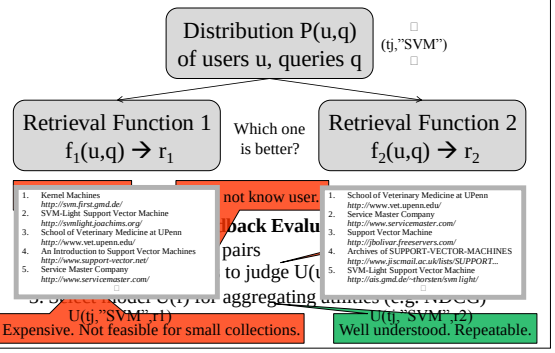
How does Clickthrough Data Reflect Retrieval Quality?

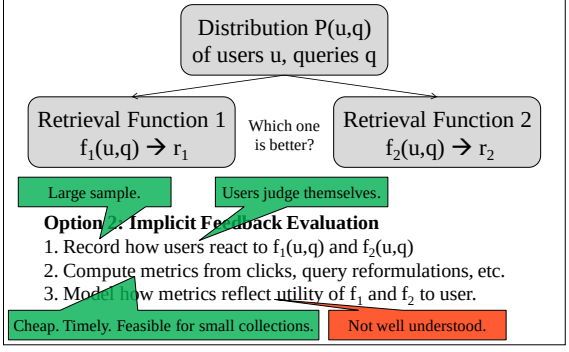
Filip Radlinski (now MSR), Madhu Kurup (now Amazon), Thorsten Joachims

Department of Computer Science
Cornell University

Evaluating Retrieval Functions



Evaluating Retrieval Functions



Research Questions

- **Question 1: Absolute Metrics**
  - Do metrics derived from observed user behavior provide absolute feedback about retrieval quality of f?
  - For example:
    - U(f) ~ numClicks(f)
    - U(f) ~ 1/abandonment(f)
- **Question 2: Paired Comparison Tests**
  - Do paired comparison tests provide relative preferences between two retrieval functions f_1 and f_2?
  - For example:
    - $f_1 \succ f_2 \Leftrightarrow \text{pairedCompTest}(f_1, f_2) > 0$

Does User Behavior Reflect Retrieval Quality?

User Study in ArXiv.org

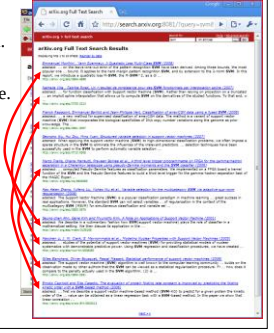
- Natural user and query population.
- User in natural context, not lab.
- Live and operational search engine.
- Ground truth by construction

ORIG ≻ SWAP2 ≻ SWAP4

- ORIG: Hand-tuned fielded
- SWAP2: ORIG with 2 pairs swapped
- SWAP4: ORIG with 4 pairs swapped

ORIG ≻ FLAT ≻ RAND

- ORIG: Hand-tuned fielded
- FLAT: No field weights
- RAND : Top 10 of FLAT shuffled



Absolute Metrics: Experiment Setup

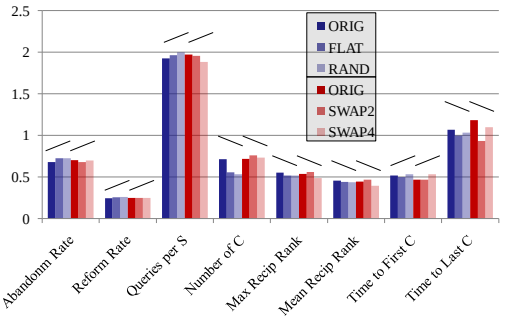
- **Experiment Setup**
  - Phase I: 19.12.2007 – 25.01.2008
    - Users randomly receive ranking from ORIG, FLAT, RAND
  - Phase II: 27.01.2008 – 25.02.2008
    - Users randomly receive ranking from ORIG, SWAP2, SWAP4
  - User are permanently assigned to one experimental condition based on IP address and browser.
- **Basic Statistics**
  - ~700 queries per day / ~300 distinct users per day
- **Quality Control and Data Cleaning**
  - Test run from 03.11.07 – 05.12.2007
  - Heuristics to identify bots and spammers
  - All evaluation code was written twice and cross-validated

