Validating Simulations of User Query Variants

ECIR '22, 10-14 April 2022 | Stavanger, Norway

Timo Breuer, Norbert Fuhr, Philipp Schaer

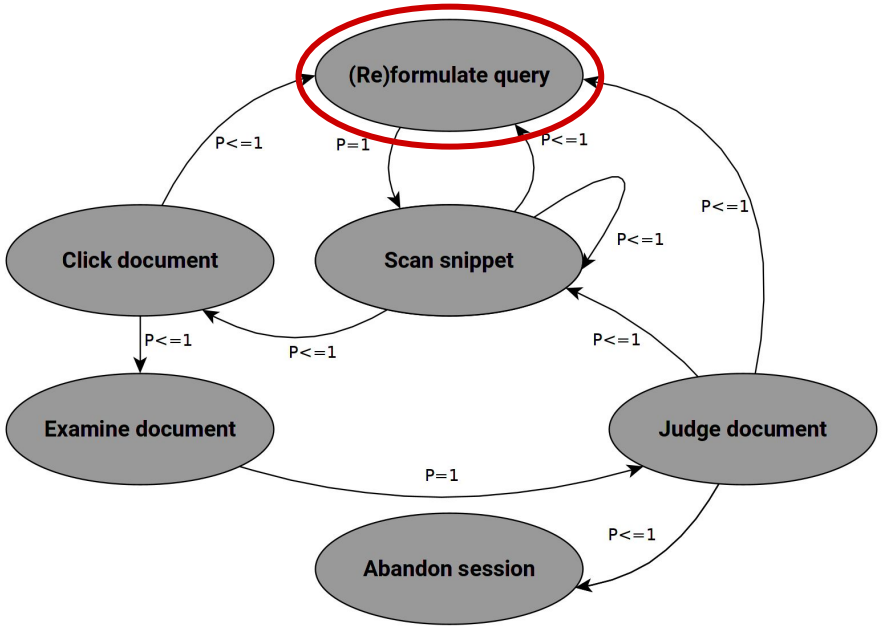
Introduction

Motivations

- Simulated user interactions enhance **system-oriented** evaluations with more **user-oriented** directives in a cost-efficient way
- **Research on query simulations is underrepresented**

Contributions

- **Validation** of (conventional) **query simulation** methods based on TREC test collections
- **Framework** covering different **evaluation** perspectives



Carterette et al., ICTIR 2015;
Maxwell and Azzopardi, CIKM 2016;
Pääkönen et al., Information Retrieval Journal 2017;
Zhang et al., ICTIR 2017

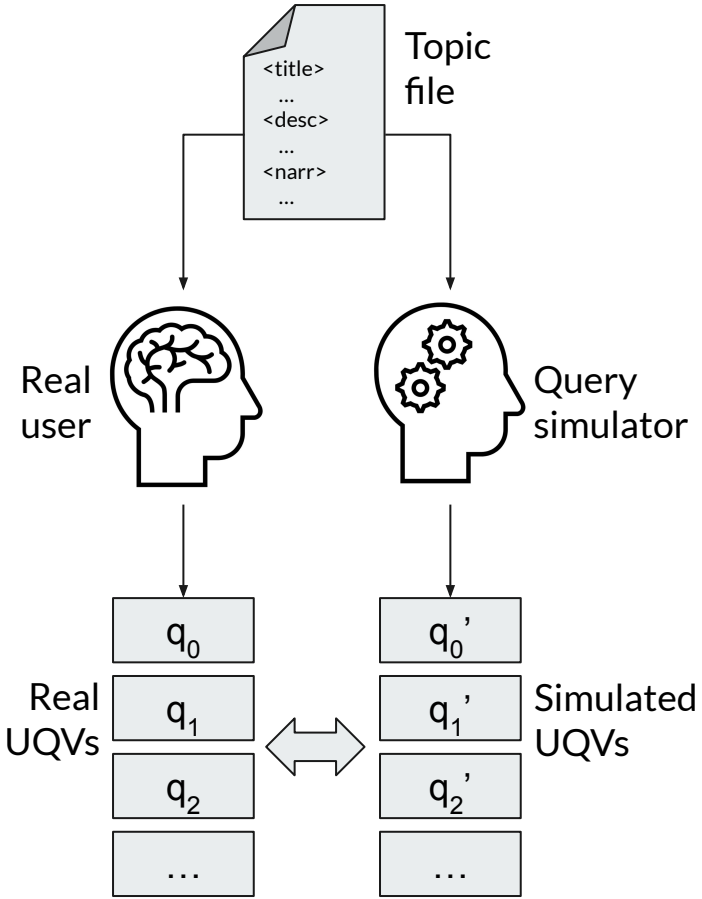
Research questions

RQ1 *How do real user queries relate to simulated queries made from topic texts and known-items in terms of retrieval effectiveness?*

RQ2 *To which degree do simulated queries reproduce real queries provided that only resources of the test collection are considered for the query simulation?*

Simulating user query variants

- Simulations of **user query variants (UQVs)** for a given topic of a TREC test collection
- **Two stage-approach**:
  1. Term candidate generation
  2. Query modification strategy
- Evaluation of **16 query simulators**:
  - 2 term candidate generation methods (TTS and KIS)
  - 8 query modification strategies (S1 - S4'')



Conventional term candidate generation

TREC Topic Searcher (TTS)

$T_{\text{topic}} = \{t_1, ..., t_n\}$ with $t_1, ..., t_n$ ordered by

the term sequence of the concatenated topic's `title`, `description`, and `narrative`

Known-item Searcher (KIS)

$T_{\text{rel}} = \{t_1, ..., t_n\}$ with $t_1, ..., t_n$ ordered by

$$P(t|D_{\text{rel}}) = (1-\lambda)P_{\text{topic}}(t|D_{\text{rel}}) + \lambda P_{\text{background}}(t)$$

Controlled Query Generation [Jordan et al., JCDL 2006]

Conventional query modification strategy

Strategy	Query modifications
S1	$q_1 = \{t_1\}; q_2 = \{t_2\}; q_3 = \{t_3\}; \ldots$
S2	$q_1 = \{t_1, t_2\}; q_2 = \{t_1, t_3\}; q_3 = \{t_1, t_4\}; \ldots$
S2'	$q_1 = \{t_1, t_2, t_3\}; q_2 = \{t_1, t_2, t_4\}; q_3 = \{t_1, t_2, t_5\}; \ldots$
S3	$q_1 = \{t_1\}; q_2 = \{t_1, t_2\}; q_3 = \{t_1, t_2, t_3\}; \ldots$
S3'	$q_1 = \{t_1, t_2, t_3\}; q_2 = \{t_1, t_2, t_3, t_4\}; q_3 = \{t_1, t_2, t_3, t_4, t_5\}; \ldots$

Combining the term generation methods with these strategies results in **10 query simulators**, denoted as TTS_{S1}, KIS_{S1}, TTS_{S2}, KIS_{S2}, ...

