Modern Retrieval Evaluations

Hongning Wang
CS@UVa

What we have known about IR evaluations

- Three key elements for IR evaluation
  - A document collection
  - A test suite of information needs
  - A set of relevance judgments
- Evaluation of unranked retrieval sets
  - Precision/Recall
- Evaluation of ranked retrieval sets
  - P@k, MAP, MRR, NDCG
- Statistic significance
  - Avoid randomness in evaluation

Rethink retrieval evaluation

- Goal of any IR system
  - Satisfying users' information need
- Core quality measure criterion
  - *"how well a system meets the information needs of its users." – wiki*
- Are traditional IR evaluations qualified for this purpose?
  - What is missing?

Do user preferences and evaluation measures line up? [Sanderson et al. SIGIR'10]

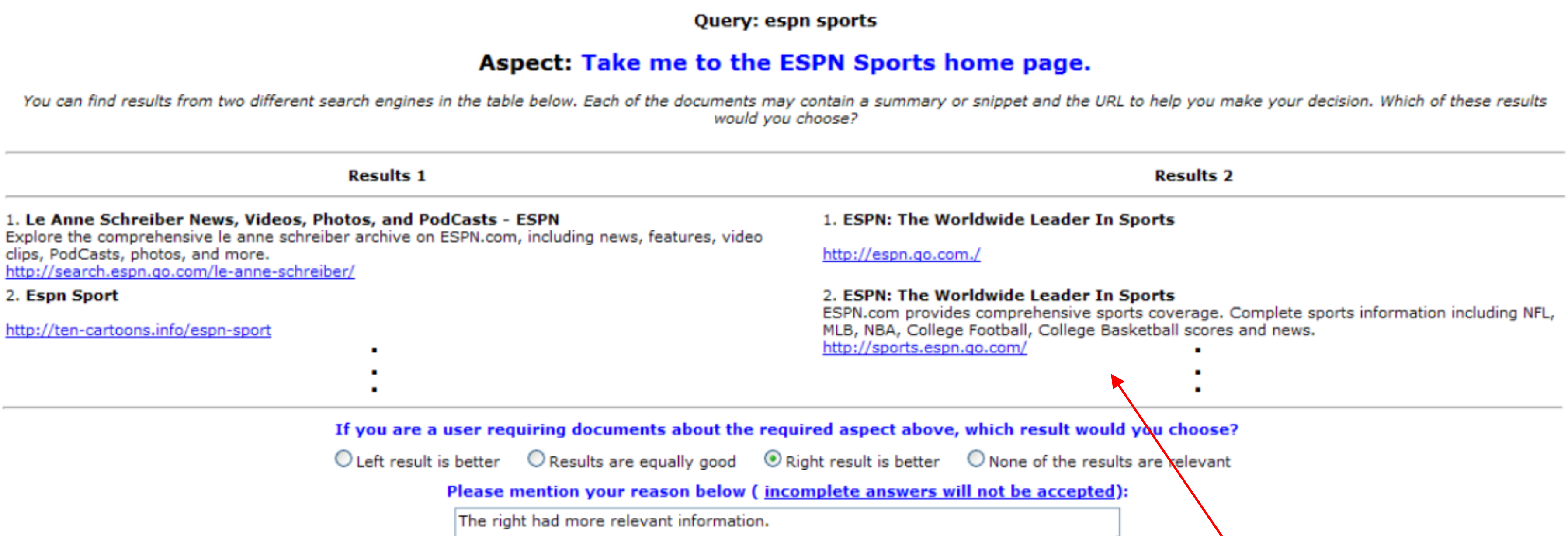
- Research question
  1. Does effectiveness measured on a test collection predict user preferences for one IR system over another?
  2. If such a predictive power exists, does the strength of prediction vary across different search tasks and topic types?
  3. If present, does the predictive power vary when different effectiveness measures are employed?
  4. When choosing one system over another, what are the reasons given by users for their choice?

Experiment settings

- User population
  - Crowd sourcing
    - Mechanical Turk
    - 296 ordinary users
- Test collection
  - TREC'09 Web track
    - 50 million documents from ClueWeb09
  - 30 topics
    - Each included several sub-topics
    - Binary relevance judgment against the sub-topics

Experiment settings

- IR systems
  - 19 runs of submissions to the TREC evaluation



Query: espn sports

Aspect: Take me to the ESPN Sports home page.

You can find results from two different search engines in the table below. Each of the documents may contain a summary or snippet and the URL to help you make your decision. Which of these results would you choose?

