult” ranker pairs
- Simulate noisy click feedback
- Measure accuracy after n queries (averaged over 25 runs x 5 folds)



Robustness to noise

- Compare “difficult” ranker pairs
- Simulate noisy click feedback
- Measure accuracy after runs x

Probabilistic interleave is more robust to noise in user clicks than other methods



Outline for 3. Interleaved comparison

3.1 Inferring Feedback

Baseline & Motivation

Probabilistic Interleave

Increasing Efficiency: Marginalization

Increasing Efficiency: Data Reuse

3.2 Online Learning with Data Reuse

3.3 Summary & Outlook

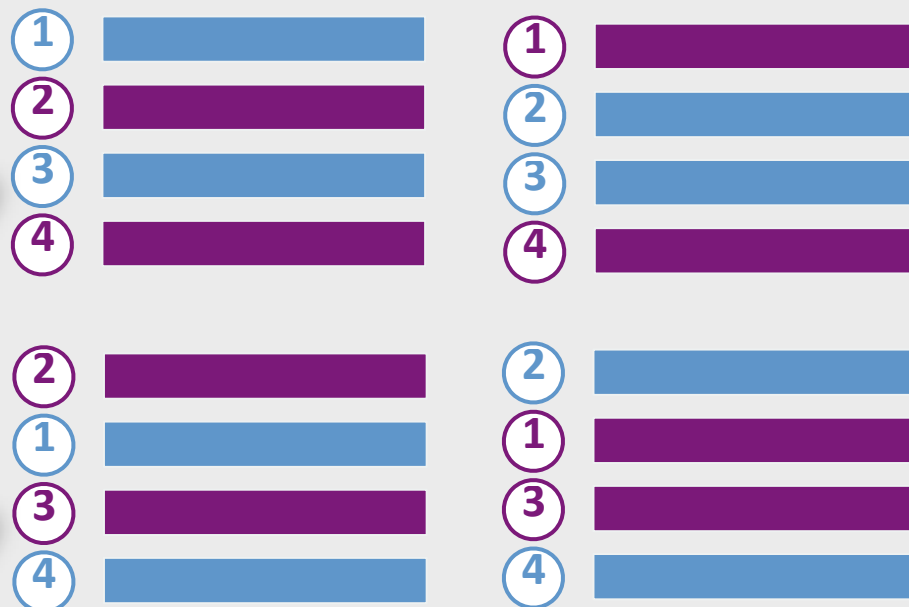
Data Reuse for Interleaved Comparisons

- So far, comparisons are based on “live” samples; data is used once and then discarded
 - Ranker comparisons are limited by the amount of search engine traffic
 - New, risky ideas have to be tested using different evaluation setup
- Instead, can we reuse historical interaction data to compare new rankers?

How can interaction data be reused?

- Example: team draft – need to observe exactly the interleaved list that would be generated from new rankers