Controlled Query Generation combined with Query Change Model

Modified two-stage approach:

1. **Query generation** with n-grams and Controlled Query Generation [Jordan et al., JCDL 2006]
2. **Query ranking** by Query Change Model [Yang et al., TOIS 2015]

Controlled Query Generation combined with Query Change Model

Modified two-stage approach:

1. **Query generation** with n-grams and Controlled Query Generation [Jordan et al., JCDL 2006]
2. **Query ranking** by Query Change Model [Yang et al., TOIS 2015]

n-grams made from $\begin{cases} T_{\text{topic+rel}} \Rightarrow \text{TTS}_{\text{S4-S4''}} \\ T_{\text{rel}} \Rightarrow \text{KIS}_{\text{S4-S4''}} \end{cases}$

$$T_{\text{topic+rel}} = \underbrace{(T_{\text{topic}} \cap T_{\text{rel}})}_{\text{topic terms in } T_{\text{rel}}} \cup \underbrace{(T_{\text{rel}} \setminus T_{\text{topic}})_k}_{\text{- top k terms of relevant documents not in topic terms; - k models user's vocabulary and domain knowledge}}$$

Controlled Query Generation combined with Query Change Model

Modified two-stage approach:

1. **Query generation** with n-grams and Controlled Query Generation [Jordan et al., JCDL 2006]
2. **Query ranking** by Query Change Model [Yang et al., TOIS 2015]

n-grams ranked by $\frac{\sum_j^{|q|} \Theta_j}{|q|}$ (in reference to previous query, starting with q_{title})

$$\Theta_j = \begin{cases} \alpha(1 - P(t_j|D_{\text{rel}})), & t_j \in q_{\text{title}} \\ 1 - \beta P(t_j|D_{\text{rel}}) & t_j \in +\Delta q \wedge t_j \in T_{\text{topic}} \\ \epsilon \, \text{idf}(t_j) & t_j \in +\Delta q \wedge t_j \notin T_{\text{topic}} \\ -\delta P(t_j|D_{\text{rel}}) & t_j \in -\Delta q \end{cases}$$

prefer title terms

prefer topic terms

prefer other terms

prefer previous terms

Controlled Query Generation combined with Query Change Model

Query generation with 3,4,5-gram candidates and **query ranking** by three different parameterizations

Strategy	α	β	ϵ	δ	Description
S4	2.2	0.2	0.05	0.6	prefers title and topic terms, keeps previous query terms
S4'	2.2	0.2	0.25	0.1	mostly keeps previous query terms, tends to include other terms
S4"	0.2	0.2	0.025	0.5	sticks to topic terms, more variation between reformulations

Resulting in 6 more query simulators, denoted as TTS_{S4}, KIS_{S4}, $TTS_{S4'}$, $KIS_{S4'}$, ...

Evaluation framework

- Retrieval performance
  Average retrieval performance, Root-Mean-Square-Error, p-values of t-tests
- Shared task utility
  Relative system orderings compared by Kendall's tau
- Effort and effect
  Session Discounted Cumulative Gain (sDCG),
  Trade-offs between queries and browsing depth
- Query term similarity
  Jaccard similarity

UQV dataset & implementations

- User query variants dataset by Benham and Culpepper:
  `https://culpepper.io/publications/robust-uqv.txt.gz`
  - 8 users formulated 3,152 queries for 250 topics
- TREC Common Core 2017 test collection
  - Each of the 8 users formulated at least one query for the 50 topics
  - 5th user (denoted as UQV_5) formulated 500 queries, i.e., 10 queries for 50 topics
- **Anserini's** indexing and **Pyserini's** interactive search feature
- BM25 with Anserini's default parameters (b=0.4, k=0.9)

Code and query dataset:
`https://github.com/irgroup/ecir2022-uqv-sim`

Retrieval performance

Table A.1: Average retrieval performance over q queries

	All queries				First queries				Best queries			
	q	nDCG	P@10	AP	q	nDCG	P@10	AP	q	nDCG	P@10	AP
UQV_1	150	.3787	.4507	.1581	50	.4293	.5040	.2003	50	.4969	.6320	.2429
UQV_2	52	.4221	.5058	.2020	50	.4096	.4880	.1894	50	.4103	.4900	.1896
UQV_3	68	.3922	.4353	.1780	50	.3979	.4560	.1813	50	.4117	.4800	.1878
UQV_4	123	.4126	.4894	.1888	50	.4469	.5220	.2099	50	.5146	.6300	.2644
UQV_5	500	.3922	.4330	.1649	50	.4447	.4920	.2043	50	.5353	.7240	.2807
UQV_6	136	.4030	.4713	.1843	50	.4488	.5080	.2197	50	.4980	.5980	.2515
UQV_7	50	.4980	.5720	.2418	50	.4980	.5720	.2418	50	.4980	.5720	.2418
UQV_8	156	.3814	.4545	.1645	50	.4046	.4500	.1799	50	.4556	.5620	.2193
TTS_{S1}	500	.0479	.0306	.0127	50	.1705	.1280	.0541	50	.3066	.2360	.0971
TTS_{S2}	500	.1964	.1716	.0688	50	.3592	.3900	.1604	50	.4391	.5100	.2097
$TTS_{S2'}$	500	.3387	.3426	.1413	50	.3895	.4020	.1821	50	.4639	.5940	.2283
TTS_{S3}	500	.3323	.3632	.1388	50	.1705	.1280	.0541	50	.4776	.6080	.2383
$TTS_{S3'}$	500	.3499	.3874	.1474	50	.3592	.3900	.1604	50	.4709	.6060	.2311
TTS_{S4}	500	.4493	.5168	.2088	50	.4409	.4920	.2072	50	.5945	.7620	.3282
$TTS_{S4'}$	500	.4788	.5626	.2288	50	.4976	.5940	.2429	50	.6207	.8040	.3554
$TTS_{S4''}$	500	.3780	.4224	.1644	50	.4393	.4860	.2065	50	.5812	.7680	.3222
KIS_{S1}	500	.1334	.1044	.0314	50	.2836	.2040	.0813	50	.4087	.4400	.1492
KIS_{S2}	500	.3969	.3972	.1615	50	.5096	.5400	.2535	50	.5988	.7460	.3429
$KIS_{S2'}$	500	.5114	.5666	.2507	50	.5474	.6220	.2870	50	.6336	.7980	.3762
KIS_{S3}	500	.5598	.6336	.3009	50	.2836	.2040	.0813	50	.6907	.8620	.4299
$KIS_{S3'}$	500	.5941	.6882	.3285	50	.5096	.5400	.2535	50	.6922	.8620	.4337
KIS_{S4}	500	.5216	.5976	.2604	50	.5146	.5960	.2630	50	.6461	.8200	.3902
$KIS_{S4'}$	500	.5008	.5888	.2416	50	.5033	.5980	.2400	50	.6269	.8080	.3703
$KIS_{S4''}$	500	.4859	.5584	.2293	50	.5191	.6020	.2644	50	.6401	.8360	.3781