Results 1	Results 2
1. **Le Anne Schreiber News, Videos, Photos, and PodCasts - ESPN** Explore the comprehensive le anne schreiber archive on ESPN.com, including news, features, video clips, PodCasts, photos, and more. http://search.espn.go.com/le-anne-schreiber/	1. **ESPN: The Worldwide Leader In Sports** http://espn.go.com./
2. **Espn Sport** http://ten-cartoons.info/espn-sport	2. **ESPN: The Worldwide Leader In Sports** ESPN.com provides comprehensive sports coverage. Complete sports information including NFL, MLB, NBA, College Football, College Basketball scores and news. http://sports.espn.go.com/
⋮	⋮

If you are a user requiring documents about the required aspect above, which result would you choose?

○ Left result is better ○ Results are equally good ◉ Right result is better ○ None of the results are relevant

Please mention your reason below (incomplete answers will not be accepted):

The right had more relevant information.

Users need to make side-by-side comparison to give their preferences over the ranking results

Experimental results

- User preferences v.s. retrieval metrics

Users	nDCG		MRR		P(10)		ERR	
Agree	160	65%	159	67%	131	62%	164	66%
Rnk eql	21	9%	21	9%	18	9%	21	9%
Disgree	66	27%	57	24%	61	29%	62	25%
	247		**237**		**210**		**247**	

  – Metrics generally match users' preferences, no significant differences between metrics

Experimental results

- Zoom into nDCG
  - Separate the comparison into groups of small differences and large differences

Compare to mean difference

Users	nDCG		Small Δ		Large Δ	
Agree	160	65%	96	62%	64	70%
Rank equal	21	9%	16	10%	5	5%
Disagree	66	27%	43	28%	23	25%
	247		**155**		**92**	

  - Users tend to agree more when the difference between the ranking results is <u>large</u>

Experimental results

- What if when one system did not retrieve anything relevant

Users	nDCG		MRR		P(10)		ERR	
Agree	88	72%	88	72%	88	72%	88	72%
Rnk eql	10	8%	10	8%	10	8%	10	8%
Disagree	24	20%	24	20%	24	20%	24	20%
	122		**122**		**122**		**122**	

  – All metrics tell the same and mostly align with the users

Experimental results

- What if when both systems retrieved something relevant at top positions

Users	nDCG		MRR		P(10)		ERR	
Agree	72	56%	71	55%	43	34%	76	59%
Rnk eql	11	9%	11	9%	8	6%	11	9%
Disagree	42	33%	33	26%	37	29%	38	30%
Ties	3	2%	13	10%	40	31%	3	2%
	128		**128**		**128**		**128**	

  – P@10 cannot distinguish the difference between systems

Conclusions of this study

- IR evaluation metrics measured on a test collection predicted user preferences for one IR system over another
- The correlation is strong when the performance difference is large
- Effectiveness of different metrics vary

How does clickthrough data reflect retrieval quality [Radlinski CIKM'08]

- User behavior oriented retrieval evaluation
  - Low cost
  - Large scale
  - Natural usage context and utility
- Common practice in modern search engine systems
  - A/B test

A/B test

- Two-sample hypothesis testing
  - Two versions (A and B) are compared, which are identical except for one variation that might affect a user's behavior
    - E.g., BM25 with different parameter settings
  - Randomized experiment
    - Separate the population into equal size groups
      - 10% random users for system A and 10% random users for system B
    - Null hypothesis: no difference between system A and B
      - Z-test, t-test

Behavior-based metrics

- Abandonment Rate
  - Fraction of queries for which no results were clicked on
- Reformulation Rate
  - Fraction of queries that were followed by another query during the same session
- Queries per Session
  - Mean number of queries issued by a user during a session

Behavior-based metrics

- Clicks per Query
  - Mean number of results that are clicked for each query
- Max Reciprocal Rank
  - Max value of $1/r$, where r is the rank of the highest ranked result clicked on
- Mean Reciprocal Rank
  - Mean value of $\sum_i 1/r_i$, summing over the ranks r_i of all clicks for each query
- Time to First Click
  - Mean time from query being issued until first click on any result
- Time to Last Click
  - Mean time from query being issued until last click on any result

Behavior-based metrics

Metric	Change as ranking gets worse
Abandonment rate	Increase (more bad result sets)
Reformulation rate	Increase (more need to reformulate)
Queries per session	Increase (more need to reformulate)
Clicks per query	Decrease (fewer relevant results)
Max recip. rank	Decrease (top results are worse)
Mean recip. rank	Decrease (more need for many clicks)
Time to first click	Increase (good results are lower)
Time to last click	Decrease (fewer relevant results)