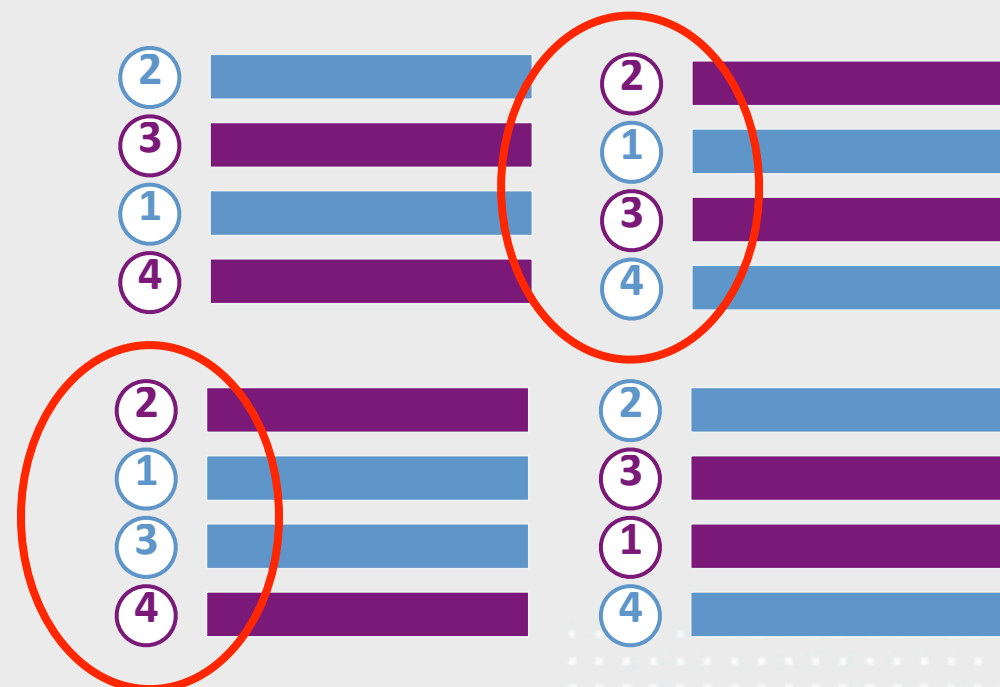
Previously observed data



New target lists



Possible lists under the target lists

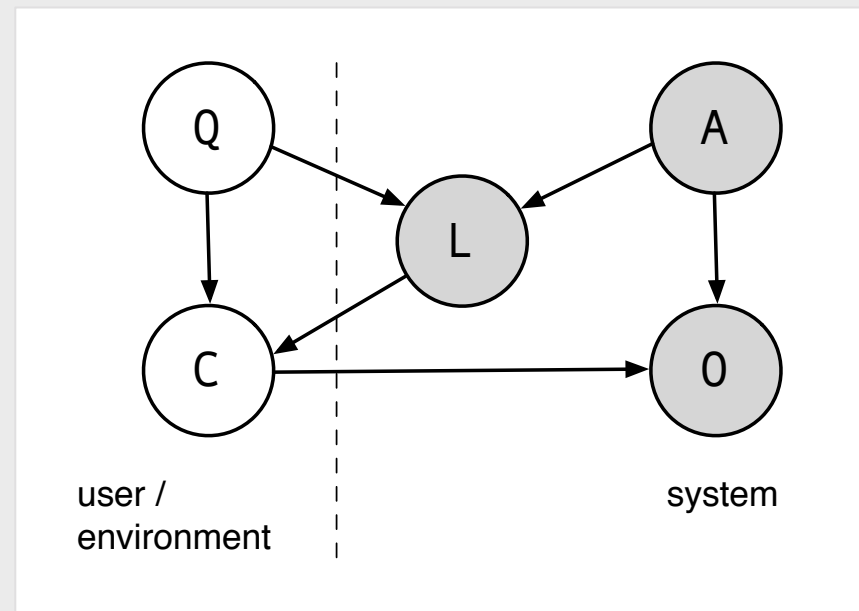


Data Reuse with Probabilistic Interleave

- **Goal:** given data collected under source distribution P_S , infer comparison outcome $E_T[O]$ that would be observed under P_T
- **Insight 1:** Probabilistic Interleave can be applied to arbitrary result lists
- But this may introduce bias ($\hat{E}_{PI} \neq E_T[O]$)
- **Insight 2:** because we defined comparison outcomes probabilistically, we can apply importance sampling to compensate for differences between distributions

Source and Target Distributions