Retrieval performance

Table A.1: Average retrieval performance over q queries

	All queries				First queries				Best queries			
	q	nDCG	P@10	AP	q	nDCG	P@10	AP	q	nDCG	P@10	AP
UQV_1	150	.3787	.4507	.1581	50	.4293	.5040	.2003	50	.4969	.6320	.2429
UQV_2	52	.4221	.5058	.2020	50	.4096	.4880	.1894	50	.4103	.4900	.1896
UQV_3	68	.3922	.4353	.1780	50	.3979	.4560	.1813	50	.4117	.4800	.1878
UQV_4	123	.4126	.4894	.1888	50	.4469	.5220	.2099	50	.5146	.6300	.2644
UQV_5	500	.3922	.4330	.1649	50	.4447	.4920	.2043	50	.5353	.7240	.2807
UQV_6	136	.4030	.4713	.1843	50	.4488	.5080	.2197	50	.4980	.5980	.2515
UQV_7	50	.4980	.5720	.2418	50	.4980	.5720	.2418	50	.4980	.5720	.2418
UQV_8	156	.3814	.4545	.1645	50	.4046	.4500	.1799	50	.4556	.5620	.2193
TTS_{S1}	500	.0479	.0306	.0127	50	.1705	.1280	.0541	50	.3066	.2360	.0971
TTS_{S2}	500	.1964	.1716	.0688	50	.3592	.3900	.1604	50	.4391	.5100	.2097
$TTS_{S2'}$	500	.3387	.3426	.1413	50	.3895	.4020	.1821	50	.4639	.5940	.2283
TTS_{S3}	500	.3323	.3632	.1388	50	.1705	.1280	.0541	50	.4776	.6080	.2383
$TTS_{S3'}$	500	.3499	.3874	.1474	50	.3592	.3900	.1604	50	.4709	.6060	.2311
TTS_{S4}	500	.4493	.5168	.2088	50	.4409	.4920	.2072	50	.5945	.7620	.3282
$TTS_{S4'}$	500	.4788	.5626	.2288	50	.4976	.5940	.2429	50	.6207	.8040	.3554
$TTS_{S4''}$	500	.3780	.4224	.1644	50	.4393	.4860	.2065	50	.5812	.7680	.3222
KIS_{S1}	500	.1334	.1044	.0314	50	.2836	.2040	.0813	50	.4087	.4400	.1492
KIS_{S2}	500	.3969	.3972	.1615	50	.5096	.5400	.2535	50	.5988	.7460	.3429
$KIS_{S2'}$	500	.5114	.5666	.2507	50	.5474	.6220	.2870	50	.6336	.7980	.3762
KIS_{S3}	500	.5598	.6336	.3009	50	.2836	.2040	.0813	50	.6907	.8620	.4299
$KIS_{S3'}$	500	.5941	.6882	.3285	50	.5096	.5400	.2535	50	.6922	.8620	.4337
KIS_{S4}	500	.5216	.5976	.2604	50	.5146	.5960	.2630	50	.6461	.8200	.3902
$KIS_{S4'}$	500	.5008	.5888	.2416	50	.5033	.5980	.2400	50	.6269	.8080	.3703
$KIS_{S4''}$	500	.4859	.5584	.2293	50	.5191	.6020	.2644	50	.6401	.8360	.3781

User query variants

Simulated queries

Retrieval performance

Table A.1: Average retrieval performance over q queries

	All queries				First queries				Best queries			
	q	nDCG	P@10	AP	q	nDCG	P@10	AP	q	nDCG	P@10	AP
UQV_1	150	.3787	.4507	.1581	50	.4293	.5040	.2003	50	.4969	.6320	.2429
UQV_2	52	.4221	.5058	.2020	50	.4096	.4880	.1894	50	.4103	.4900	.1896
UQV_3	68	.3922	.4353	.1780	50	.3979	.4560	.1813	50	.4117	.4800	.1878
UQV_4	123	.4126	.4894	.1888	50	.4469	.5220	.2099	50	.5146	.6300	.2644
UQV_5	500	.3922	.4330	.1649	50	.4447	.4920	.2043	50	.5353	.7240	.2807
UQV_6	136	.4030	.4713	.1843	50	.4488	.5080	.2197	50	.4980	.5980	.2515
UQV_7	50	.4980	.5720	.2418	50	.4980	.5720	.2418	50	.4980	.5720	.2418
UQV_8	156	.3814	.4545	.1645	50	.4046	.4500	.1799	50	.4556	.5620	.2193
TTS_{S1}	500	.0479	.0306	.0127	50	.1705	.1280	.0541	50	.3066	.2360	.0971
TTS_{S2}	500	.1964	.1716	.0688	50	.3592	.3900	.1604	50	.4391	.5100	.2097
$TTS_{S2'}$	500	.3387	.3426	.1413	50	.3895	.4020	.1821	50	.4639	.5940	.2283
TTS_{S3}	500	.3323	.3632	.1388	50	.1705	.1280	.0541	50	.4776	.6080	.2383
$TTS_{S3'}$	500	.3499	.3874	.1474	50	.3592	.3900	.1604	50	.4709	.6060	.2311
TTS_{S4}	500	.4493	.5168	.2088	50	.4409	.4920	.2072	50	.5945	.7620	.3282
$TTS_{S4'}$	500	.4788	.5626	.2288	50	.4976	.5940	.2429	50	.6207	.8040	.3554
$TTS_{S4''}$	500	.3780	.4224	.1644	50	.4393	.4860	.2065	50	.5812	.7680	.3222
KIS_{S1}	500	.1334	.1044	.0314	50	.2836	.2040	.0813	50	.4087	.4400	.1492
KIS_{S2}	500	.3969	.3972	.1615	50	.5096	.5400	.2535	50	.5988	.7460	.3429
$KIS_{S2'}$	500	.5114	.5666	.2507	50	.5474	.6220	.2870	50	.6336	.7980	.3762
KIS_{S3}	500	.5598	.6336	.3009	50	.2836	.2040	.0813	50	.6907	.8620	.4299
$KIS_{S3'}$	500	.5941	.6882	.3285	50	.5096	.5400	.2535	50	.6922	.8620	.4337
KIS_{S4}	500	.5216	.5976	.2604	50	.5146	.5960	.2630	50	.6461	.8200	.3902
$KIS_{S4'}$	500	.5008	.5888	.2416	50	.5033	.5980	.2400	50	.6269	.8080	.3703
$KIS_{S4''}$	500	.4859	.5584	.2293	50	.5191	.6020	.2644	50	.6401	.8360	.3781

UQV performance ranges between conventional query simulation methods $TTS_{S1\text{-}S3'}$ and $KIS_{S1\text{-}S3'}$

