Query Performance Prediction

the Past, the Present, and the Future

Debasis Ganguly

Asst. Professor,
University of Glasgow

Outline

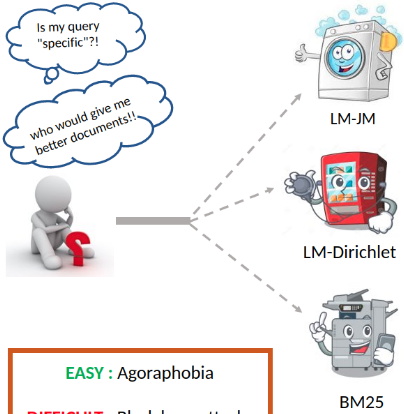
A Brief Introduction to Query Performance Prediction (QPP)

What is Query Performance Prediction (QPP)?

"If we could determine in advance which retrieval approach would work well for a given query, then hopefully, selecting the appropriate retrieval method on a [per] query basis could improve the retrieval effectiveness significantly."

– Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services 2(1), 1–89 (2010)

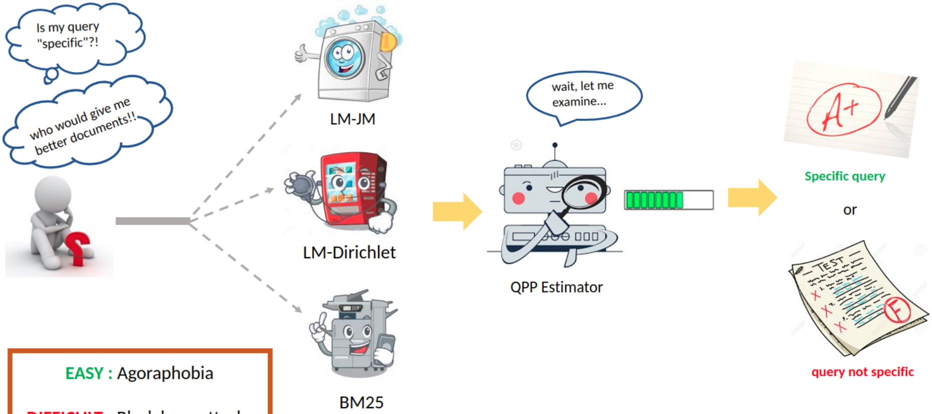
What is Query Performance Prediction (QPP)?



EASY : Agoraphobia

DIFFICULT : Black bear attacks

What is Query Performance Prediction (QPP)?



EASY : Agoraphobia

DIFFICULT : Black bear attacks

Why do we need QPP?

- There are always a number of **difficult queries that cannot be effectively addressed**.

Why do we need QPP?

- There are always a number of **difficult queries that cannot be effectively addressed**.

- Detecting hard/poor-performing queries is useful -
    - Query Reformulation
    - Feedback to system
    - Query Routing

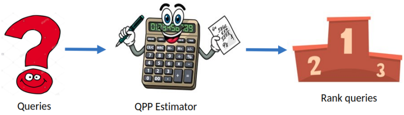
Why do we need QPP?

- There are always a number of **difficult queries that cannot be effectively addressed**.

- Detecting hard/poor-performing queries is useful -
  - Query Reformulation
  - Feedback to system
  - Query Routing

- **QPP** - Predicting the quality of retrieved documents to satisfy the information needs behind the query.

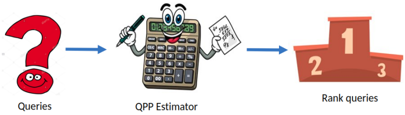
Types of QPP estimators

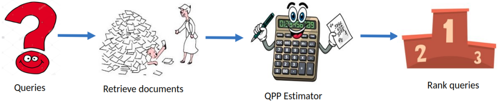


Pre-retrieval

- Predicts the performance of each query based on the content and the context of the query.
- Predictors are often derived from linguistic or statistical information.
- AvgIDF, MaxIDF.

Types of QPP estimators





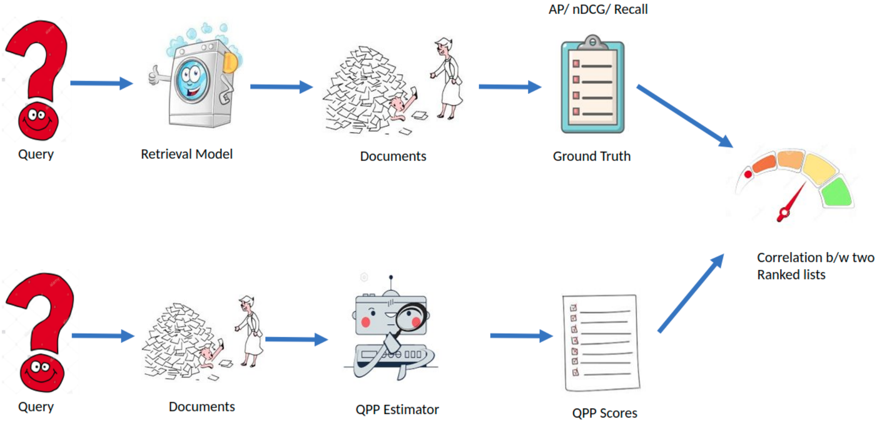
Pre-retrieval

- Predicts the performance of each query based on the content and the context of the query.
- Predictors are often derived from linguistic or statistical information.
- AvgIDF, MaxIDF.

Post-retrieval

- Estimates the query performance by analyzing the result list returned by the retrieval engine.
- Clarity-based approaches - Clarity.
- Score-based approaches - WIG, NQC.
- Robustness-based approaches - UEF.

Evaluating QPP Estimators



A Pairwise Interaction-based Supervised QPP Model (WSDM'22)

Deep-QPP: A data-driven Supervised QPP Model

A purely **data-driven supervised** QPP approach that leverages information from the semantic interactions between the terms of the query and those of the top-retrieved documents.

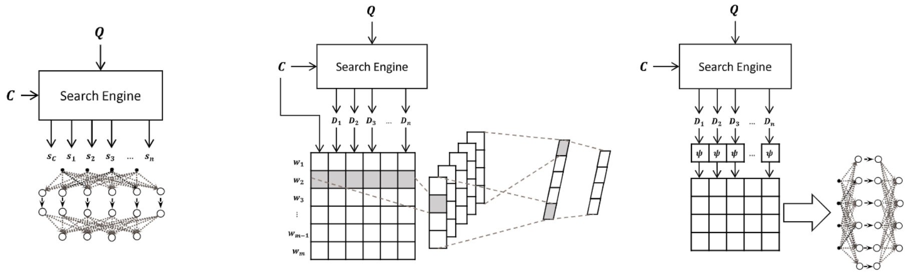
– **Datta, S.**, Ganguly, D., Greene, D., and Mitra, M. Deep-QPP: A pairwise interaction-based deep learning model for supervised query performance prediction. In proceedings of WSDM (2022), ACM, pp. 201–209.

Suchana Datta
University College Dublin

Derek Greene
University College Dublin

Mandar Mitra
Indian Statistical Institute

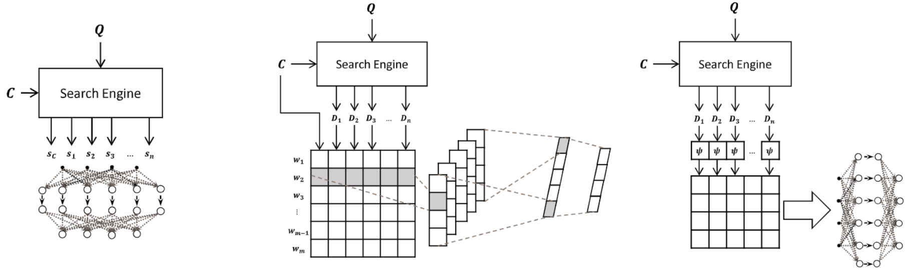
NeuralQPP - SIGIR'18



(a) Retrieval Scores Analyzer (b) Term Distribution Analyzer (c) Semantic Analyzer

Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In proceedings of SIGIR' 18. Association for Computing Machinery, 105–114.