- P_S and P_T have the same event space
- Under different ranker pairs, only L differs, and it is known for both distributions



A - assignment
C - clicks
L - list
O - outcome
Q - query

Importance sampling

- Apply importance sampling, and show that the resulting estimator is a sound w.r.t. $E_T[O]$
- Resulting importance sampling estimator:

$$E_T[O] \approx \frac{1}{n} \sum_{i=1}^n \sum_{\mathbf{a} \in \mathbf{A}} \sum_{o \in O} o P(o|\mathbf{a}, \mathbf{c}_i) P(\mathbf{a}|\mathbf{l}_i, q_i) \frac{P_T(\mathbf{l}_i|q_i)}{P_S(\mathbf{l}_i|q_i)}$$

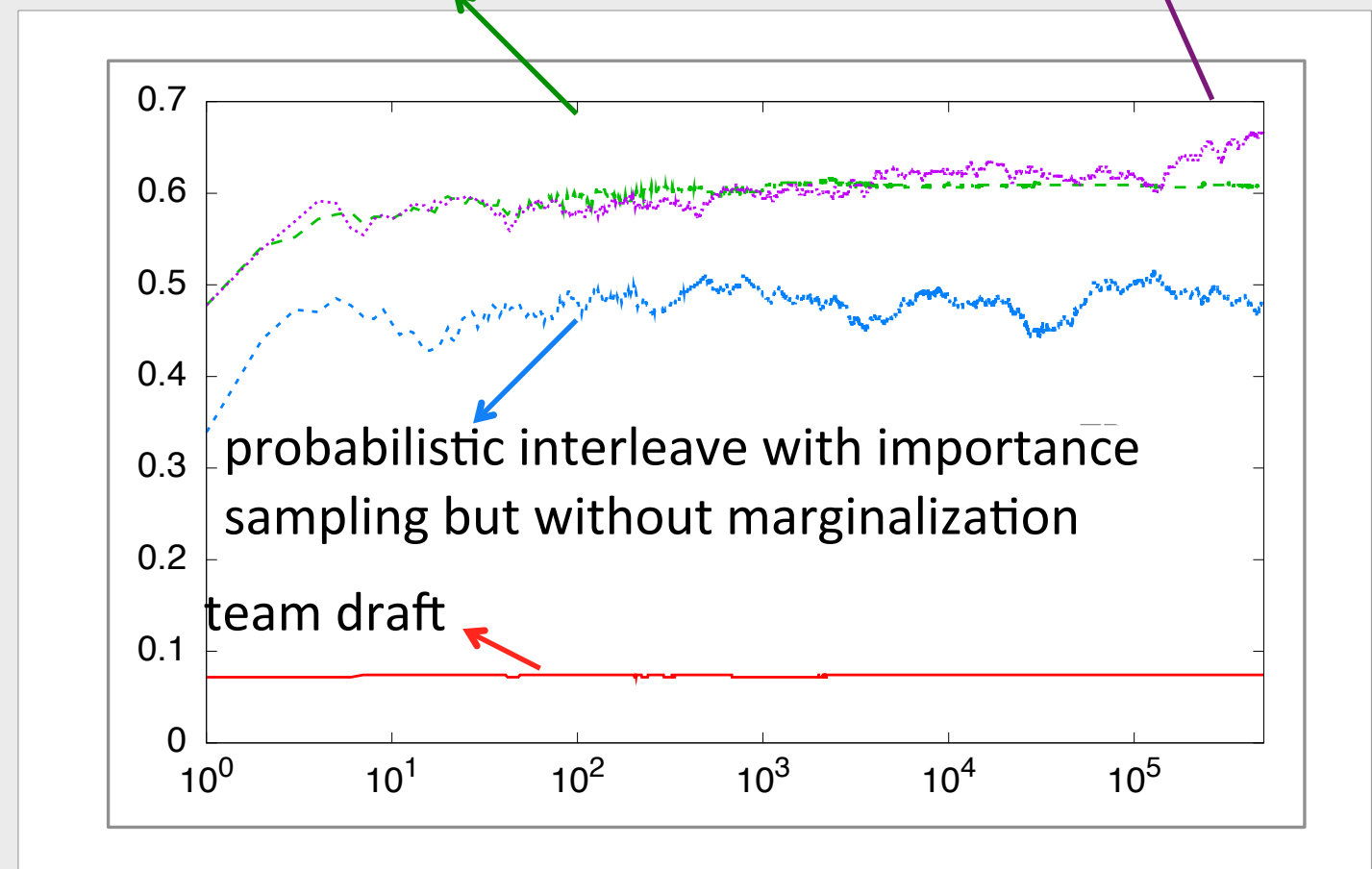
Probability of the observed list under the **target distribution**