Lower-bound performance

Upper-bound performance

Retrieval performance

Table A.1: Average retrieval performance over q queries

	All queries				First queries				Best queries			
	q	nDCG	P@10	AP	q	nDCG	P@10	AP	q	nDCG	P@10	AP
UQV_1	150	.3787	.4507	.1581	50	.4293	.5040	.2003	50	.4969	.6320	.2429
UQV_2	52	.4221	.5058	.2020	50	.4096	.4880	.1894	50	.4103	.4900	.1896
UQV_3	68	.3922	.4353	.1780	50	.3979	.4560	.1813	50	.4117	.4800	.1878
UQV_4	123	.4126	.4894	.1888	50	.4469	.5220	.2099	50	.5146	.6300	.2644
UQV_5	500	.3922	.4330	.1649	50	.4447	.4920	.2043	50	.5353	.7240	.2807
UQV_6	136	.4030	.4713	.1843	50	.4488	.5080	.2197	50	.4980	.5980	.2515
UQV_7	50	.4980	.5720	.2418	50	.4980	.5720	.2418	50	.4980	.5720	.2418
UQV_8	156	.3814	.4545	.1645	50	.4046	.4500	.1799	50	.4556	.5620	.2193
TTS_{S1}	500	.0479	.0306	.0127	50	.1705	.1280	.0541	50	.3066	.2360	.0971
TTS_{S2}	500	.1964	.1716	.0688	50	.3592	.3900	.1604	50	.4391	.5100	.2097
$TTS_{S2'}$	500	.3387	.3426	.1413	50	.3895	.4020	.1821	50	.4639	.5940	.2283
TTS_{S3}	500	.3323	.3632	.1388	50	.1705	.1280	.0541	50	.4776	.6080	.2383
$TTS_{S3'}$	500	.3499	.3874	.1474	50	.3592	.3900	.1604	50	.4709	.6060	.2311
TTS_{S4}	500	.4493	.5168	.2088	50	.4409	.4920	.2072	50	.5945	.7620	.3282
$TTS_{S4'}$	500	.4788	.5626	.2288	50	.4976	.5940	.2429	50	.6207	.8040	.3554
$TTS_{S4''}$	500	.3780	.4224	.1644	50	.4393	.4860	.2065	50	.5812	.7680	.3222
KIS_{S1}	500	.1334	.1044	.0314	50	.2836	.2040	.0813	50	.4087	.4400	.1492
KIS_{S2}	500	.3969	.3972	.1615	50	.5096	.5400	.2535	50	.5988	.7460	.3429
$KIS_{S2'}$	500	.5114	.5666	.2507	50	.5474	.6220	.2870	50	.6336	.7980	.3762
KIS_{S3}	500	.5598	.6336	.3009	50	.2836	.2040	.0813	50	.6907	.8620	.4299
$KIS_{S3'}$	500	.5941	.6882	.3285	50	.5096	.5400	.2535	50	.6922	.8620	.4337
KIS_{S4}	500	.5216	.5976	.2604	50	.5146	.5960	.2630	50	.6461	.8200	.3902
$KIS_{S4'}$	500	.5008	.5888	.2416	50	.5033	.5980	.2400	50	.6269	.8080	.3703
$KIS_{S4''}$	500	.4859	.5584	.2293	50	.5191	.6020	.2644	50	.6401	.8360	.3781

UQV performance is most similar to $TTS_{S4\text{-}S4''}$

Similar performance range

⇨ focus on **$TTS_{S4\text{-}S4''}$**

Upper-bound performance

Retrieval performance: Root-Mean-Square-Error

$$\mathrm{RMSE}(M(r), M(r')) = \sqrt{\frac{1}{n}\sum_{i}^{n}(M_i(r) - M_i(r'))^2}$$

M Evaluation measure (e.g. P@10, nDCG, AP)

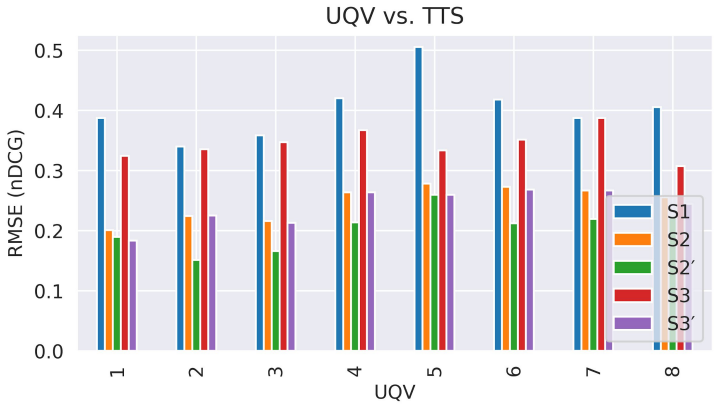
r Run made from real user queries

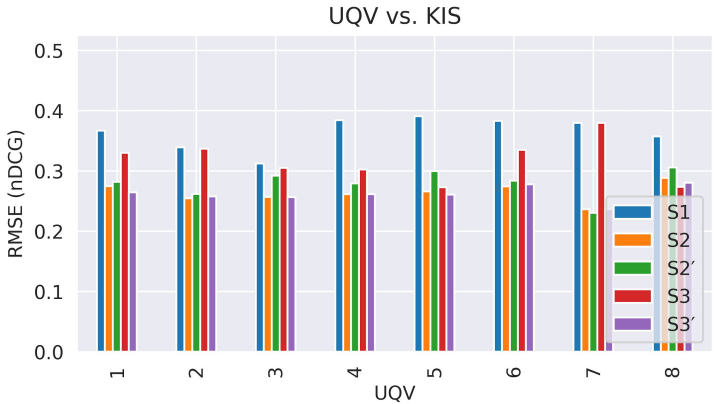
r' Run made from simulated user queries

$M_i(r)$ Score of the i-th topic

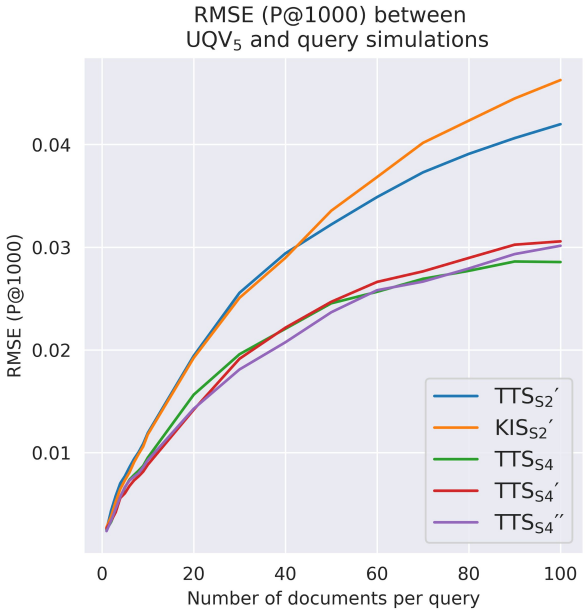
⇨ focus on **TTS**$_{S2'}$ and **KIS**$_{S2'}$

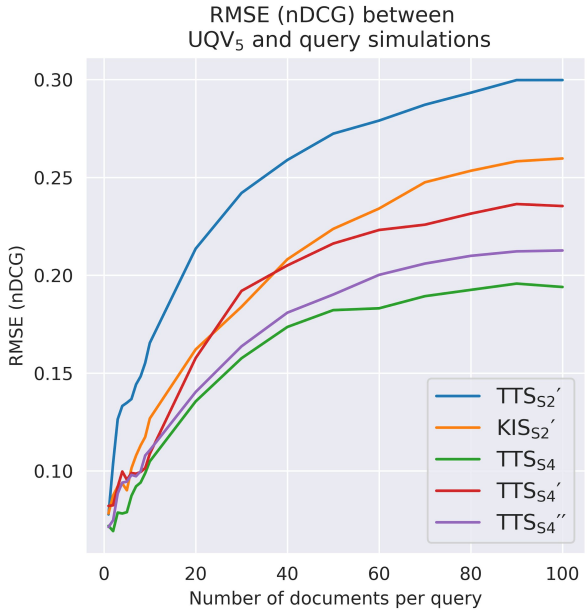
S2': $q_1 = \{t_1, t_2, t_3\}; q_2 = \{t_1, t_2, t_4\}; q_3 = \{t_1, t_2, t_5\}; \ldots$