Experiment setup

- Philosophy
  - Given systems with known relative ranking performance
  - Test which metric can recognize such difference

Reverse thinking of hypothesis testing

- In hypothesis testing, we choose system by test statistics
- In this study, we choose test statistics by systems

Constructing comparison systems

- Orig > Flat > Rand
  - Orig: original ranking algorithm from arXiv.org
  - Flat: remove structure features (known to be important) in original ranking algorithm
  - Rand: random shuffling of Flat's results
- Orig > Swap2 > Swap4
  - Swap2: randomly selects two documents from top 5 and swaps them with two random documents from rank 6 through 10 (the same for next page)
  - Swap4: similar to Swap2, but select four documents for swap

Result for A/B test

- 1/6 users of arXiv.org are routed to each of the testing system in one month period

	$\mathcal{H}_1$	ORIG≻FLAT≻RAND		
		ORIG	FLAT	RAND
Abandonment Rate (Mean)	$<$	0.680 ± 0.021	0.725 ± 0.020	0.726 ± 0.020
Reformulation Rate (Mean)	$<$	0.247 ± 0.021	0.257 ± 0.021	0.260 ± 0.021
Queries per Session (Mean)	$<$	1.925 ± 0.098	1.963 ± 0.100	2.000 ± 0.115
Clicks per Query (Mean)	$>$	0.713 ± 0.091	0.556 ± 0.081	0.533 ± 0.077
Max Reciprocal Rank (Mean)	$>$	0.554 ± 0.029	0.520 ± 0.029	0.518 ± 0.030
Mean Reciprocal Rank (Mean)	$>$	0.458 ± 0.027	0.442 ± 0.027	0.439 ± 0.028
Time (s) to First Click (Median)	$<$	31.0 ± 3.3	30.0 ± 3.3	32.0 ± 4.0
Time (s) to Last Click (Median)	$>$	64.0 ± 19.0	60.0 ± 14.0	62.0 ± 9.0

Result for A/B test

- 1/6 users of arXiv.org are routed to each of the testing system in one month period

	$\mathcal{H}_1$	ORIG≻SWAP2≻SWAP4		
		ORIG	SWAP2	SWAP4
Abandonment Rate (Mean)	$<$	0.704 ± 0.021	0.680 ± 0.021	0.698 ± 0.021
Reformulation Rate (Mean)	$<$	0.248 ± 0.021	0.250 ± 0.021	0.248 ± 0.021
Queries per Session (Mean)	$<$	1.971 ± 0.110	1.957 ± 0.099	1.884 ± 0.091
Clicks per Query (Mean)	$>$	0.720 ± 0.098	0.760 ± 0.127	0.734 ± 0.125
Max Reciprocal Rank (Mean)	$>$	0.538 ± 0.029	0.559 ± 0.028	0.488 ± 0.029
Mean Reciprocal Rank (Mean)	$>$	0.444 ± 0.027	0.467 ± 0.027	0.394 ± 0.026
Time (s) to First Click (Median)	$<$	28.0 ± 2.2	28.0 ± 3.0	32.0 ± 3.5
Time (s) to Last Click (Median)	$>$	71.0 ± 19.0	56.0 ± 10.0	66.0 ± 15.0

Result for A/B test

- Few of such comparisons are significant

	weak		signif	
	✓	↯	✓	↯
Abandonment Rate (Mean)	4	2	2	0
Reformulation Rate (Mean)	4	2	0	0
Queries per Session (Mean)	3	3	0	0
Clicks per Query (Mean)	4	2	2	0
Max Reciprocal Rank (Mean)	5	1	3	0
Mean Reciprocal Rank (Mean)	5	1	2	0
Time (s) to First Click (Median)	4	1	0	0
Time (s) to Last Click (Median)	4	2	1	1

Interleave test

- Design principle from sensory analysis
  - Instead of giving absolute ratings, ask for relative comparison between alternatives
    - E.g., is A better than B?
  - Randomized experiment
    - <u>Interleave</u> results from both A and B
    - Giving interleaved results to the same population and ask for their preference
    - Hypothesis test over preference votes