Absolute Metrics: Metrics

Name	Description	Aggre-gation	Hypothesized Change with Decreased Quality
Abandonment Rate	% of queries with no click	N/A	Increase
Reformulation Rate	% of queries that are followed by reformulation	N/A	Increase
Queries per Session	Session = no interruption of more than 30 minutes	Mean	Increase
Clicks per Query	Number of clicks	Mean	Decrease
Max Reciprocal Rank*	1/rank for highest click	Mean	Decrease
Mean Reciprocal Rank*	Mean of 1/rank for all clicks	Mean	Decrease
Time to First Click*	Seconds before first click	Median	Increase
Time to Last Click*	Seconds before final click	Median	Decrease

(*) only queries with at least one click count

Absolute Metrics: Results



Absolute Metrics: Results

- **How w… order ORIG ≻ …WAP4?**

Absolute Metrics: Summary and Conclusions

- None of the absolute metrics reflects expected order.
- Most differences not significant after one month of data.
- Absolute metrics not suitable for ArXiv-sized search engines.

			signif ✓	signif ↯
Aban…			2	0
Refor…			0	0
Queri…			0	0
Clicks…			2	0
Max R…			3	0
Mean…			2	0
Time (s) to First Click (Median)	4	1	0	0
Time (s) to Last Click (Median)	4	2	1	1

Research Questions

- **Question 1: Absolute Metrics**
  - Do metrics derived from observed user behavior provide absolute feedback about retrieval quality of f?
  - For example:
    - U(f) ~ numClicks(f)
    - U(f) ~ 1/abandonment(f)
- **Question 2: Paired Comparison Tests**
  - Do paired comparison tests provide relative preferences between two retrieval functions f_1 and f_2?
  - For example:
    - $f_1 \succ f_2 \Leftrightarrow$ pairedCompTest(f_1, f_2) > 0

Paired Comparisons: What to Measure?

(u=tj,q="svm")

$f_1(u,q) \rightarrow r_1$ $\quad$ $f_2(u,q) \rightarrow r_2$

Interpretation: $(r_1 \succ r_2) \leftrightarrow \text{clicks}(r_1) > \text{clicks}(r_2)$

Paired Comparisons: Balanced Interleaving

(u=tj,q="svm")

$f_1(u,q) \rightarrow r_1$ $\quad$ $f_2(u,q) \rightarrow r_2$

Interleaving(r_1,r_2)

Invariant: For all k, top k of balanced interleaving is union of top k_1 of r_1 and top k_2 of r_2 with k_1=k_2 ±1.

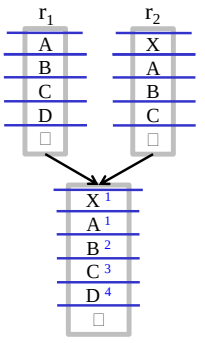
Interpretation: $(r_1 \succ r_2) \leftrightarrow \text{clicks}(\text{topk}(r_1)) > \text{clicks}(\text{topk}(r_2))$

[Joachims/01]

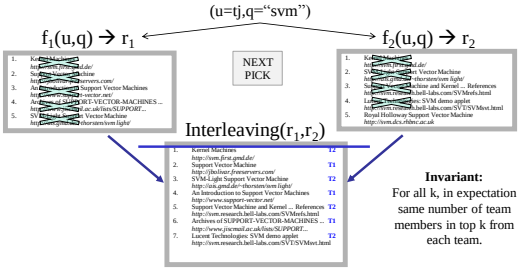
Balanced Interleaving: a Problem

- **Example:**
  - Two rankings r_1 and r_2 that are identical up to one insertion (X)
  - "Random user" clicks uniformly on results in interleaved ranking
    1. "X" → r_2 wins
    2. "A" → r_1 wins
    3. "B" → r_1 wins
    4. "C" → r_1 wins
    5. "D" → r_1 wins