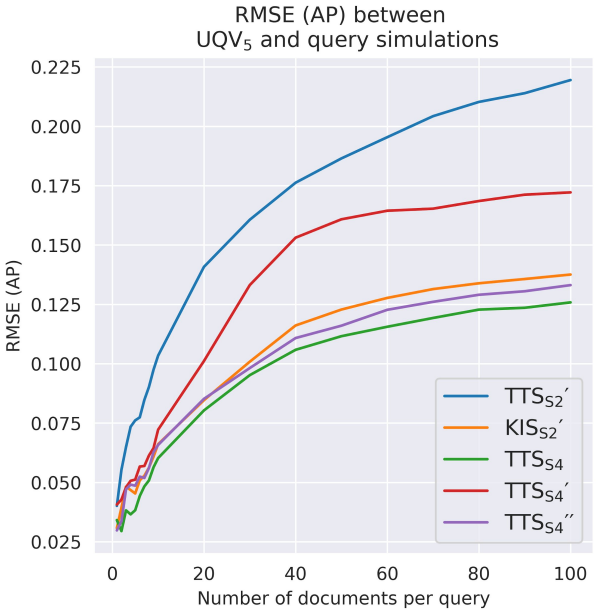




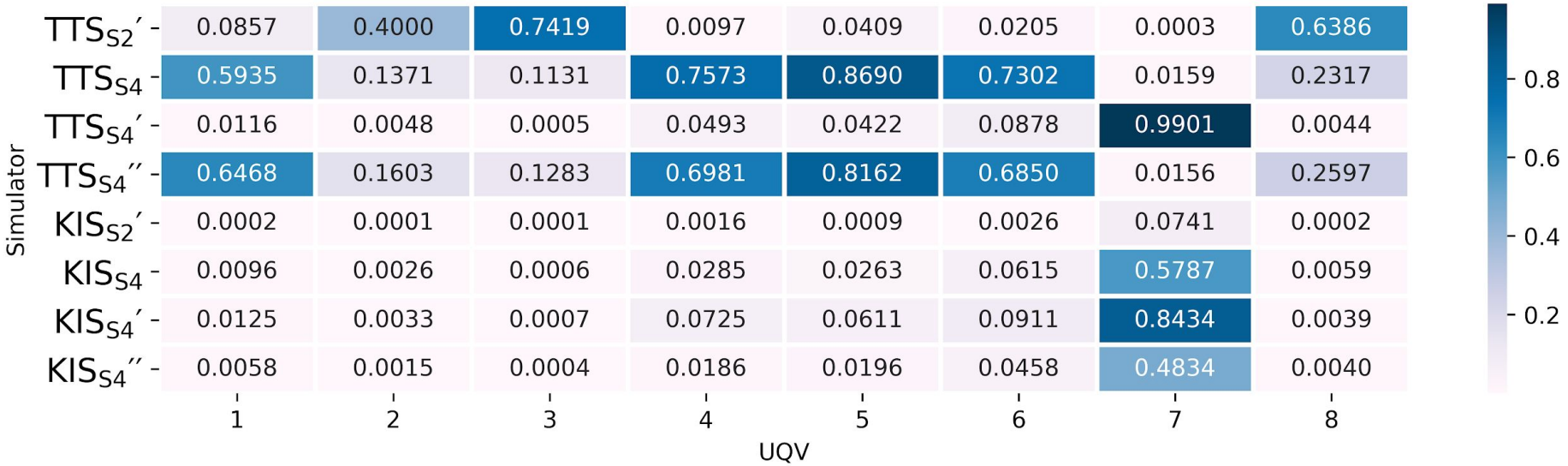
Retrieval performance: Root-Mean-Square-Error



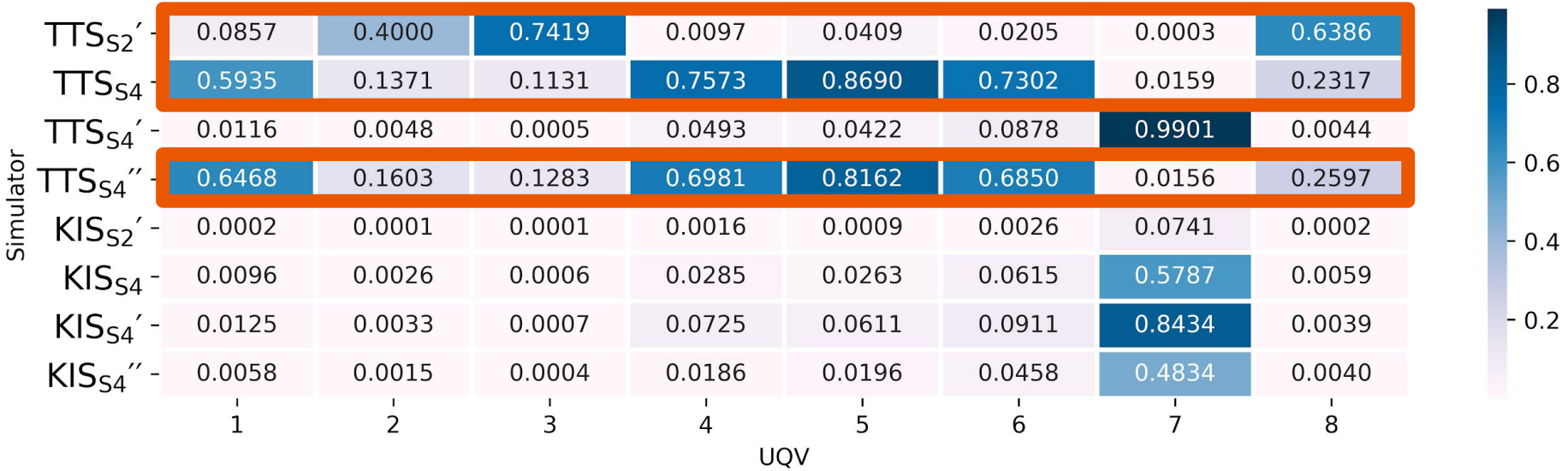




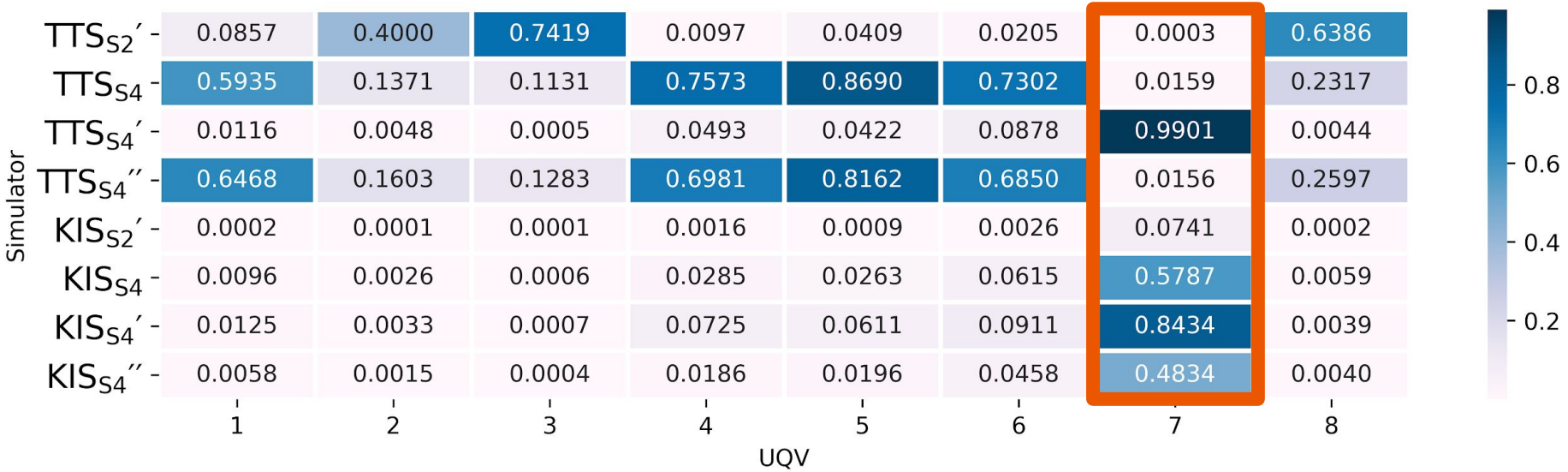
Retrieval performance: p-values of paired t-tests