NeuralQPP - SIGIR'18



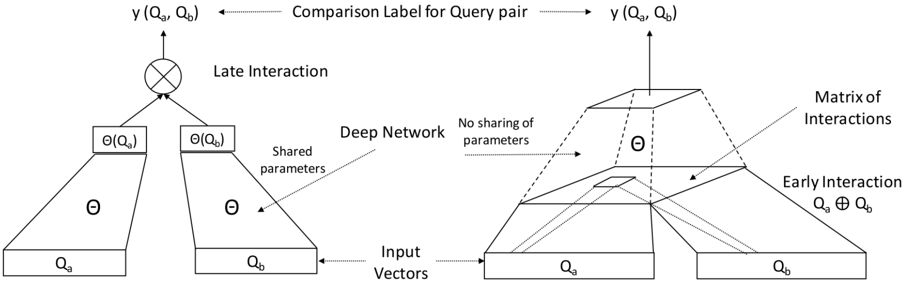
(a) Retrieval Scores Analyzer (b) Term Distribution Analyzer (c) Semantic Analyzer

- A weakly supervised model.
- Uses a combination of features (e.g. retrieval scores) and word embedded vectors to learn an optimal combination.
- A major limitation - training procedure involves weak supervision over a number of estimators.

Advantages of our Proposed Method

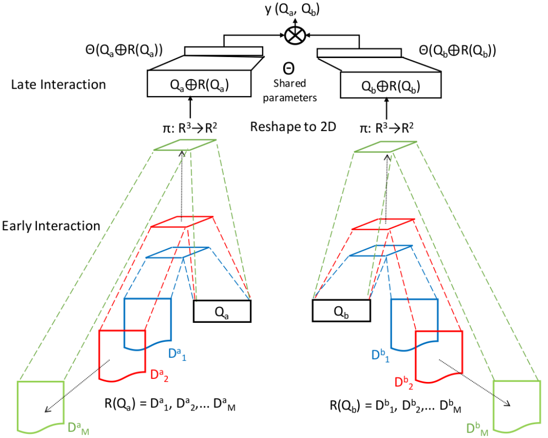
- An end-to-end **strictly supervised** QPP model.
- Solely **data-driven** because it does not rely on other estimators.
- Early interactions between query-document pairs.

Types of Interactions in Pairwise Models



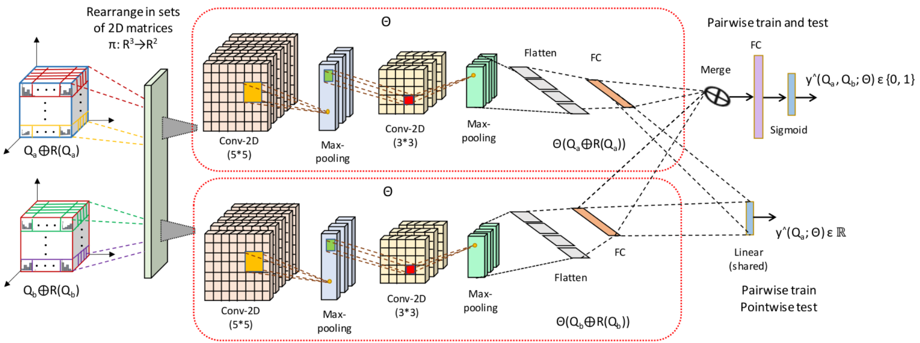
- Representation-based models rely on *late interaction* involving shared parameters (left).
- Interaction-based models make use of *early interactions* transforming paired instances into a single input (right).

Interaction between Queries and Top-Docs

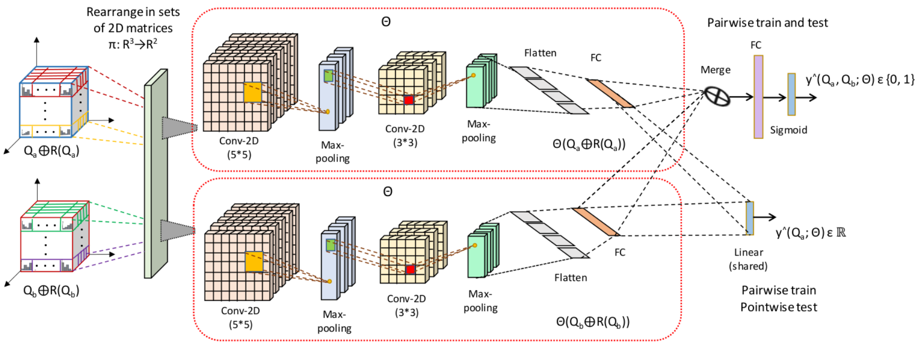


- Combines the benefits of both early and late interactions.
- Includes interaction of the terms in the top-retrieved documents of a query with the constituent terms of the query.
- Incorporates the characteristic pattern of these interactions to estimate the comparison function $y(Q_a, Q_b)$ between a pair of queries.

End-to-end Architecture of Deep-QPP



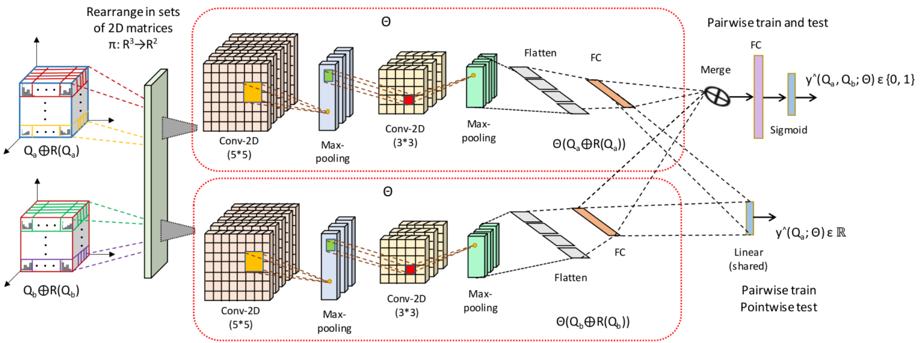
End-to-end Architecture of Deep-QPP



Pairwise: $\mathcal{L}(Q_a, Q_b) = (y(Q_a, Q_b) - \hat{y}(Q_a, Q_b; \Theta))^2$

Pointwise: $\mathcal{L}(Q_a, Q_b) = \max(0, 1 - \mathrm{sgn}(y(Q_a, Q_b) \cdot (\hat{y}(Q_a; \Theta) - \hat{y}(Q_b; \Theta))))$

End-to-end Architecture of Deep-QPP



Pairwise: $\mathcal{L}(Q_a, Q_b) = (y(Q_a, Q_b) - \hat{y}(Q_a, Q_b; \Theta))^2$

Pointwise: $\mathcal{L}(Q_a, Q_b) = \max(0, 1 - \operatorname{sgn}(y(Q_a, Q_b) \cdot (\hat{y}(Q_a; \Theta) - \hat{y}(Q_b; \Theta))))$