Interleave for IR evaluation

- Team-draft interleaving

Input: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
Init: $I \leftarrow (); TeamA \leftarrow \emptyset; TeamB \leftarrow \emptyset;$
while $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ **do**
  if $(|TeamA| < |TeamB|) \vee$
    $((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ **then**
    $k \leftarrow \min_i \{i : A[i] \notin I\}$ ..... *top result in A not yet in I*
    $I \leftarrow I + A[k];$ .... *append it to I*
    $TeamA \leftarrow TeamA \cup \{A[k]\}$ ...... *clicks credited to A*
  else
    $k \leftarrow \min_i \{i : B[i] \notin I\}$ ..... *top result in B not yet in I*
    $I \leftarrow I + B[k]$ .... *append it to I*
    $TeamB \leftarrow TeamB \cup \{B[k]\}$ ...... *clicks credited to B*
  end if
end while
Output: Interleaved ranking I, $TeamA$, $TeamB$

Interleave for IR evaluation

- Team-draft interleaving

Ranking A:	2	3	1	4	5	7	8	6
Ranking B:	1	2	5	3	6	8	7	4

RND = 0

Interleaved ranking	1	2	3	5	4	6

Result for interleaved test

- 1/6 users of arXiv.org are routed to each of the testing system in one month period
  - Test which group receives more clicks

Comparison Pair	Query Based			User Based		
A $\succ$ B	A wins	B wins	# queries	A wins	B wins	# users
ORIG $\succ$ FLAT	**47.7%**	**37.3%**	1272	**49.6%**	**36.0%**	667
FLAT $\succ$ RAND	**46.7%**	**39.7%**	1376	**46.3%**	**36.8%**	646
ORIG $\succ$ RAND	**55.6%**	**29.8%**	1095	**58.7%**	**28.6%**	622
ORIG $\succ$ SWAP2	44.4%	40.3%	1170	**44.7%**	**37.4%**	693
SWAP2 $\succ$ SWAP4	44.2%	40.3%	1202	45.1%	39.8%	703
ORIG $\succ$ SWAP4	**47.7%**	**37.8%**	1332	**47.2%**	**35.0%**	697

Conclusions

- Interleaved test is more accurate and sensitive
  - 4 out of 6 experiments follows our expectation
- Only click count is utilized in this interleaved test
  - More aspects can be evaluated
    - E.g., dwell-time, reciprocal rank, if leads to download, is last click, is first click

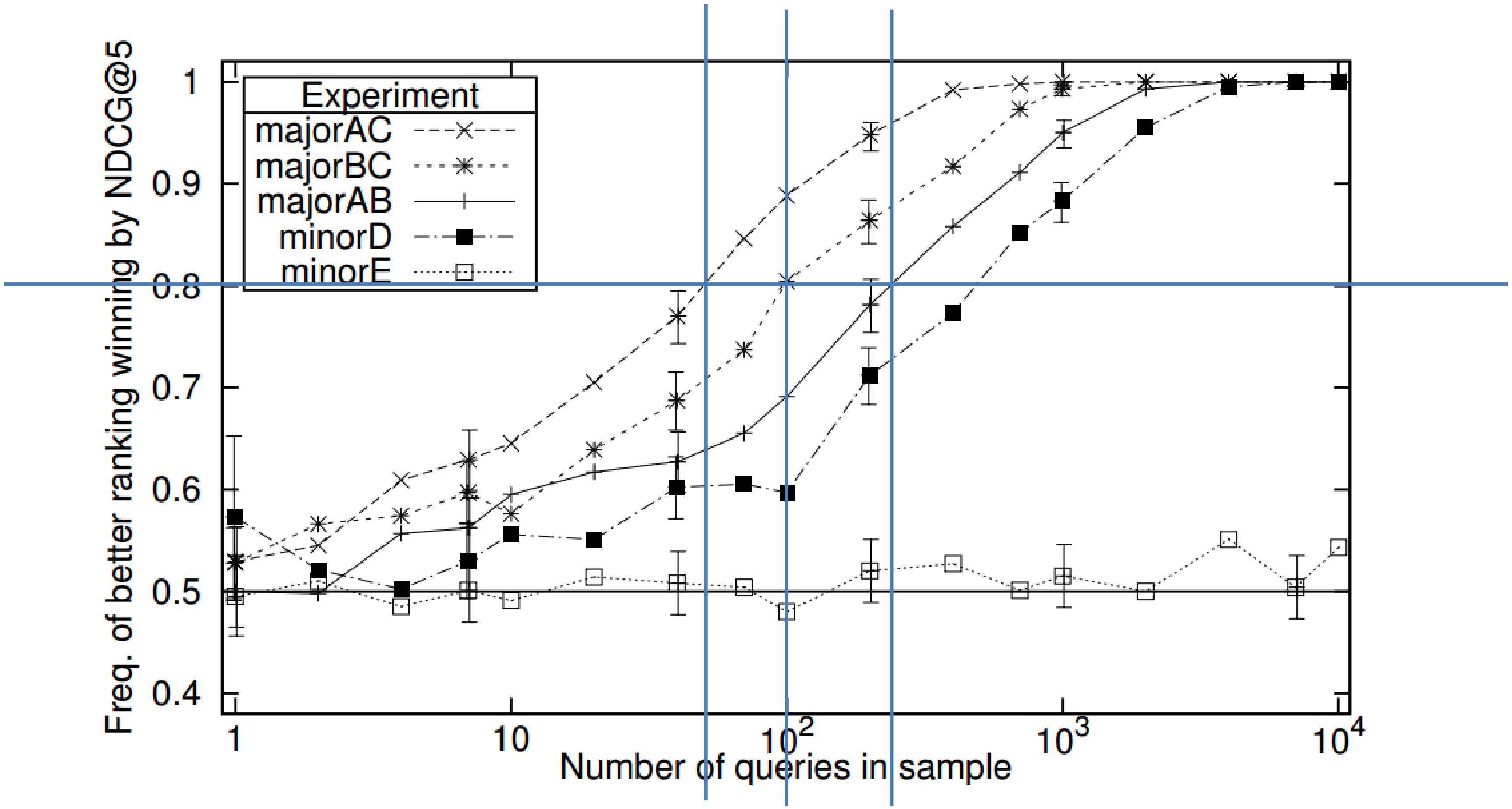
Comparing the sensitivity of information retrieval metrics [Radlinski & Craswell, SIGIR'10]