Retrieval performance: p-values of paired t-tests

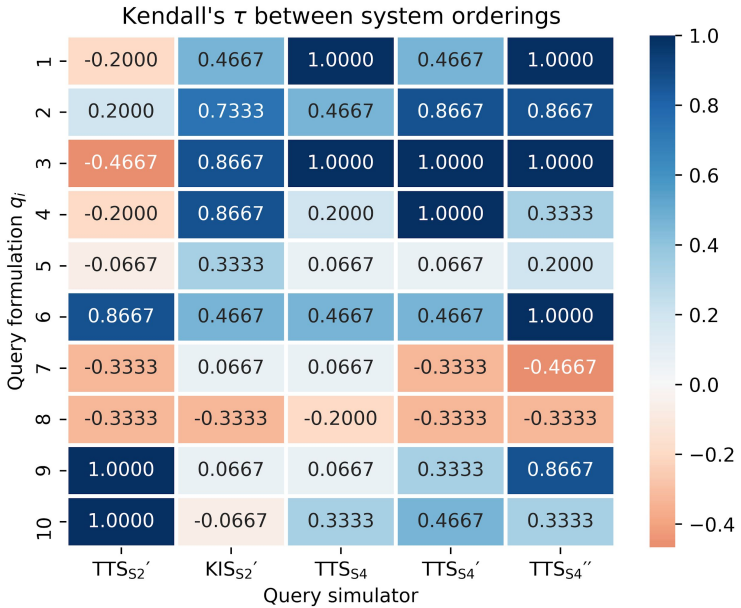


Retrieval performance: p-values of paired t-tests



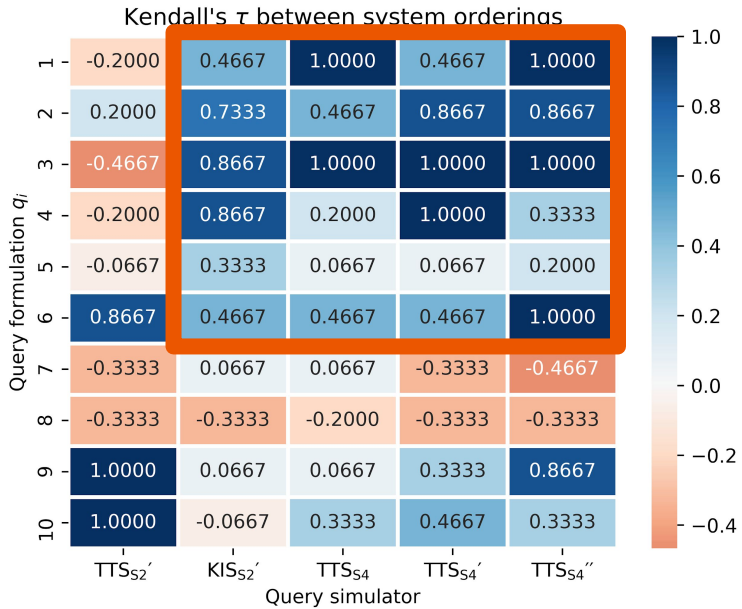
Shared task utility

- Usefulness for shared task evaluations (Huurnik et al., CLEF 2010)
- Comparison of system rankings by Kendall's tau
- Voorhees (SIGIR, 1998) recommends Kendall's tau >0.9 as a rule of thumb
- Five systems based on Query Likelihood with Dirichlet smoothing and different parameterizations



Shared task utility

- Usefulness for shared task evaluations (Huurnik et al., CLEF 2010)
- Comparison of system rankings by Kendall's tau
- Voorhees (SIGIR, 1998) recommends Kendall's tau >0.9 as a rule of thumb
- Five systems based on Query Likelihood with Dirichlet smoothing and different parameterizations



Effort and effect:
Session-based Discounted Cumulative Gain (sDCG)

Järvelin et al., ECIR 2008

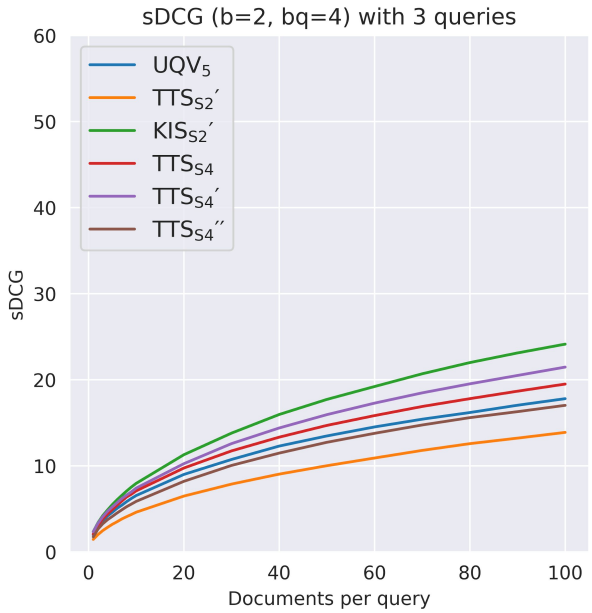
$$\mathrm{sDCG}(q_i) = \frac{\mathrm{DCG}}{1 + \log_{bq}(i)}$$

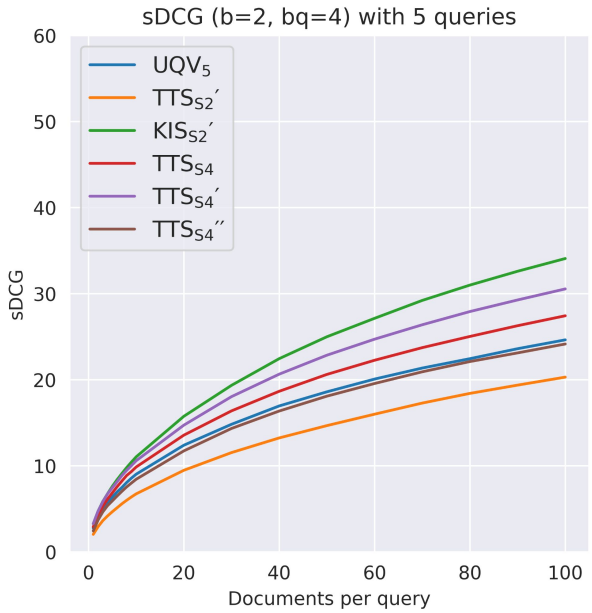
bq logarithm base for the query discount

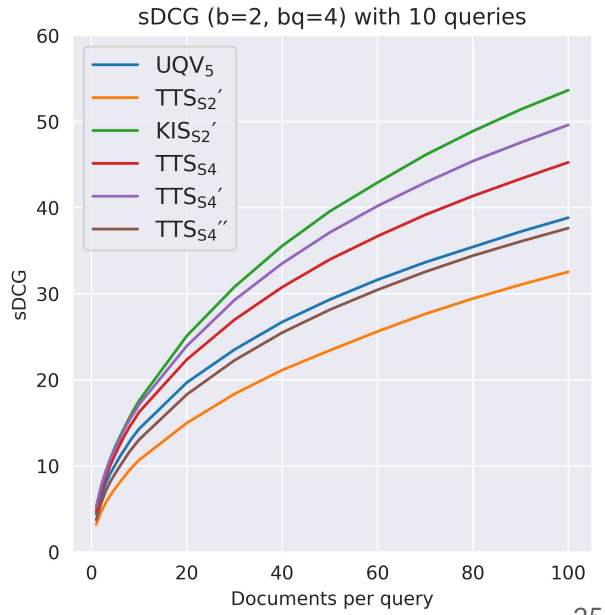
q_i query at the i-th position in a session