Available at: `https://github.com/suchanadatta/DeepQPP.git`

A comparative evaluation of Deep-QPP

Methods	Metric : AP@100						Metric : nDCG@20					
	TREC-Robust			ClueWeb09B			TREC-Robust			ClueWeb09B		
	Pairwise	Pointwise		Pairwise	Pointwise		Pairwise	Pointwise		Pairwise	Pointwise	
	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ
Clarity	0.6251	0.4863	0.3140	0.6120	0.1911	0.0641	0.6118	0.3529	0.2462	0.6101	0.0923	0.0714
NQC	0.6720	0.5269	0.4041	0.7030	0.2654	0.1518	0.6689	0.4261	0.3017	0.6916	0.3105	0.1987
WIG	0.6613	0.5440	0.4279	0.6829	0.2492	0.1920	0.6629	0.3915	0.2407	0.6710	0.2780	0.1823
UEF	0.6941	0.5523	0.4154	0.7217	0.3162	0.1959	0.6792	0.5029	0.3510	0.6925	0.3320	0.1854
SN-BERT	0.6613	0.5208	0.4169	0.6902	0.2317	0.1441	0.6529	0.5023	0.3624	0.6724	0.2241	0.1334
SN-SG	0.6349	0.5112	0.3987	0.6273	0.2110	0.1154	0.6147	0.4736	0.3561	0.6231	0.2049	0.1283
DRMM	0.5871	0.4730	0.3710	0.6023	0.2014	0.1141	0.5629	0.4038	0.3119	0.6004	0.1927	0.1201
WS-NeurQPP	0.8123	0.7215	0.5090	0.7727	0.5192	0.2828	0.7973	0.5913	0.4126	0.7614	0.3928	0.2337
Deep-QPP (MDMQ)	0.7857	0.6988	0.4981	0.7414	0.4636	0.2495	0.7632	0.5649	0.3619	0.7189	0.3509	0.2185
Deep-QPP (SDSQ)	0.7210	0.6303	0.4018	0.6844	0.4208	0.2401	0.7284	0.5112	0.3065	0.6753	0.3124	0.2014
Deep-QPP (MDSQ)	0.8006	0.7203	0.4989	0.7426	0.4840	0.2575	0.7824	0.5601	0.3245	0.7037	0.3518	0.2100
Deep-QPP (SDMQ)	**0.8420**	**0.7404**	**0.5434**	**0.8045**	**0.5532**	**0.3130**	**0.8371**	**0.6315**	**0.4614**	**0.7903**	**0.4431**	**0.2554**

Weakly Supervised Method outperforms Unsupervised Baselines.

Performance of Deep-QPP

Methods	Metric : AP@100						Metric : nDCG@20					
	TREC-Robust			ClueWeb09B			TREC-Robust			ClueWeb09B		
	Pairwise	Pointwise		Pairwise	Pointwise		Pairwise	Pointwise		Pairwise	Pointwise	
	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ
Clarity	0.6251	0.4863	0.3140	0.6120	0.1911	0.0641	0.6118	0.3529	0.2462	0.6101	0.0923	0.0714
NQC	0.6720	0.5269	0.4041	0.7030	0.2654	0.1518	0.6689	0.4261	0.3017	0.6916	0.3105	0.1987
WIG	0.6613	0.5440	0.4279	0.6829	0.2492	0.1920	0.6629	0.3915	0.2407	0.6710	0.2780	0.1823
UEF	0.6941	0.5523	0.4154	0.7217	0.3162	0.1959	0.6792	0.5029	0.3510	0.6925	0.3320	0.1854
SN-BERT	0.6613	0.5208	0.4169	0.6902	0.2317	0.1441	0.6529	0.5023	0.3624	0.6724	0.2241	0.1334
SN-SG	0.6349	0.5112	0.3987	0.6273	0.2110	0.1154	0.6147	0.4736	0.3561	0.6231	0.2049	0.1283
DRMM	0.5871	0.4730	0.3710	0.6023	0.2014	0.1141	0.5629	0.4038	0.3119	0.6004	0.1927	0.1201
WS-NeurQPP	0.8123	0.7215	0.5090	0.7727	0.5192	0.2828	0.7973	0.5913	0.4126	0.7614	0.3928	0.2337
Deep-QPP (MDMQ)	0.7857	0.6988	0.4981	0.7414	0.4636	0.2495	0.7632	0.5649	0.3619	0.7189	0.3509	0.2185
Deep-QPP (SDSQ)	0.7210	0.6303	0.4018	0.6844	0.4208	0.2401	0.7284	0.5112	0.3065	0.6753	0.3124	0.2014
Deep-QPP (MDSQ)	0.8006	0.7203	0.4989	0.7426	0.4840	0.2575	0.7824	0.5601	0.3245	0.7037	0.3518	0.2100
Deep-QPP (SDMQ)	**0.8420**	**0.7404**	**0.5434**	**0.8045**	**0.5532**	**0.3130**	**0.8371**	**0.6315**	**0.4614**	**0.7903**	**0.4431**	**0.2554**

Strictly supervised QPP model **Deep-QPP** outperforms the Weakly Supervised QPP Model.

Performance of Deep-QPP

Methods	Metric : AP@100						Metric : nDCG@20					
	TREC-Robust			ClueWeb09B			TREC-Robust			ClueWeb09B		
	Pairwise	Pointwise		Pairwise	Pointwise		Pairwise	Pointwise		Pairwise	Pointwise	
	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ	Accuracy	P-ρ	K-τ
Clarity	0.6251	0.4863	0.3140	0.6120	0.1911	0.0641	0.6118	0.3529	0.2462	0.6101	0.0923	0.0714
NQC	0.6720	0.5269	0.4041	0.7030	0.2654	0.1518	0.6689	0.4261	0.3017	0.6916	0.3105	0.1987
WIG	0.6613	0.5440	0.4279	0.6829	0.2492	0.1920	0.6629	0.3915	0.2407	0.6710	0.2780	0.1823
UEF	0.6941	0.5523	0.4154	0.7217	0.3162	0.1959	0.6792	0.5029	0.3510	0.6925	0.3320	0.1854
SN-BERT	0.6613	0.5208	0.4169	0.6902	0.2317	0.1441	0.6529	0.5023	0.3624	0.6724	0.2241	0.1334
SN-SG	0.6349	0.5112	0.3987	0.6273	0.2110	0.1154	0.6147	0.4736	0.3561	0.6231	0.2049	0.1283
DRMM	0.5871	0.4730	0.3710	0.6023	0.2014	0.1141	0.5629	0.4038	0.3119	0.6004	0.1927	0.1201
WS-NeurQPP	0.8123	0.7215	0.5090	0.7727	0.5192	0.2828	0.7973	0.5913	0.4126	0.7614	0.3928	0.2337
Deep-QPP (MDMQ)	0.7857	0.6988	0.4981	0.7414	0.4636	0.2495	0.7632	0.5649	0.3619	0.7189	0.3509	0.2185
Deep-QPP (SDSQ)	0.7210	0.6303	0.4018	0.6844	0.4208	0.2401	0.7284	0.5112	0.3065	0.6753	0.3124	0.2014
Deep-QPP (MDSQ)	0.8006	0.7203	0.4989	0.7426	0.4840	0.2575	0.7824	0.5601	0.3245	0.7037	0.3518	0.2100
Deep-QPP (SDMQ)	**0.8420**	**0.7404**	**0.5434**	**0.8045**	**0.5532**	**0.3130**	**0.8371**	**0.6315**	**0.4614**	**0.7903**	**0.4431**	**0.2554**

Purely representation-based or purely interaction-based approaches perform worse than **Deep-QPP**.

A Pointwise-Query:Listwise-Document based QPP Approach (SIGIR'22)

A BERT-based End-to-end Model

An end-to-end neural cross-encoder-based approach - trained **pointwise** on individual queries, but **listwise** over the top ranked documents (split into chunks).

– **Datta, S.**, MacAvaney, S., Ganguly, D., Greene, D. A 'Pointwise-Query, Listwise-Document' based Query Performance Prediction Approach (to appear in the proceedings of SIGIR'22).