- How sensitive are those IR evaluation metrics?
  - How many queries do we need to get a confident comparison result?
  - How quickly it can recognize the difference between different IR systems?

Experiment setup

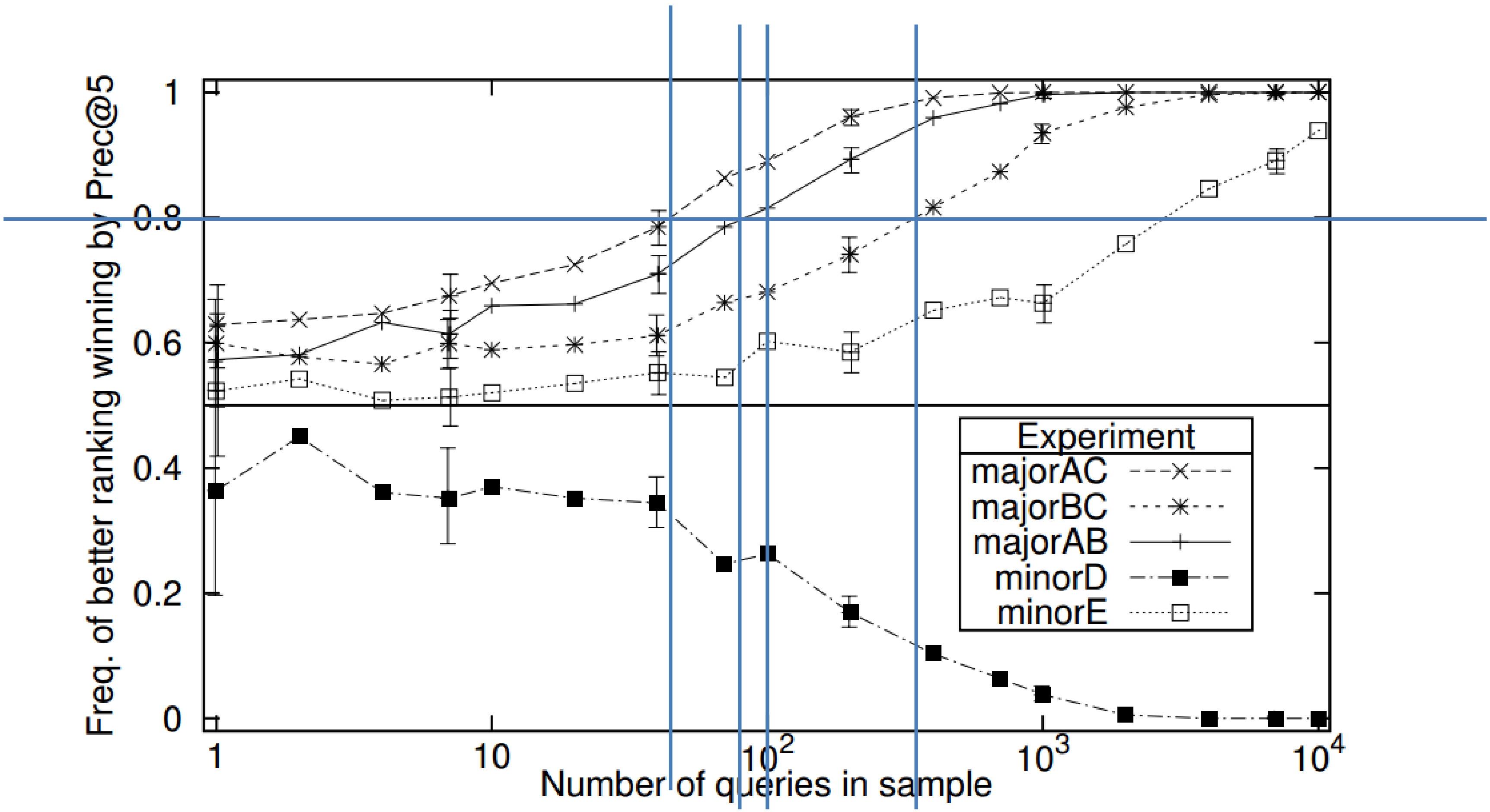
- IR systems with known search effectiveness
- Large set of annotated corpus
  - 12k queries
  - Each retrieved document is labeled into 5-grade level
- Large collection of real users' clicks from a major commercial search engine
- Approach
  - Gradually increase evaluation query size to investigate the conclusion of metrics