DCG discounted cumulative gain

Effort and effect: Session-based Discounted Cumulative Gain (sDCG)

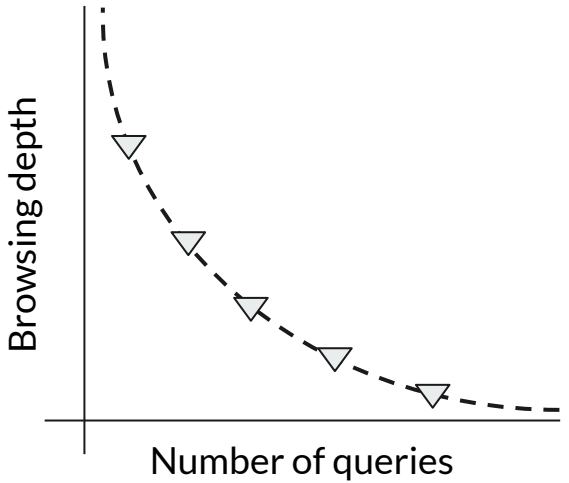




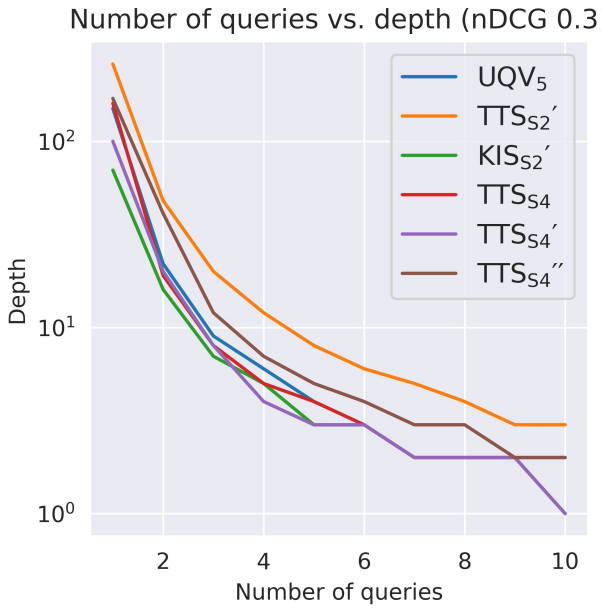


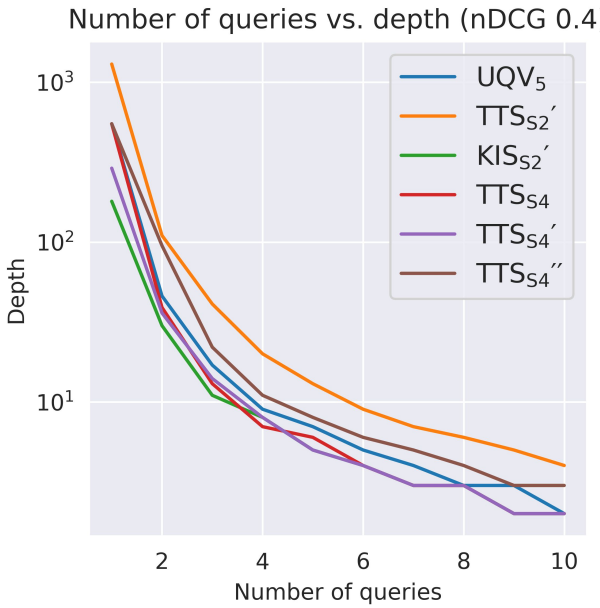
Effort and effect: microeconomics

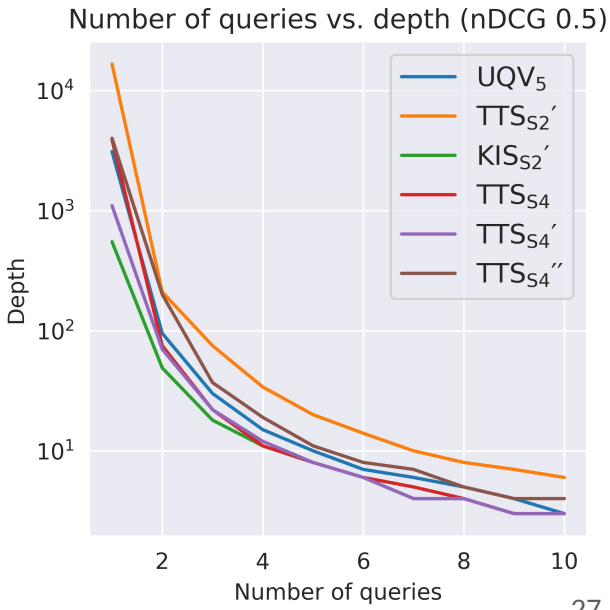
- Azzopardi (SIGIR 2011) applied microeconomics to interactive IR
- Isoquant between **queries** and **browsing depth** for a predefined level of gain (nDCG)
- Distance measure: Mean-Squared-Logarithmic-Error



Effort and effect: microeconomics







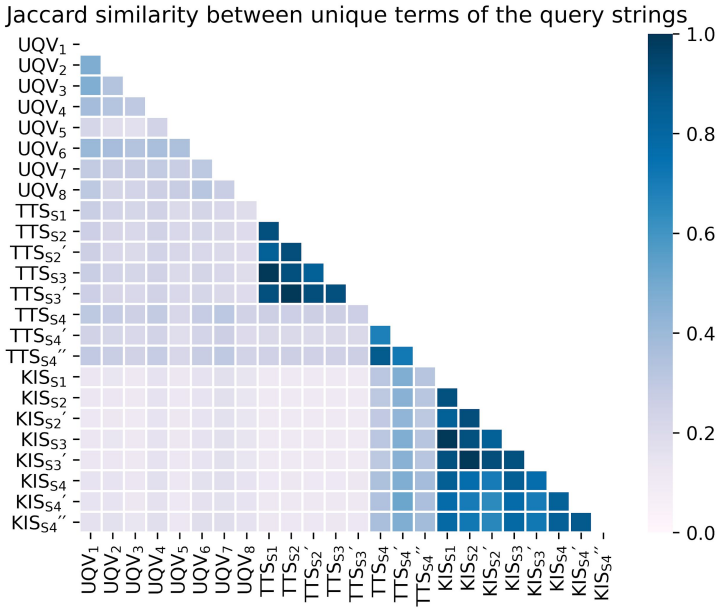
Query term similarity

Jaccard similarity as a measure of variance
[Liu et al., ICTIR 2019; Mackenzie and Moffat, ICTIR 2021]