  → biased



Paired Comparisons: Team-Game Interleaving



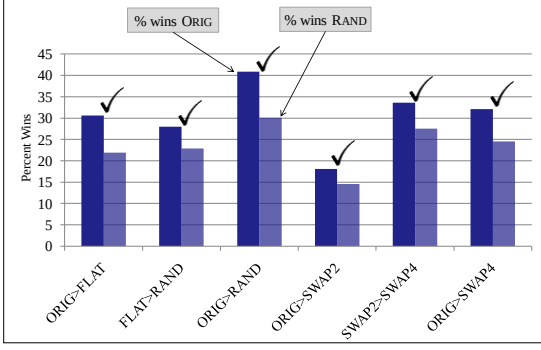
Invariant: For all k, in expectation same number of team members in top k from each team.

Interpretation: $(r_1 \succ r_2) \leftrightarrow \text{clicks}(T_1) > \text{clicks}(T_2)$

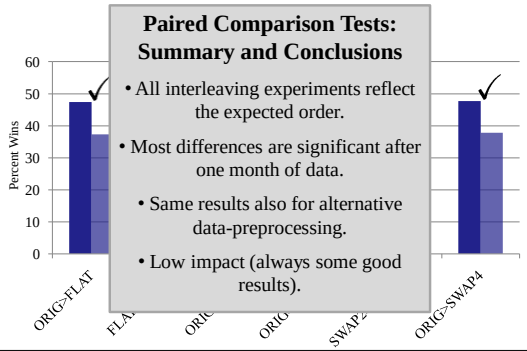
Paired Comparisons: Experiment Setup

- **Experiment Setup**
  - Phase I: 19.12.2007 – 25.01.2008
    - Balanced Interleaving of (ORIG,FLAT) (FLAT,RAND) (ORIG,RAND)
  - Phase II: 27.01.2008 – 25.02.2008
    - Balanced Interleaving of (ORIG,SWAP2) (SWAP2,SWAP4) (ORIG,SWAP4)
  - Phase III: 15.03.2008 – 20.04.2008
    - Team-Game Interleaving of (ORIG,FLAT) (FLAT,RAND) (ORIG,RAND)
    - Team-Game Interleaving of (ORIG,SWAP2) (SWAP2,SWAP4) (ORIG,SWAP4)
- **Quality Control and Data Cleaning**
  - Same as for absolute metrics

Balanced Interleaving: Results



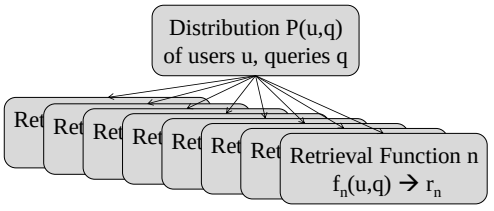
Team-Game Interleaving: Results



Paired Comparison Tests: Summary and Conclusions

- All interleaving experiments reflect the expected order.
- Most differences are significant after one month of data.
- Same results also for alternative data-preprocessing.
- Low impact (always some good results).

Future Work: Evaluation = Learning



Learning Task: Find the f_i that gives best retrieval quality over P(u,q)?

- Algorithm can perform pairwise comparison tests (f_i,f_j)
- Algorithm has to decide on which (f_i,f_j) to compare so that
  - the results from (f_i,f_j) are of good quality
  - the algorithm eventually finds the best retrieval function f*

→ Regret minimization problem "Dueling Bandit Problem"

Summary

- **Interpreting User Interactions as Absolute Feedback**
  - Not reliable for ArXiv-sized retrieval systems.
- **Paired Comparison Tests for Eliciting Relative Feedback**
  - Reliable and significant.
  - Further support for simple decision theoretic user model.
- **Open Question**
  - Verification in other domains (e.g. intranet, desktop, web) Osmot Search Engine (http://radlinski.org/osmot)
  - Beyond clicks for pairwise comparisons
  - The importance of good abstracts (e.g. image search)
  - Robustness to malicious spam
  - Active learning with pairwise comparisons