Suchana Datta
University College Dublin

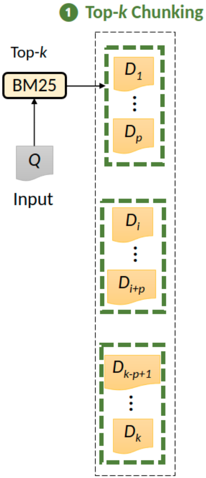
Sean MacAvaney
University of Glasgow

Derek Greene
University College Dublin

A BERT-based End-to-end Model

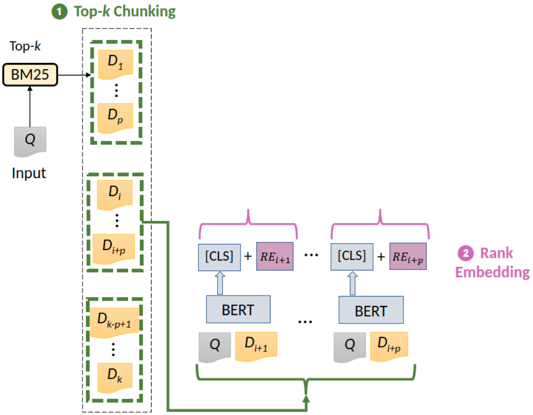
- A novel architecture and objective function for a **pointwise** neural QPP.
- Transformed the pointwise QPP objective into a **classification task**, not a regression model.
- Models the top-ranked documents as a **sequence of chunks (Listwise)**, not as a whole set.
- Incorporates the **relative Positions (or ranks)** of the top documents.

End-to-end Architecture of qppBERT-PL



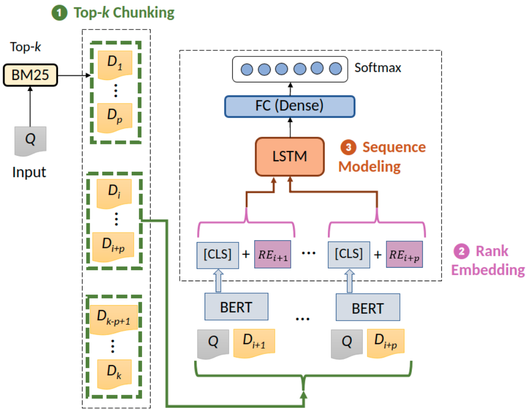
- Popular unsupervised QPP methods (e.g. NQC, WIG) work well when information used from the top-100 documents.
- Encoding long sequences of 100 documents is likely to be noisy.
- Top-ranked set is segmented into equal sized partitions (chunks).

End-to-end Architecture of qppBERT-PL



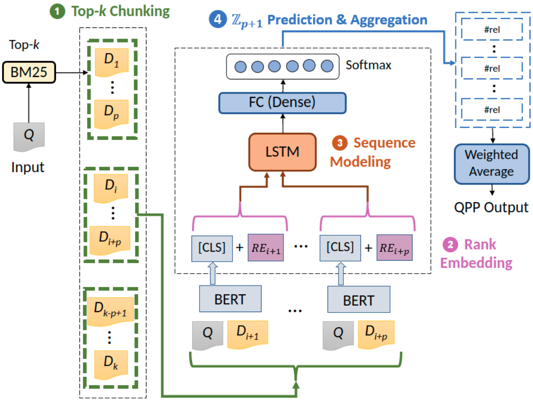
- BERT-based cross-encoder is used to model the interactions between the query and the document terms of each chunk.
- LSTM-encoded representation of this interaction sequence.
- Ranks are encoded via BERT positional embeddings.

End-to-end Architecture of qppBERT-PL



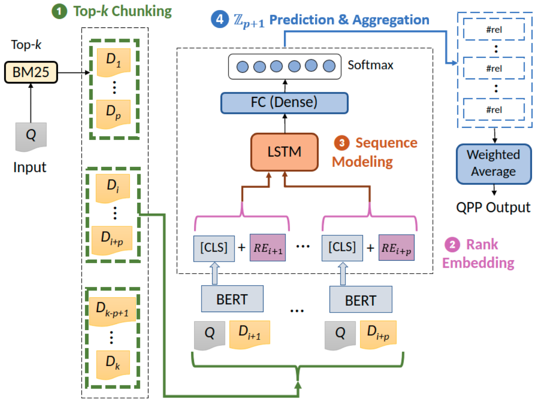
- Passed through a fully connected layer (FC).
- Terminates at a $p+1$ dimensional $\mathrm{Softmax}$ representing the probability of finding r relevant documents within this p-sized chunk ($r \in \{0, 1, \ldots, p\}$).

End-to-end Architecture of qppBERT-PL



- Compute a weighted average from the outputs of the network, predicted for each p-sized partition of the top documents.
- Aggregated scores are used to sort the queries in descending order.

End-to-end Architecture of qppBERT-PL



- Compute a weighted average from the outputs of the network, predicted for each p-sized partition of the top documents.
- Aggregated scores are used to sort the queries in descending order.

Available at: `https://github.com/suchanadatta/qppBERT-PL.git`

Performance of qppBERT-PL

Type	Models	MS MARCO Dev				TREC-DL'19				TREC-DL'20			
		MRR@10		AP@100		MRR@10		AP@100		MRR@10		AP@100	
		P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ
Baselines	NQC	0.331	0.298	0.285	0.227	0.239	0.185	0.183	0.107	0.259	0.243	0.179	0.124
	Clarity	0.173	0.248	0.172	0.207	0.156	0.147	0.096	0.113	0.239	0.215	0.107	0.129
	WIG	0.193	0.215	0.215	0.203	0.192	0.133	0.133	0.089	0.260	0.241	0.143	0.096
	UEF(NQC)	0.347	0.313	0.294	0.227	0.254	0.235	0.189	0.112	0.275	0.291	0.200	0.126
	SCNQC	0.334	0.310	0.304	0.228	0.261	0.251	0.204	0.123	0.284	0.298	0.215	0.141
	NeuralQPP	0.215	0.197	0.173	0.193	0.156	0.126	0.129	0.133	0.271	0.253	0.133	0.112
	BERT-QPP	0.520	0.411	0.326	0.301	0.350	0.363	0.268	0.202	0.343	0.341	0.233	0.195
	+ Seq.	0.463	0.360	0.301	0.312	0.345	0.333	0.265	0.193	0.277	0.218	0.258	0.190
	+ Seq. + RankEmb	0.473	0.370	0.328	0.285	0.323	0.332	0.253	0.167	0.303	0.236	0.252	0.172
Ours	qppBERT-PL	**0.562**	**0.448**	**0.354**	**0.327**	**0.413**	**0.403**	**0.301**	**0.247**	**0.422**	**0.392**	**0.303**	**0.251**
	– Seq.	0.512	0.386	0.303	0.283	0.357	0.349	0.274	0.193	0.345	0.320	0.271	0.200
	– Chunked	0.520	0.413	0.331	0.274	0.373	0.326	0.290	0.225	0.370	0.333	0.297	0.231
	– RankEmb	0.519	0.392	0.320	0.267	0.361	0.328	0.285	0.232	0.352	0.331	0.293	0.215
	– Chunked – RankEmb	0.405	0.329	0.293	0.285	0.309	0.299	0.260	0.159	0.217	0.198	0.199	0.184

qppBERT-PL is more effective at predicting query performance than other supervised and unsupervised methods.