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q real user query terms

Q' simulated query terms



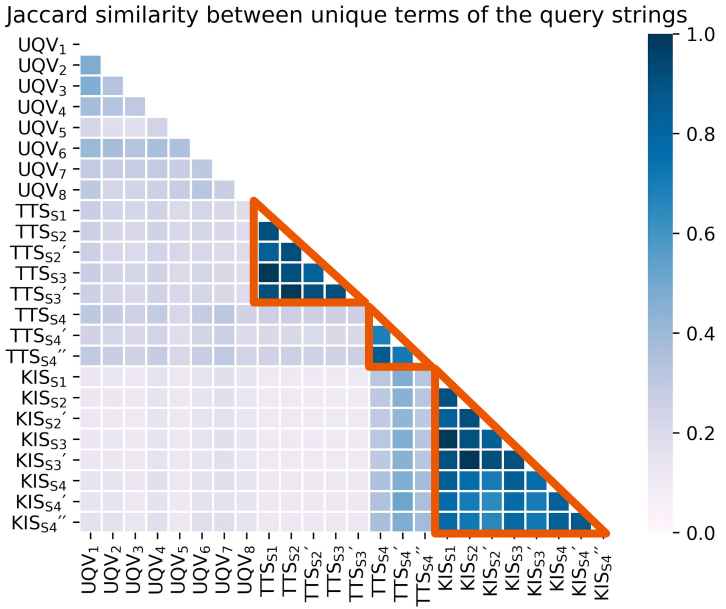
Query term similarity

Jaccard similarity as a measure of variance
[Liu et al., ICTIR 2019; Mackenzie and Moffat, ICTIR 2021]

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q real user query terms

Q' simulated query terms



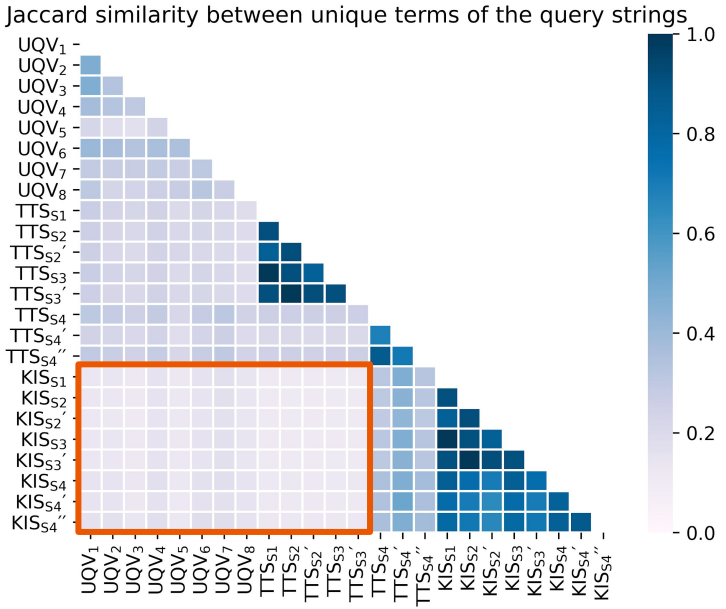
Query term similarity

Jaccard similarity as a measure of variance
[Liu et al., ICTIR 2019; Mackenzie and Moffat, ICTIR 2021]

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q real user query terms

Q' simulated query terms



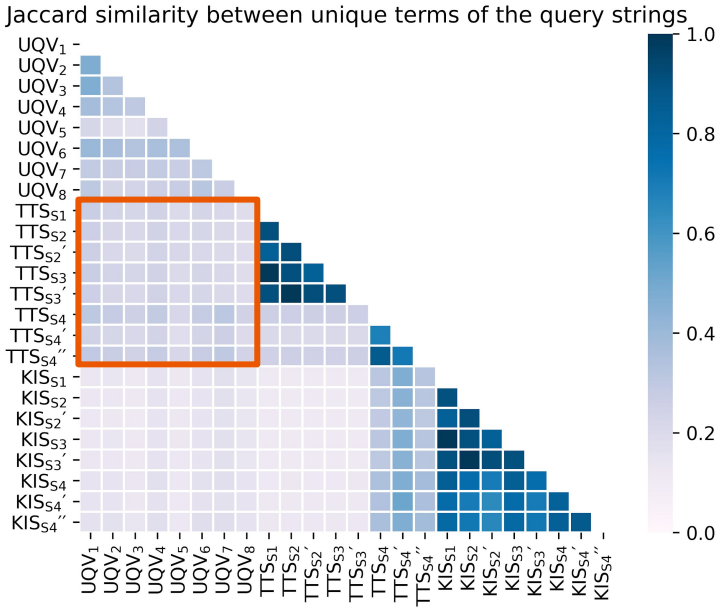
Query term similarity

Jaccard similarity as a measure of variance
[Liu et al., ICTIR 2019; Mackenzie and Moffat, ICTIR 2021]

$$J(Q, Q') = \frac{|Q \cap Q'|}{|Q \cup Q'|}$$

Q real user query terms

Q' simulated query terms



RQ1 *How do real user queries relate to simulated queries made from topic texts and known-items in terms of retrieval effectiveness?*

- The retrieval performance of real user queries ranges between that of conventional query simulation methods
  - Lower bound performance estimates: **TREC Topic Searcher** with strategies S1 - S3'
  - Upper bound performance estimates: **Known-item Searcher** with strategies S1 - S3'

- Better approximations of the retrieval performance can be made by simulating queries with **Controlled Query Generation** and **Query Change Model**
  - Strategies S4-S4'' results in the most similar retrieval performance compared to real UQVs

RQ2 *To which degree do simulated queries reproduce real queries provided that only resources of the test collection are considered for the query simulation?*

- Simulated queries $TTS_{S4\text{-}S4''}$ reproduce real queries:
  - Comparable retrieval performance
  - Lower Root-Mean-Square-Error
  - High p-values for some real user queries
- Shared task utility:
  - More similar relative system orderings for the first queries
  - Later query reformulations have different system orderings
- Economic properties:
  - Session-oriented evaluations show similarities wrt. sDCG and isoquant
- Query term similarity:
  - Only slight overlap between terms of real and simulated queries
  - Highest overlap between simulated queries $TTS_{S1\text{-}S3''}$ and $KIS_{S1\text{-}S3''}$

Thank you!

Many thanks to the SIGIR Student Travel Grant program!

References

Azzopardi; The economics in interactive information retrieval; SIGIR 2011.

Carterette, Bah, Zengin; Dynamic test collections for retrieval evaluation; ICTIR 2015.

Huurnik, Hofmann, de Rijke, Bron; Validating query simulators: an experiment using commercial searches and purchases; CLEF 2010.

Järvelin, Price, Delcambre, Nielsen; Discounted cumulated gain based evaluation of multiple-query IR sessions; ECIR 2008.

Jordan, Watters, Gao; Using controlled query generation to evaluate blind relevance feedback algorithms; JCDL 2006.

Liu, Craswell, Lu, Kurland, Culpepper; A comparative analysis of human and automatic query variants; SIGIR 2019.

Mackenzie and Moffat; Modality effects when simulating user querying tasks; ICTIR 2021.

Maxwell and Azzopardi; Agents, simulated users and humans: an analysis of performance and behaviour; CIKM 2016.

Pääkönen, Kekäläinen, Keskustalo, Azzopardi, Maxwell, Järvelin;Validating simulated interaction for retrieval evaluation; Information Retrieval Journal 2017.

Voorhees; Variations in relevance judgments and the measurement of retrieval effectiveness; SIGIR 1998.

Yang, Guan, Zhang; The query change model: modeling session search as a Markov decision process; TOIS 2015.

Zhang, Liu, Zhai; Information retrieval evaluation as search simulation: a general formal framework for IR evaluation; ICTIR 2017.