Probability of the observed list under the **source distribution**

Experiment & Results

- Experiment:
randomly sample
query, source, and
target rankers; run
200k comparisons
(1000 times)

probabilistic interleave with importance sampling
probabilistic interleave without importance sampling



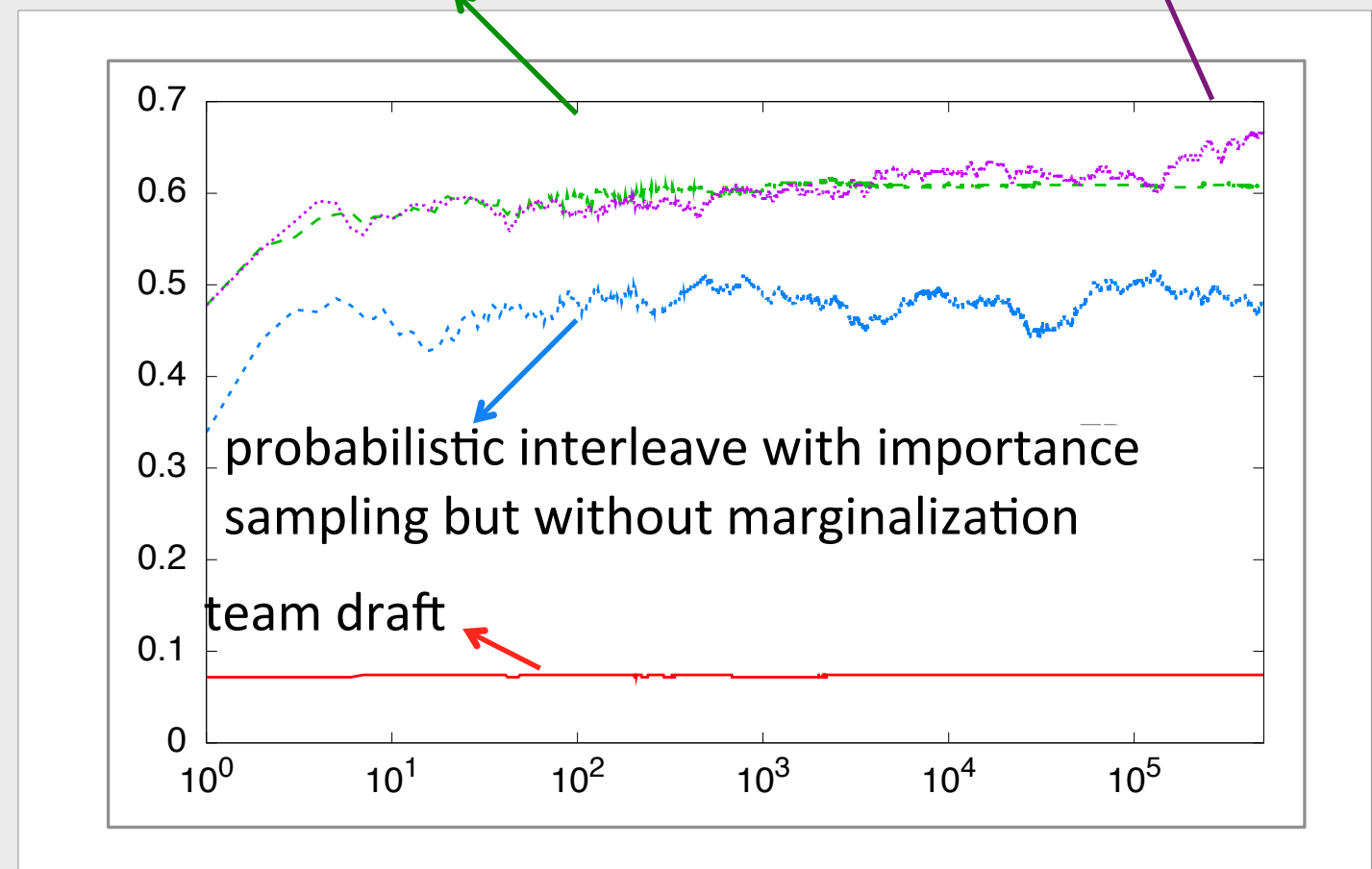
Results: comparison accuracy over 1000 repetitions