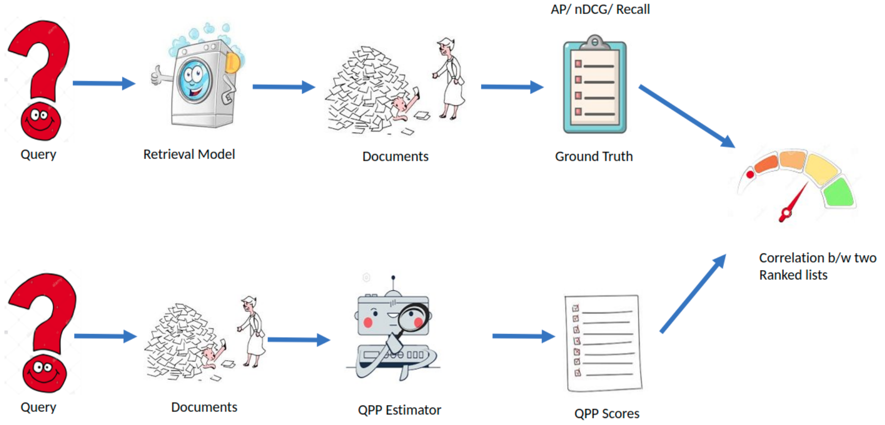
Performance of qppBERT-PL

Type	Models	MS MARCO Dev				TREC-DL'19				TREC-DL'20			
		MRR@10		AP@100		MRR@10		AP@100		MRR@10		AP@100	
		P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ
Baselines	NQC	0.331	0.298	0.285	0.227	0.239	0.185	0.183	0.107	0.259	0.243	0.179	0.124
	Clarity	0.173	0.248	0.172	0.207	0.156	0.147	0.096	0.113	0.239	0.215	0.107	0.129
	WIG	0.193	0.215	0.215	0.203	0.192	0.133	0.133	0.089	0.260	0.241	0.143	0.096
	UEF(NQC)	0.347	0.313	0.294	0.227	0.254	0.235	0.189	0.112	0.275	0.291	0.200	0.126
	SCNQC	0.334	0.310	0.304	0.228	0.261	0.251	0.204	0.123	0.284	0.298	0.215	0.141
	NeuralQPP	0.215	0.197	0.173	0.193	0.156	0.126	0.129	0.133	0.271	0.253	0.133	0.112
	BERT-QPP	0.520	0.411	0.326	0.301	0.350	0.363	0.268	0.202	0.343	0.341	0.233	0.195
	+ Seq.	0.463	0.360	0.301	0.312	0.345	0.333	0.265	0.193	0.277	0.218	0.258	0.190
	+ Seq. + RankEmb	0.473	0.370	0.328	0.285	0.323	0.332	0.253	0.167	0.303	0.236	0.252	0.172
Ours	qppBERT-PL	**0.562**	**0.448**	**0.354**	**0.327**	**0.413**	**0.403**	**0.301**	**0.247**	**0.422**	**0.392**	**0.303**	**0.251**
	– Seq.	0.512	0.386	0.303	0.283	0.357	0.349	0.274	0.193	0.345	0.320	0.271	0.200
	– Chunked	0.520	0.413	0.331	0.274	0.373	0.326	0.290	0.225	0.370	0.333	0.297	0.231
	– RankEmb	0.519	0.392	0.320	0.267	0.361	0.328	0.285	0.232	0.352	0.331	0.293	0.215
	– Chunked – RankEmb	0.405	0.329	0.293	0.285	0.309	0.299	0.260	0.159	0.217	0.198	0.199	0.184

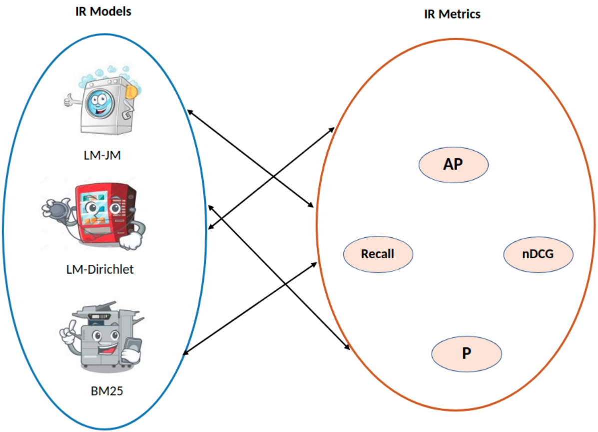
Sequence modeling, chunking and Rank Embeddings are critical components of qppBERT-PL.

Analyzing the Sensitivity of QPP Evaluation (ECIR'22)

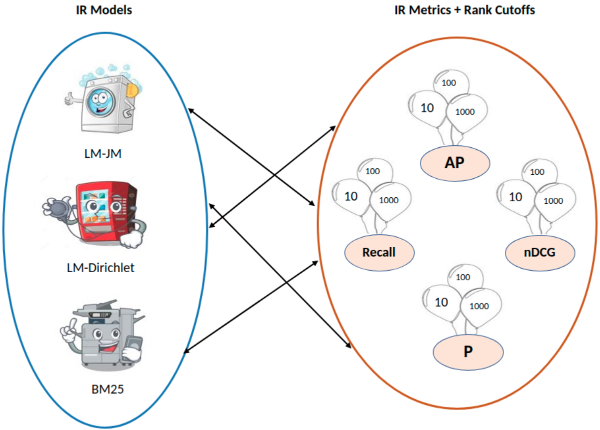
How Do We Evaluate QPP Estimators?



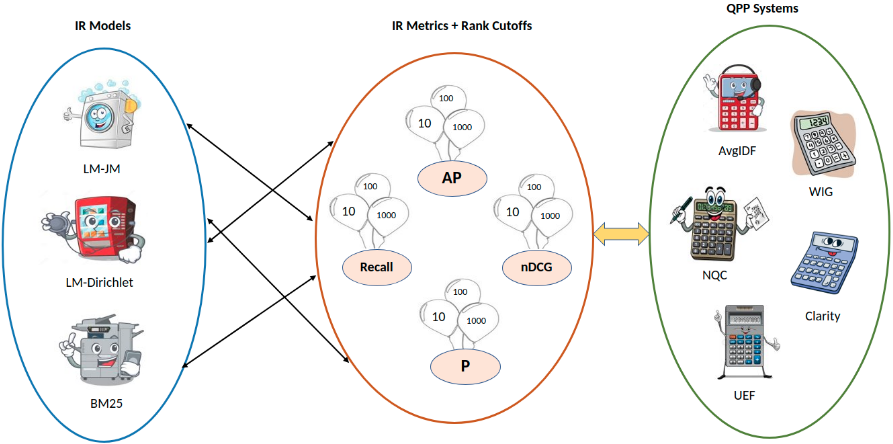
There are too many combinations!



There are too many combinations!



There are too many combinations!



Research Question

RQ1: Do variations in the QPP context, in terms of the **IR metric**, the **IR model**, and the **rank cut-off** used to construct the QPP evaluation ground-truth, lead to significant differences in outcome of a QPP method?

Research Question

RQ1: Do variations in the QPP context, in terms of the **IR metric**, the **IR model**, and the **rank cut-off** used to construct the QPP evaluation ground-truth, lead to significant differences in outcome of a QPP method?

- We measure the sensitivity of QPP results with variations in the IR evaluation metric and the IR model for the QPP methods.
- We compute the *standard deviations* in the observed values for different QPP experiment setup.

What Do We Observe?

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
	LMJM	0.3795	0.3966	0.3869	0.3311	**0.0291**
r	BM25	0.5006	0.4879	0.4813	0.2525	**0.1190**
	LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
	$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
	LMJM	0.4553	0.4697	0.4663	0.3067	0.0788
ρ	BM25	0.4526	0.4700	0.4736	0.2842	0.0911
	LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
	$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
	LMJM	0.3175	0.3285	0.3278	0.2193	0.0529
τ	BM25	0.3144	0.3162	0.3319	0.2040	0.0589
	LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
	$\sigma(\mathcal{S})$	0.0087	0.0123	**0.0084**	0.0120	

(a) AvgIDF

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
	LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
r	BM25	0.3563	0.4118	0.4495	0.2707	0.0777
	LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
	$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
	LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
ρ	BM25	0.4618	0.4887	0.5137	0.3308	0.0814
	LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
	$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
	LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
τ	BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
	LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
	$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
	LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
r	BM25	0.4488	0.4563	0.4386	0.3485	0.0502
	LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
	$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
	LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
ρ	BM25	0.4520	0.4601	0.4505	0.3586	0.0480
	LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
	$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
	LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
τ	BM25	0.3116	0.3181	0.3125	0.2549	0.0297
	LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
	$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
	LMJM	0.4746	0.4763	0.4646	0.3573	0.0575
r	BM25	0.5386	0.5476	0.5263	0.4182	0.0603
	LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
	$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
	LMJM	0.4385	0.4477	0.4472	0.3682	0.0384
ρ	BM25	0.5334	0.5429	0.5316	0.4231	0.0567
	LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
	$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
	LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
τ	BM25	0.3677	0.3754	0.3688	0.3008	0.0351
	LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
	$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