Sensitivity of NDCG@5



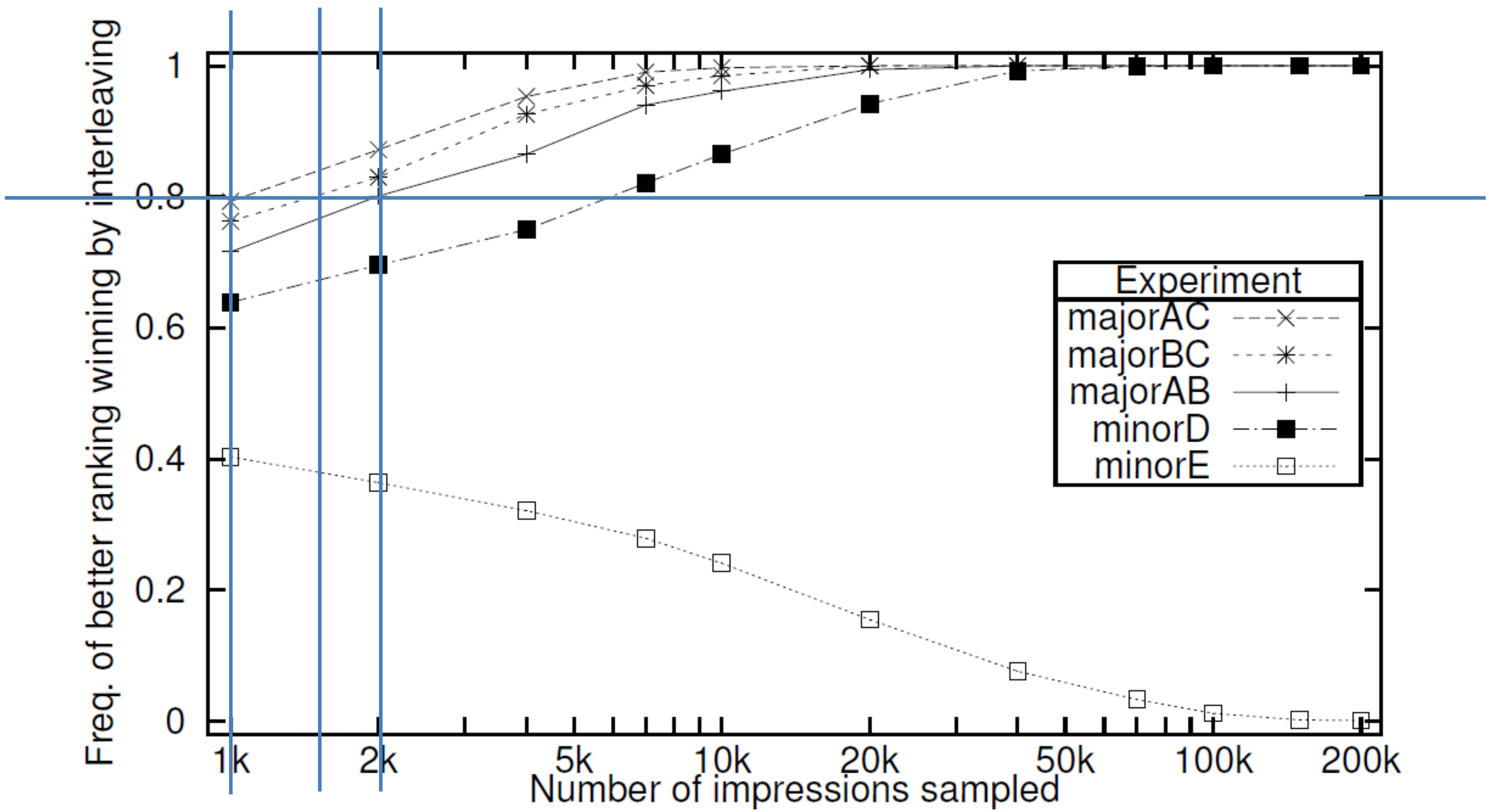
System effectiveness: A>B>C

Sensitivity of P@5



System effectiveness: A>B>C

Sensitivity of interleaving

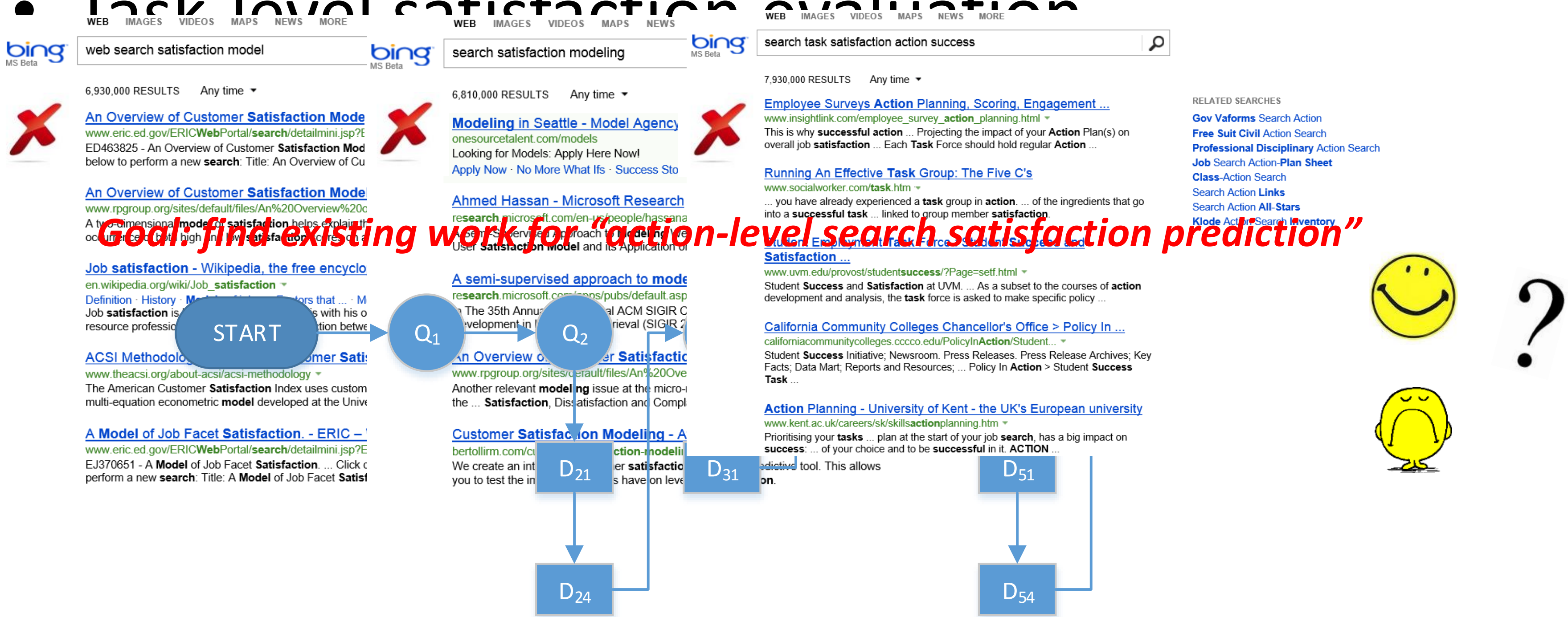


Correlation between IR metrics and interleaving

Inter'l Scoring	IR Metric	Correlation	p-value
Per impression	NDCG@5	0.882	0.048
	MAP@10	0.689	0.198
	P@5	0.662	0.223
Per query	NDCG@5	0.910	0.032
	MAP@10	0.776	0.122
	P@5	0.733	0.159

How to assess search result quality?