Experiment & Results

- Experiment:
randomly sample
query, source, and
target ranks

Probabilistic interleave is effective with and without importance sampling; effect of importance sampling visible after 100k samples

probabilistic interleave with importance sampling
probabilistic interleave without importance sampling



Results: comparison accuracy over 1000 repetitions

Outline for 3. Interleaved comparison

3.1 Inferring Feedback

Motivation

Probabilistic Interleave

Increasing Efficiency: Marginalization

Increasing Efficiency: Data Reuse

3.2 Online Learning with Data Reuse

3.3 Summary & Outlook

Research Questions

- Can historical interaction data be used to improve the **online performance** of online learning to rank for IR methods?
- How does **click noise** affect online learning under data reuse?

Baseline: Dueling Bandit Gradient Descent

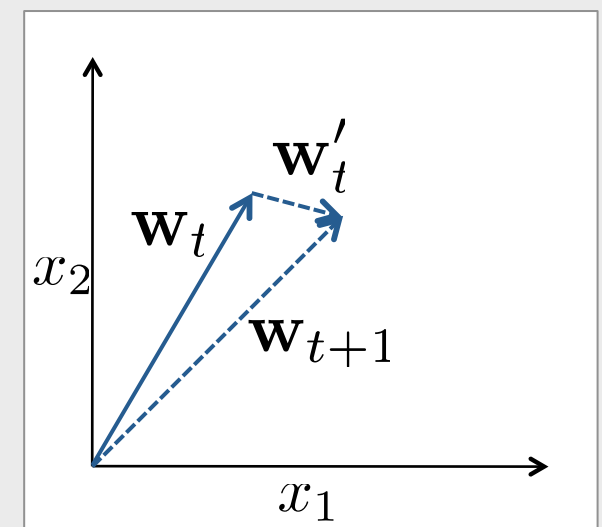
- Input: feature vectors for all candidate documents

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D\}$$

- Output: complete ranking of the candidate documents (by a score $S = \mathbf{w}\mathbf{x}(q, d)$)

Approach

- Maintain a current “best” ranking function
- On each incoming query:
 - Generate a new candidate ranking function
 - Compare to current “best”
 - If candidate is better, update “best” ranking function



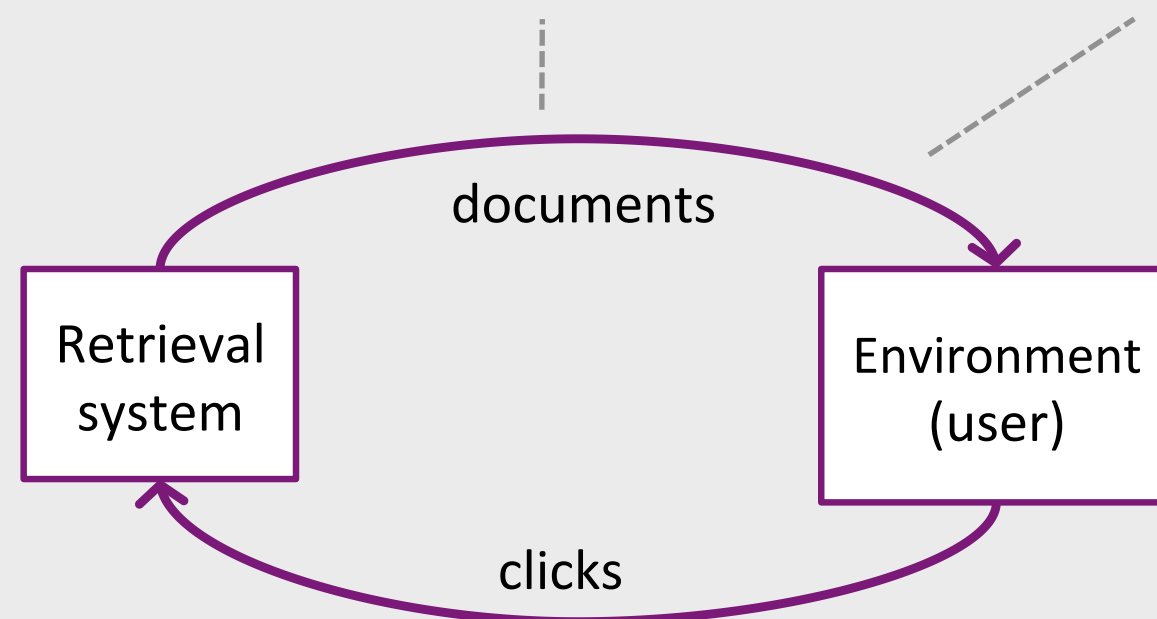
Candidate Pre-selection (CPS)

- **Idea:** use historical data to pre-select promising candidate rankers
- **Approach:**
 - Generate a pool of candidate rankers
 - Compare rankers using historical data to select the most promising one
 - Use winning candidate in live comparison

Evaluation – Simulation Framework

Evaluate on 9 data sets (LETOR)
Here: results for NP2003

Measure **cumulative reward**: quality
of all result lists presented to the
(simulated) user



Simulate user clicks [2] and vary
level of noise in click model:
“perfect”, “navigational”,
“informational”

[1] **K. Hofmann**, S. Whiteson, and M. de Rijke. *Balancing exploration and exploitation in learning to rank online*. ECIR'11, 2011.