Variations due to IR Evaluation Metrics

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3795	0.3966	0.3869	0.3311	**0.0291**
	BM25	0.5006	0.4879	0.4813	0.2525	**0.1190**
	LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
	$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
ρ	LMJM	0.4553	0.4697	0.4663	0.3067	0.0788
	BM25	0.4526	0.4700	0.4736	0.2842	0.0911
	LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
	$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
τ	LMJM	0.3175	0.3285	0.3278	0.2193	0.0529
	BM25	0.3144	0.3162	0.3319	0.2040	0.0589
	LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
	$\sigma(\mathcal{S})$	0.0087	0.0123	**0.0084**	0.0120	

(a) AvgIDF

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
	BM25	0.3563	0.4118	0.4495	0.2707	0.0777
	LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
	$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
ρ	LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
	BM25	0.4618	0.4887	0.5137	0.3308	0.0814
	LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
	$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
τ	LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
	BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
	LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
	$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
	BM25	0.4488	0.4563	0.4386	0.3485	0.0502
	LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
	$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
ρ	LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
	BM25	0.4520	0.4601	0.4505	0.3586	0.0480
	LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
	$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
τ	LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
	BM25	0.3116	0.3181	0.3125	0.2549	0.0297
	LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
	$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4746	0.4763	0.4646	0.3573	0.0575
	BM25	0.5386	0.5476	0.5263	0.4182	0.0603
	LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
	$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
ρ	LMJM	0.4385	0.4477	0.4472	0.3682	0.0384
	BM25	0.5334	0.5429	0.5316	0.4231	0.0567
	LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
	$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
τ	LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
	BM25	0.3677	0.3754	0.3688	0.3008	0.0351
	LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
	$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

- Substantial absolute differences in the QPP outcomes.

Variations due to IR Evaluation Metrics

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.3795	0.3966	0.3869	0.3311	**0.0291**
	BM25	0.5006	0.4879	0.4813	0.2525	**0.1190**
	LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
	$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
ρ	LMJM	0.4553	0.4697	0.4663	0.3067	0.0788
	BM25	0.4526	0.4700	0.4736	0.2842	0.0911
	LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
	$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
τ	LMJM	0.3175	0.3285	0.3278	0.2193	0.0529
	BM25	0.3144	0.3162	0.3319	0.2040	0.0589
	LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
	$\sigma(\mathcal{S})$	0.0087	0.0123	**0.0084**	0.0120	

(a) AvgIDF

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
	BM25	0.3563	0.4118	0.4495	0.2707	0.0777
	LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
	$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
ρ	LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
	BM25	0.4618	0.4887	0.5137	0.3308	0.0814
	LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
	$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
τ	LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
	BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
	LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
	$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
	BM25	0.4488	0.4563	0.4386	0.3485	0.0502
	LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
	$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
ρ	LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
	BM25	0.4520	0.4601	0.4505	0.3586	0.0480
	LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
	$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
τ	LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
	BM25	0.3116	0.3181	0.3125	0.2549	0.0297
	LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
	$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.4746	0.4763	0.4646	0.3573	0.0575
	BM25	0.5386	0.5476	0.5263	0.4182	0.0603
	LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
	$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
ρ	LMJM	0.4385	0.4477	0.4472	0.3682	0.0384
	BM25	0.5334	0.5429	0.5316	0.4231	0.0567
	LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
	$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
τ	LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
	BM25	0.3677	0.3754	0.3688	0.3008	0.0351
	LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
	$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

- Substantial absolute differences in the QPP outcomes.
- Lower variations with Kendall's τ.

Variations due to IR Evaluation Metrics

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3795	0.3966	0.3869	0.331	**0.0291**
	BM25	0.5006	0.4879	0.4813	0.2525	**0.1190**
	LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
	$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
ρ	LMJM	0.4553	0.4697	0.4663	0.306	0.0788
	BM25	0.4526	0.4700	0.4736	0.2842	0.0911
	LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
	$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
τ	LMJM	0.3175	0.3285	0.3278	0.219	0.0529
	BM25	0.3144	0.3162	0.3319	0.2040	0.0589
	LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
	$\sigma(\mathcal{S})$	0.0087	0.0123	**0.0084**	0.0120	

(a) AvgIDF

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
	BM25	0.3563	0.4118	0.4495	0.2707	0.0777
	LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
	$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
ρ	LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
	BM25	0.4618	0.4887	0.5137	0.3308	0.0814
	LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
	$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
τ	LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
	BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
	LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
	$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
	BM25	0.4488	0.4563	0.4386	0.3485	0.0502
	LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
	$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
ρ	LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
	BM25	0.4520	0.4601	0.4505	0.3586	0.0480
	LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
	$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
τ	LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
	BM25	0.3116	0.3181	0.3125	0.2549	0.0297
	LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
	$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4746	0.4763	0.4646	0.357	0.0575
	BM25	0.5386	0.5476	0.5263	0.4182	0.0603
	LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
	$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
ρ	LMJM	0.4385	0.4477	0.4472	0.368	0.0384
	BM25	0.5334	0.5429	0.5316	0.4231	0.0567
	LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
	$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
τ	LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
	BM25	0.3677	0.3754	0.3688	0.3008	0.0351
	LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
	$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

- Substantial absolute differences in the QPP outcomes.
- Lower variations with Kendall's τ.
- Lower variances with LMJM.

Variations due to IR Models

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3795	0.3966	0.3869	0.3311	**0.0291**
r	BM25	0.5006	0.4879	0.4813	0.2525	**0.1190**
r	LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
	$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
ρ	LMJM	0.4553	0.4697	0.4663	0.3067	0.0788
ρ	BM25	0.4526	0.4700	0.4736	0.2842	0.0911
ρ	LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
	$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
τ	LMJM	0.3175	0.3285	0.3278	0.2193	0.0529
τ	BM25	0.3144	0.3162	0.3319	0.2040	0.0589
τ	LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
	$\sigma(\mathcal{S})$	0.0087	0.0125	**0.0084**	0.0120	

(a) AvgIDF

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
r	BM25	0.3563	0.4118	0.4495	0.2707	0.0777
r	LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
	$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
ρ	LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
ρ	BM25	0.4618	0.4887	0.5137	0.3308	0.0814
ρ	LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
	$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
τ	LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
τ	BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
τ	LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
	$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
r	BM25	0.4488	0.4563	0.4386	0.3485	0.0502
r	LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
	$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
ρ	LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
ρ	BM25	0.4520	0.4601	0.4505	0.3586	0.0480
ρ	LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
	$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
τ	LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
τ	BM25	0.3116	0.3181	0.3125	0.2549	0.0297
τ	LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
	$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4746	0.4763	0.4646	0.3573	0.0575
r	BM25	0.5386	0.5476	0.5263	0.4182	0.0603
r	LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
	$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
ρ	LMJM	0.4385	0.4477	0.4472	0.3682	0.0384
ρ	BM25	0.5334	0.5429	0.5316	0.4231	0.0567
ρ	LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
	$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
τ	LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
τ	BM25	0.3677	0.3754	0.3688	0.3008	0.0351
τ	LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
	$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

- Lower variations with Kendall's τ.