- Query-level relevance evaluation
  - Metrics: MAP, NDCG, MRR
- Task-level satisfaction evaluation



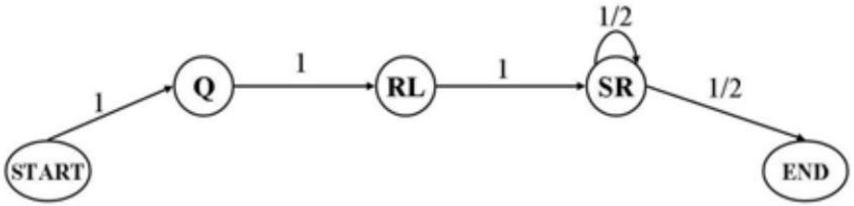
Example of search task

- Information need: *find out what metal can float on water*

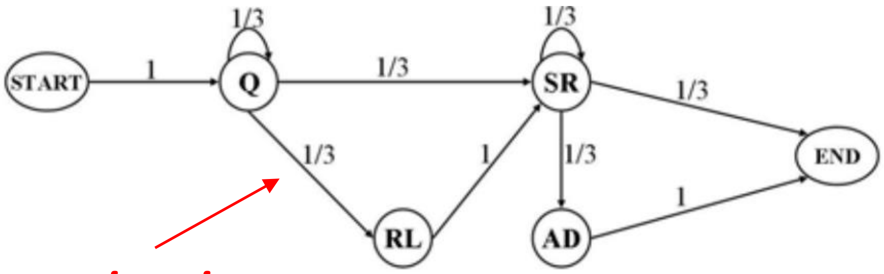
Search Actions	Engine	Time	
Q: metals float on water	Google	10s	
SR: wiki.answers.com		2s	quick back
BR: blog.sciseek.com		3s	
Q: which metals float on water	Google	31s	query reformulation
Q: metals floating on water	Google	16s	
SR: www.blurtit.com		5s	
Q: metals floating on water	Bing	53s	search engine switch
Q: lithium sodium potassium float on water	Google	38s	
SR: www.docbrown.info		15s	

Beyond DCG: User Behavior as a Predictor of a Successful Search [Ahmed et al. WSDM'10]

- Modeling users' sequential search behaviors with Markov models
  - A model for successful search patterns



  - A model for unsuccessful search patterns

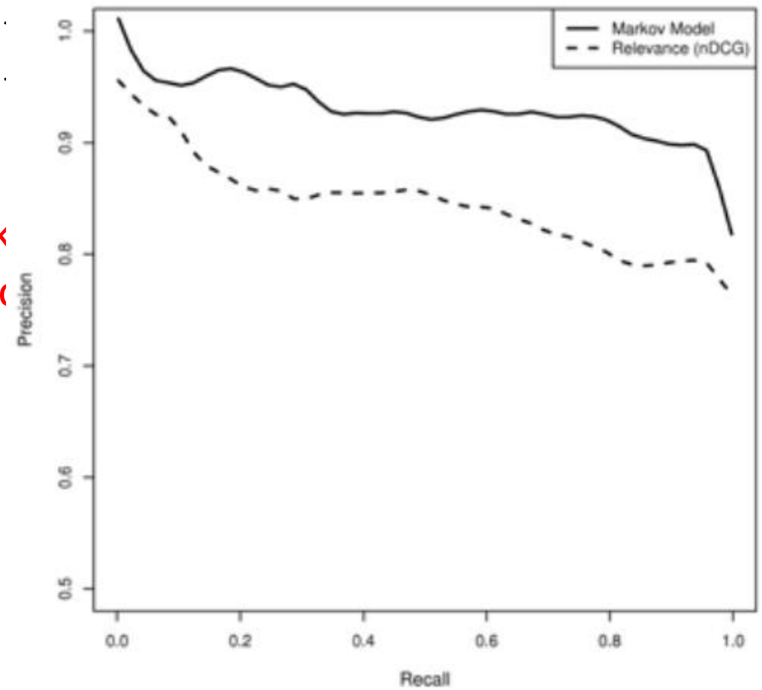


ML for parameter estimation on annotated data set

Predict user satisfaction

- Choose the model that better explains users' search behavior
  - $P(S = 1|B$ … $)$ / … $0|B)p(S=0)$

Lik … or: difficulty of this task,
m … users' expertise of search



Prediction performance for search task satisfaction

What you should know

- IR evaluation metrics generally aligns with users' result preferences
- A/B test v.s. interleaved test
- Sensitivity of evaluation metrics
- Direct evaluation of search satisfaction