[2] F. Guo, C. Liu, Y. M. Wang. *Efficient multiple-click models in web search*. WSDM '09, 2009

Results

Click Model	Live data only		With data reuse	
	TD	PI	PI + history	CPS
perfect	98.15	88.67	89.05	108.04
navigational	85.07	74.77	80.47	107.58
informational	52.88	45.38	64.25	103.34

Results: online performance after 1000 queries / interactions, averaged over 5 folds x 25 repetitions

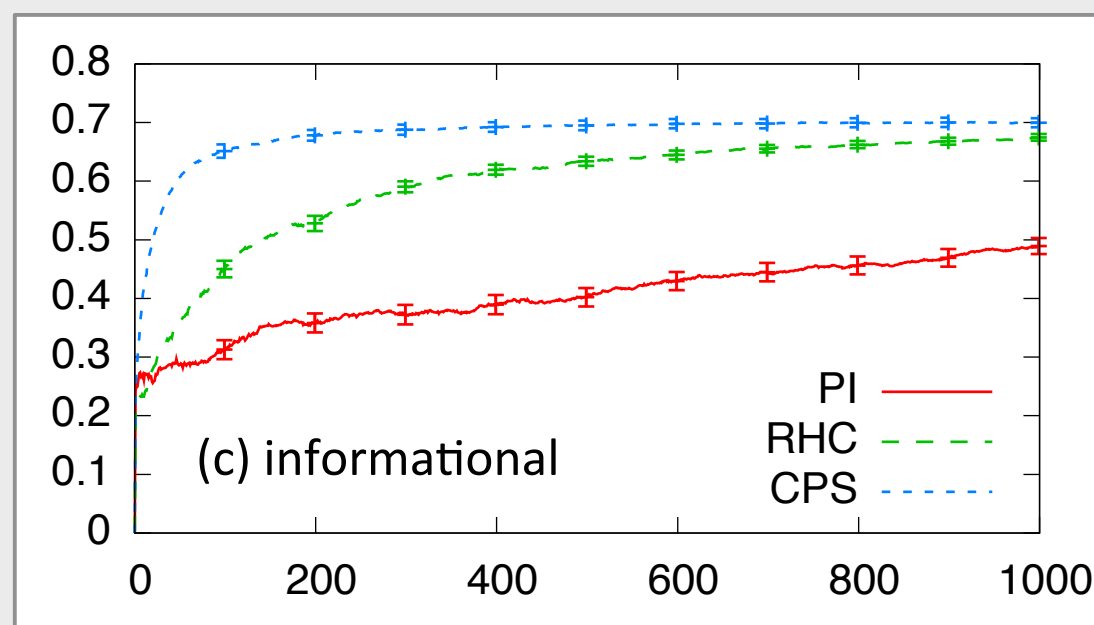
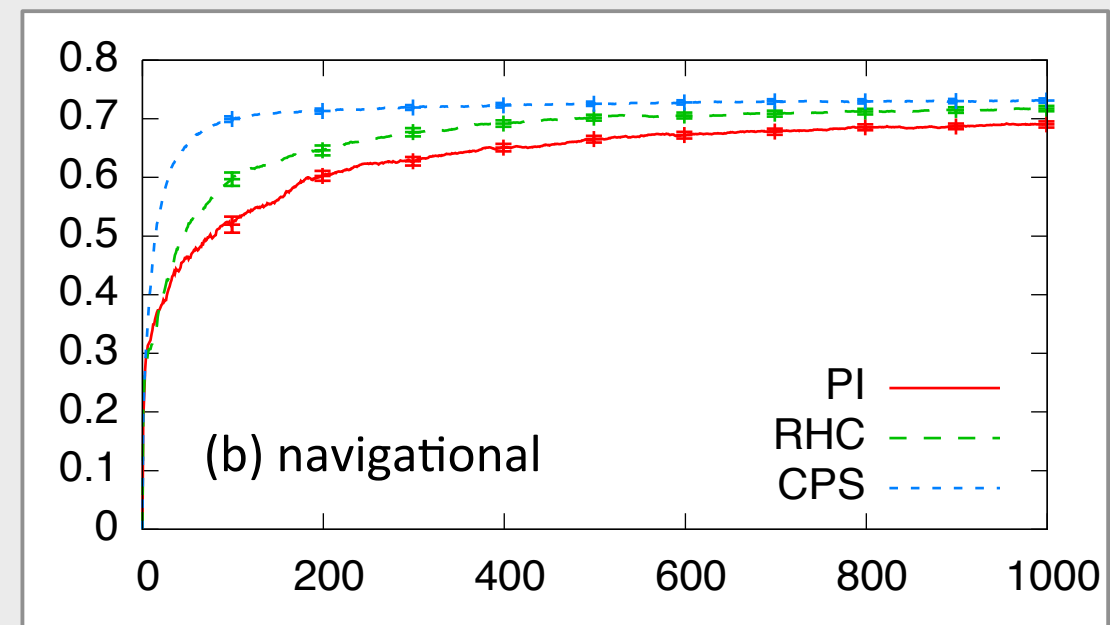
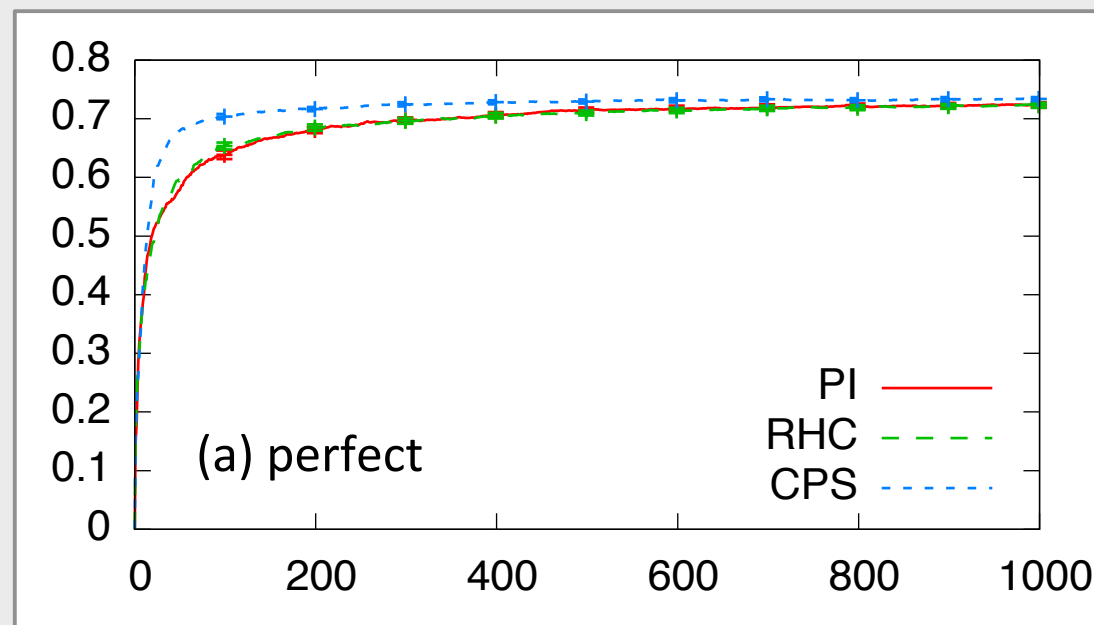
Results