Variations due to IR Models

	Model($\mathcal{S}$)	IR Evaluation Metric (θ) AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.3795	0.3966	0.3869	0.3311	**0.0291**
	BM25	0.5006	0.4879	0.4813	0.2526	**0.1190**
	LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
	$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
ρ	LMJM	0.4553	0.4697	0.4663	0.3067	0.0788
	BM25	0.4526	0.4700	0.4736	0.2842	0.0911
	LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
	$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
τ	LMJM	0.3175	0.3285	0.3278	0.2193	0.0529
	BM25	0.3144	0.3162	0.3319	0.2040	0.0589
	LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
	$\sigma(\mathcal{S})$	0.0087	0.0123	**0.0084**	0.0120	

(a) AvgIDF

	Model($\mathcal{S}$)	IR Evaluation Metric (θ) AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
	BM25	0.3563	0.4118	0.4495	0.2707	0.0777
	LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
	$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
ρ	LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
	BM25	0.4618	0.4887	0.5137	0.3308	0.0814
	LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
	$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
τ	LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
	BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
	LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
	$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	Model($\mathcal{S}$)	IR Evaluation Metric (θ) AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
	BM25	0.4488	0.4563	0.4386	0.3485	0.0502
	LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
	$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
ρ	LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
	BM25	0.4520	0.4601	0.4505	0.3586	0.0480
	LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
	$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
τ	LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
	BM25	0.3116	0.3181	0.3125	0.2549	0.0297
	LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
	$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	Model($\mathcal{S}$)	IR Evaluation Metric (θ) AP	nDCG	R	P@10	$\sigma(\theta)$
r	LMJM	0.4746	0.4763	0.4646	0.3573	0.0575
	BM25	0.5386	0.5476	0.5263	0.4182	0.0603
	LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
	$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
ρ	LMJM	0.4385	0.4477	0.4472	0.3682	0.0384
	BM25	0.5334	0.5429	0.5316	0.4231	0.0567
	LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
	$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
τ	LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
	BM25	0.3677	0.3754	0.3688	0.3008	0.0351
	LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
	$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

- Lower variations with Kendall's τ.
- Lower variations across IR models than IR metrics.

Variations due to IR Models

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3795	0.3966	0.3869	0.3311	**0.0291**
	BM25	0.5006	0.4879	0.4813	0.2525	**0.1190**
	LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
	$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
ρ	LMJM	0.4553	0.4697	0.4663	0.3067	0.0788
	BM25	0.4526	0.4700	0.4736	0.2842	0.0911
	LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
	$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
τ	LMJM	0.3175	0.3285	0.3278	0.2193	0.0529
	BM25	0.3144	0.3162	0.3319	0.2040	0.0589
	LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
	$\sigma(\mathcal{S})$	0.0087	0.0123	**0.0084**	0.0120	

(a) AvgIDF

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
	BM25	0.3563	0.4118	0.4495	0.2707	0.0777
	LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
	$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
ρ	LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
	BM25	0.4618	0.4887	0.5137	0.3308	0.0814
	LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
	$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
τ	LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
	BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
	LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
	$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
	BM25	0.4488	0.4563	0.4386	0.3485	0.0502
	LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
	$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
ρ	LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
	BM25	0.4520	0.4601	0.4505	0.3586	0.0480
	LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
	$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
τ	LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
	BM25	0.3116	0.3181	0.3125	0.2549	0.0297
	LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
	$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
		IR Evaluation Metric (θ)				
r	LMJM	0.4746	0.4763	0.4646	0.3573	0.0575
	BM25	0.5386	0.5476	0.5263	0.4182	0.0603
	LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
	$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
ρ	LMJM	0.4385	0.4477	0.4472	0.3682	0.0384
	BM25	0.5334	0.5429	0.5316	0.4231	0.0567
	LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
	$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
τ	LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
	BM25	0.3677	0.3754	0.3688	0.3008	0.0351
	LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
	$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

- Lower variations with Kendall's τ.
- Lower variations across IR models than IR metrics.
- Lack of consistency on which combination of QPP method with IR evaluation context yields the least variance.

Research Question

RQ2: Do these variations lead to **significant differences in the relative ranks of different QPP methods**?

Model	Metric	AP@100	AP@1000	R@10	R@100	R@1000	nDCG@10	nDCG@100	nDCG@1000
LMJM	AP@10	0.4286	0.3333	0.9048	0.2381	**-0.1429**	1.0000	0.2381	0.3333
BM25	AP@10	1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir	AP@10	1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	AP@100		0.9048	0.5238	0.8095	0.4286	0.4286	0.8095	0.9048
BM25	AP@100		0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir	AP@100		0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	AP@1000			0.4286	0.8095	0.5238	0.3333	0.9048	1.0000
BM25	AP@1000			0.9048	0.8095	0.3333	0.9048	0.9048	0.8095
LMDir	AP@1000			0.9048	0.8095	0.5238	0.9048	0.9048	0.8095
LMJM	R@10				0.3333	-0.0476	0.9048	0.3333	0.4286
BM25	R@10				0.9048	0.4286	1.0000	1.0000	0.7143
LMDir	R@10				0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	R@100					0.6190	0.2381	1.0000	0.9048
BM25	R@100					0.5238	0.9048	0.9048	0.6190
LMDir	R@100					0.5238	0.9048	0.9048	0.6190
LMJM	R@1000						**-0.1429**	0.6190	0.5238
BM25	R@1000						0.4286	0.4286	0.5238
LMDir	R@1000						0.4286	0.4286	0.5238
LMJM	nDCG@10							0.2381	0.3333
BM25	nDCG@10							1.0000	0.7143
LMDir	nDCG@10							1.0000	0.7143
LMJM	nDCG@100								0.9048
BM25	nDCG@100								0.7143
LMDir	nDCG@100								0.7143

- Each cell indicates the correlation (Kendall's τ) between QPP systems ranked in order by their evaluated effectiveness.
- A total of 7 QPP systems were used in these experiments - AvgIDF, Clarity, WIG, NQC, UEF(Clarity), UEF(WIG) and UEF(NQC).
- The lowest correlation for each group is in red and the lowest correlations, overall, are bold-faced.

Variations due to IR Evaluation Metrics

Model	Metric	AP@100	AP@1000	R@10	R@100	R@1000	nDCG@10	nDCG@100	nDCG@1000
LMJM	AP@10	0.4286	0.3333	0.9048	0.2381	**-0.1429**	1.0000	0.2381	0.3333
BM25	AP@10	1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir	AP@10	1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	AP@100		0.9048	0.5238	0.8095	0.4286	0.4286	0.8095	0.9048
BM25	AP@100		0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir	AP@100		0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	AP@1000			0.4286	0.8095	0.5238	0.3333	0.9048	1.0000
BM25	AP@1000			0.9048	0.8095	0.3333	0.9048	0.9048	0.8095
LMDir	AP@1000			0.9048	0.8095	0.5238	0.9048	0.9048	0.8095
LMJM	R@10				0.3333	-0.0476	0.9048	0.3333	0.4286
BM25	R@10				0.9048	0.4286	1.0000	1.0000	0.7143
LMDir	R@10				0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	R@100					0.6190	0.2381	1.0000	0.9048
BM25	R@100					0.5238	0.9048	0.9048	0.6190
LMDir	R@100					0.5238	0.9048	0.9048	0.6190
LMJM	R@1000						**-0.1429**	0.6190	0.5238
BM25	R@1000						0.4286	0.4286	0.5238
LMDir	R@1000						0.4286	0.4286	0.5238
LMJM	nDCG@10							0.2381	0.3333
BM25	nDCG@10							1.0000	0.7143
LMDir	nDCG@10							1.0000	0.7143
LMJM	nDCG@100								0.9048
BM25	nDCG@100								0.7143
LMDir	nDCG@100								0.7143

- LMJM leads to the most instability in the relative ranks.