Click Model	Live data only		With data reuse	
	TD	PI	PI + history	CPS
perfect	98.15	88.67	89.05	108.04
navigational	85.07		80.47	107.58
informational			64.25	103.34

Results: online performance averaged over 5 folds x 25 repetitions

CPS performs significantly better than existing methods under reliable and noisy user feedback

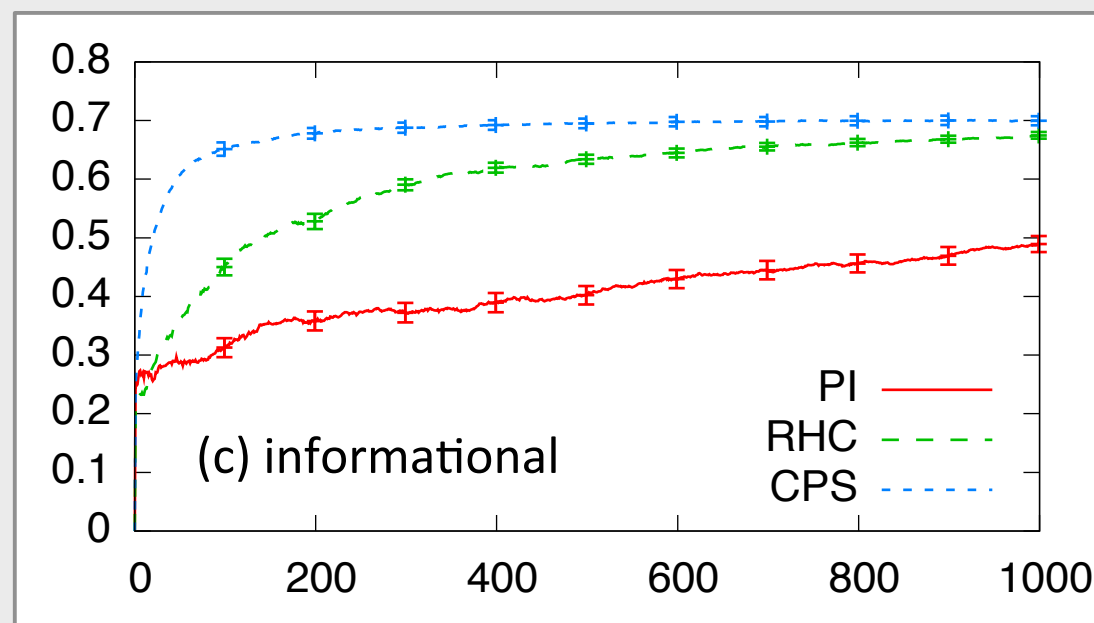
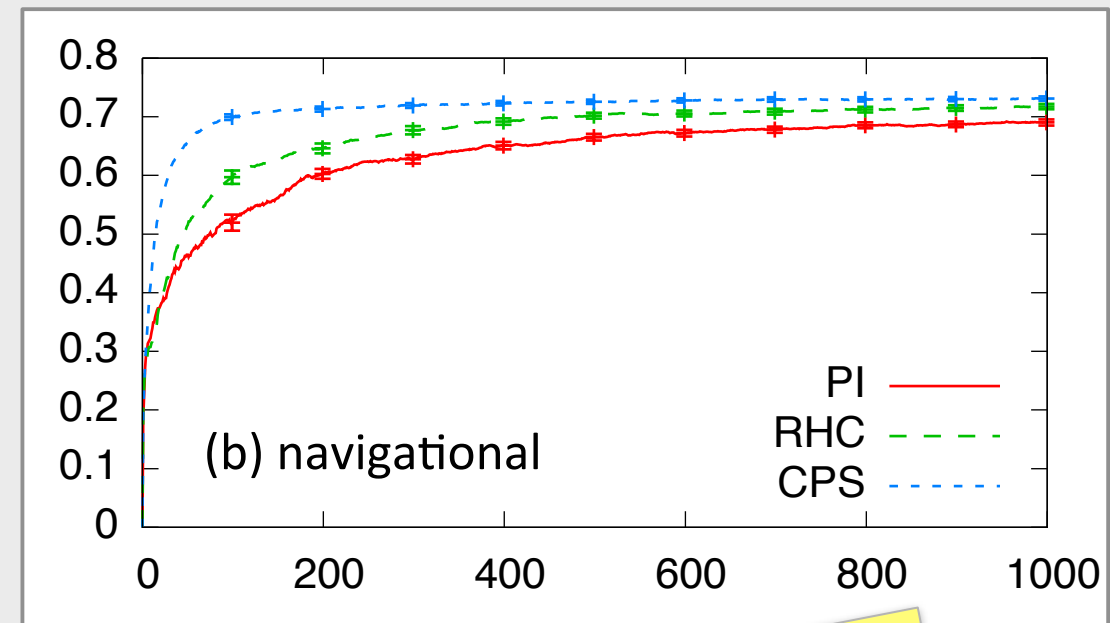
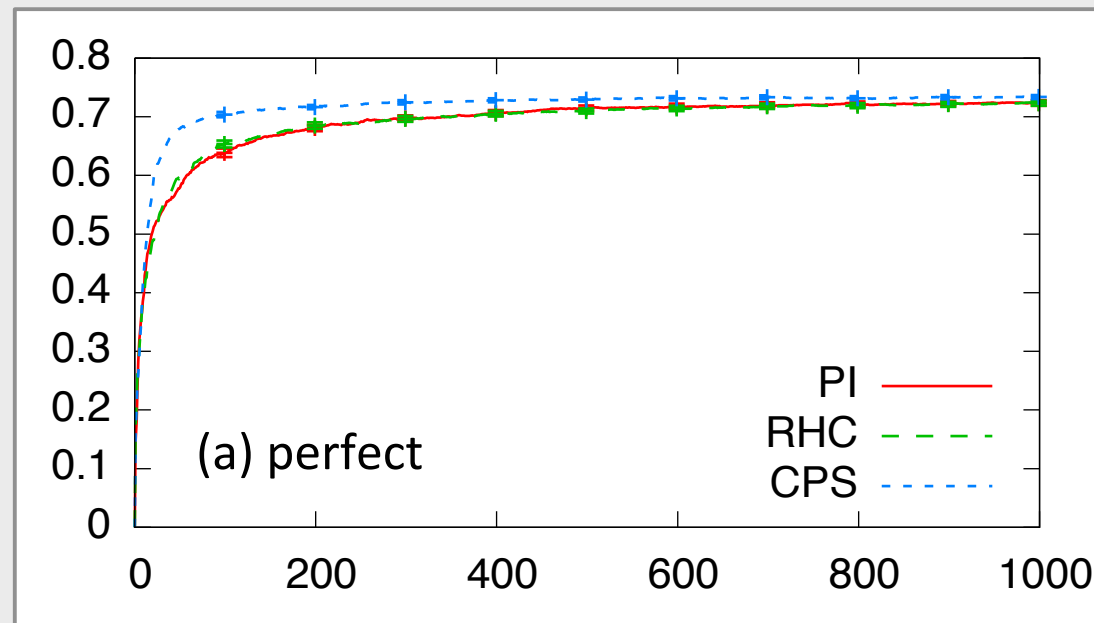
Analysis: Offline Performance



Offline performance of

- **PI**: probabilistic interleave
 - **RHC**: probabilistic interleave with history
 - **CPS**: candidate pre-selection
- on data set NP2003 under varying reliability of click feedback

Analysis: Offline Performance



Offline performance

- **PI:** performance on data set
- **RHC:** performance on data set with history
- **CPS:** performance on data set with increasing noise

Outline for 3. Interleaved comparison

3.1 Inferring Feedback

Motivation

Probabilistic Interleave

Increasing Efficiency: Marginalization

Increasing Efficiency: Data Reuse

3.2 Online Learning with Data Reuse

3.3 Summary & Outlook

Summary

- Online learning to rank for IR methods need to learn effectively from noisy, relative feedback
- We looked at methods for learning quickly and reliably:
 - Probabilistic interleave (for live and historical data)
 - CPS for data reuse in online learning to rank
- **It is possible to learn quickly and reliably from user interactions with an IR system**

Applying Online Learning to Rank for IR

- **So far:** Focus on principles, evaluate methods using simulation
- **Next:** Move experiments towards real-life settings to prepare practical applications in personalized, enterprise, (etc.) search
- **Key challenges:** User reactions to dynamic systems are poorly understood (except: recommender systems and games) – How should interactions be designed? How fast should systems learn? How quickly do user preferences / behavior change?

Smart Exploration

- **So far:** Learning approach explores random direction, but pre-selecting candidates shows promise
- **Idea:** Explore several promising areas of the solution space in parallel (e.g., using population-based approaches), utilize historical data to quickly zoom in on most promising areas
- **Key challenges:** How to compare large sets of rankers as efficiently as possible? Can result lists be constructed to be maximally informative for multiple ranker comparisons?

Long-term Learning and Planning

- **So far:** Assume queries are independent of previous interactions (contextual bandit setting)
- **Next:** Learn more complex interaction patterns, e.g., to optimize system actions throughout sessions
- **Key challenges:** What kind of feedback can be inferred from user interactions? How can it best be used for learning?

This lecture

■ Use naturally occurring side-products ...

- ... of user interactions
- ... of edited or user generated content creation

for training, tuning and testing purposes

■ Pseudo-test collections

■ Click models

■ Interleaved comparisons



Reading material

■ The Cranfield tradition

- M. Sanderson, Test Collection Based Evaluation of Information Retrieval Systems, Foundations and Trends in Information Retrieval, 6, 2010
- E.M. Voorhees and D. Harman, TREC: Experiment and Evaluation in Information Retrieval, MIT Press, 2005

■ A/B testing

- R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery, 18, 2008

■ Interleaving

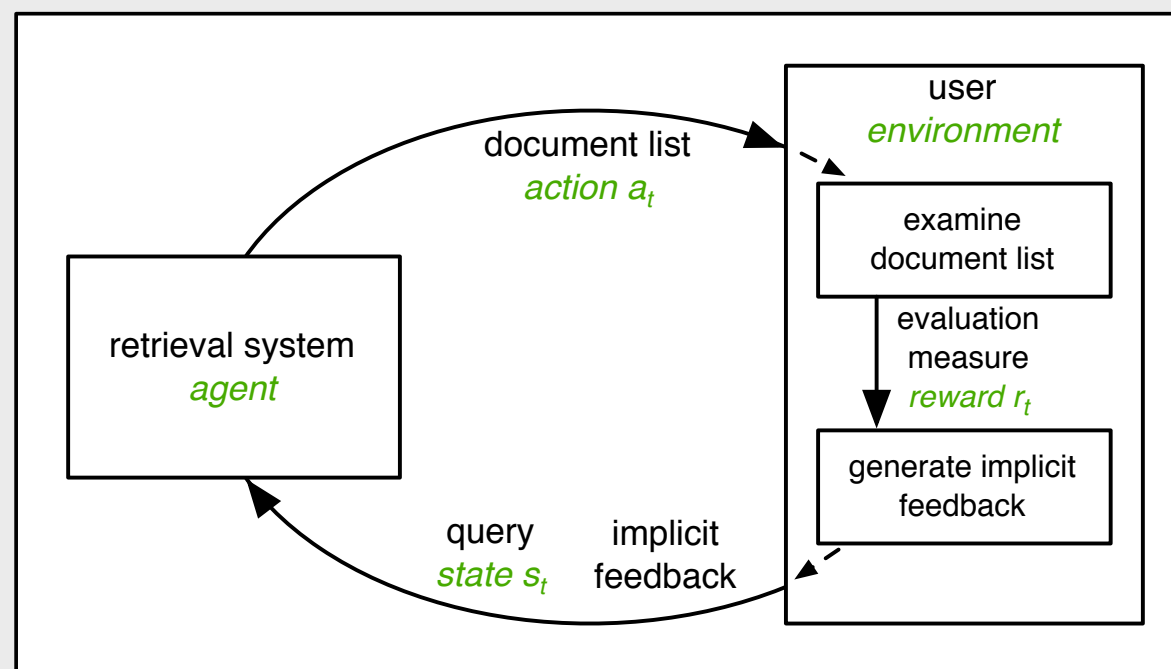
- T. Joachims. Optimizing search engines using clickthrough data. In KDD, 2002
- F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In SIGIR. ACM, 2010.
- K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. CIKM, 2011. (CIKM 2012, WSDM 2013)

■ Pseudo test collections

- I. Soboroff, C.K. Nicholas and P. Cahan. Ranking retrieval systems without relevance judgments. In SIGIR, 2001
- L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In SIGIR, 2007
- R. Berendsen, E. Tsagkias, M. de Rijke, and E. Meij. Generating pseudo test collections for learning to rank scientific articles. In CLEF, 2012

(this slide intentionally left blank)

Problem Formulation for online learning to rank



The IR problem modeled as a contextual bandit problem with RL terminology in *green* and IR terminology in black.

Reinforcement learning (RL) Approach

Learn by trying out actions (document lists), and observing implicit feedback

- **Formulation:** contextual bandit problem
 - context = query – document features: $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_D\}$
 - queries are independent
- **Goal:** present result lists \mathbf{l}_t that maximize discounted cumulative reward:

$$C = \sum_{t=0}^{\infty} \gamma^{t-1} r_t(\mathbf{l}_t)$$

γ - discount factor