Variations due to IR Evaluation Metrics

Model	Metric	AP@100	AP@1000	R@10	R@100	R@1000	nDCG@10	nDCG@100	nDCG@1000
LMJM	AP@10	0.4286	0.3333	0.9048	0.2381	**-0.1429**	1.0000	0.2381	0.3333
BM25		1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir		1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	AP@100		0.9048	0.5238	0.8095	0.4286	0.4286	0.8095	0.9048
BM25			0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir			0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	AP@1000			0.4286	0.8095	0.5238	0.3333	0.9048	1.0000
BM25				0.9048	0.8095	0.3333	0.9048	0.9048	0.8095
LMDir				0.9048	0.8095	0.5238	0.9048	0.9048	0.8095
LMJM	R@10				0.3333	-0.0476	0.9048	0.3333	0.4286
BM25					0.9048	0.4286	1.0000	1.0000	0.7143
LMDir					0.9048	0.4286	1.0000	1.0000	0.7143
LMJM	R@100					0.6190	0.2381	1.0000	0.9048
BM25						0.5238	0.9048	0.9048	0.6190
LMDir						0.5238	0.9048	0.9048	0.6190
LMJM	R@1000						**-0.1429**	0.6190	0.5238
BM25							0.4286	0.4286	0.5238
LMDir							0.4286	0.4286	0.5238
LMJM	nDCG@10							0.2381	0.3333
BM25								1.0000	0.7143
LMDir								1.0000	0.7143
LMJM	nDCG@100								0.9048
BM25									0.7143
LMDir									0.7143

- LMJM leads to the most instability in the relative ranks.
- Some evaluation metrics are more sensitive to rank cut-off values.

Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (1.0, 1.0)	BM25 (0.3, 0.7)	LMDir (100)	LMDir (500)	LMDir (1000)
AP@100		1.0000	0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100	LMJM	1.0000	0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100	(0.3)	0.9048	0.8095	0.9048	1.0000	1.0000	0.9048	0.9048
P@10		1.0000	0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100			0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100	LMJM		0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100	(0.6)		0.9048	1.0000	0.9048	0.9048	1.0000	1.0000
P@10			0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100				0.9048	0.9048	1.0000	1.0000	1.0000
nDCG@100	BM25			0.9048	0.9048	0.9048	1.0000	1.0000
R@100	(0.7, 0.3)			0.9048	0.8095	0.8095	0.9048	0.9048
P@10				0.8095	1.0000	0.9048	0.9048	0.8095
AP@100					0.9048	0.9048	0.9048	0.9048
nDCG@100	BM25				1.0000	1.0000	0.9048	0.9048
R@100	(1.0, 1.0)				0.9048	0.9048	1.0000	1.0000
P@10					0.8095	**0.7143**	**0.7143**	1.0000
AP@100						1.0000	1.0000	1.0000
nDCG@100	BM25					1.0000	0.9048	0.9048
R@100	(0.3, 0.7)					1.0000	0.9048	0.9048
P@10						0.9048	0.9048	0.8095
AP@100							1.0000	1.0000
nDCG@100	LMDir						0.9048	0.9048
R@100	(100)						0.9048	0.9048
P@10							0.8095	**0.7143**
AP@100								1.0000
nDCG@100	LMDir							1.0000
R@100	(500)							1.0000
P@10								**0.7143**

- Each cell in the table indicates the correlation (Kendall's τ) between QPP systems ranked in order by their evaluated effectiveness.
- 7 QPP systems were used in these experiments.
- The lowest correlation for each group is in red and the lowest correlations, overall, are bold-faced.

Variations due to IR Models

Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (1.0, 1.0)	BM25 (0.3, 0.7)	LMDir (100)	LMDir (500)	LMDir (1000)
AP@100	LMJM (0.3)	1.0000	0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100		1.0000	0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100		0.9048	0.8095	0.9048	1.0000	1.0000	0.9048	0.9048
P@10		1.0000	0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100	LMJM (0.6)		0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100			0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100			0.9048	1.0000	0.9048	0.9048	1.0000	1.0000
P@10			0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100	BM25 (0.7, 0.3)			0.9048	0.9048	1.0000	1.0000	1.0000
nDCG@100				0.9048	0.9048	0.9048	1.0000	1.0000
R@100				0.9048	0.8095	0.8095	0.9048	0.9048
P@10				0.8095	1.0000	0.9048	0.9048	0.8095
AP@100	BM25 (1.0, 1.0)				0.9048	0.9048	0.9048	0.9048
nDCG@100					1.0000	1.0000	0.9048	0.9048
R@100					0.9048	0.9048	1.0000	1.0000
P@10					0.8095	**0.7143**	**0.7143**	1.0000
AP@100	BM25 (0.3, 0.7)					1.0000	1.0000	1.0000
nDCG@100						1.0000	0.9048	0.9048
R@100						1.0000	0.9048	0.9048
P@10						0.9048	0.9048	0.8095
AP@100	LMDir (100)						1.0000	1.0000
nDCG@100							0.9048	0.9048
R@100							0.9048	0.9048
P@10							0.8095	**0.7143**
AP@100	LMDir (500)							1.0000
nDCG@100								1.0000
R@100								1.0000
P@10								**0.7143**

- Relative ranks of QPP systems are quite stable across IR models.

Variations due to IR Models

Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (1.0, 1.0)	BM25 (0.3, 0.7)	LMDir (100)	LMDir (500)	LMDir (1000)
AP@100	LMJM (0.3)	1.0000	0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100		1.0000	0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100		0.9048	0.8095	0.9048	1.0000	1.0000	0.9048	0.9048
P@10		1.0000	0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100	LMJM (0.6)		0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100			0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100			0.9048	1.0000	0.9048	0.9048	1.0000	1.0000
P@10			0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100	BM25 (0.7, 0.3)			0.9048	0.9048	1.0000	1.0000	1.0000
nDCG@100				0.9048	0.9048	0.9048	1.0000	1.0000
R@100				0.9048	0.8095	0.8095	0.9048	0.9048
P@10				0.8095	1.0000	0.9048	0.9048	0.8095
AP@100	BM25 (1.0, 1.0)				0.9048	0.9048	0.9048	0.9048
nDCG@100					1.0000	1.0000	0.9048	0.9048
R@100					0.9048	0.9048	1.0000	1.0000
P@10					0.8095	**0.7143**	**0.7143**	1.0000
AP@100	BM25 (0.3, 0.7)					1.0000	1.0000	1.0000
nDCG@100						1.0000	0.9048	0.9048
R@100						1.0000	0.9048	0.9048
P@10						0.9048	0.9048	0.8095
AP@100	LMDir (100)						1.0000	1.0000
nDCG@100							0.9048	0.9048
R@100							0.9048	0.9048
P@10							0.8095	**0.7143**
AP@100	LMDir (500)							1.0000
nDCG@100								1.0000
R@100								1.0000
P@10								**0.7143**

- Relative ranks of QPP systems are quite stable across IR models.
- LMJM leads to more instability in the QPP outcomes.

Variations due to IR Models

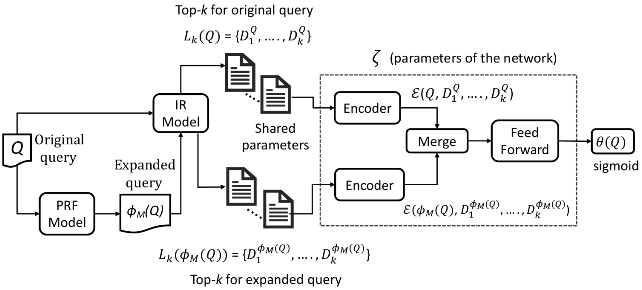
Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (1.0, 1.0)	BM25 (0.3, 0.7)	LMDir (100)	LMDir (500)	LMDir (1000)
AP@100	LMJM (0.3)	1.0000	0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100		1.0000	0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100		0.9048	0.8095	0.9048	1.0000	1.0000	0.9048	0.9048
P@10		1.0000	0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100	LMJM (0.6)		0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100			0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100			0.9048	1.0000	0.9048	0.9048	1.0000	1.0000
P@10			0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100	BM25 (0.7, 0.3)			0.9048	0.9048	1.0000	1.0000	1.0000
nDCG@100				0.9048	0.9048	0.9048	1.0000	1.0000
R@100				0.9048	0.8095	0.8095	0.9048	0.9048
P@10				0.8095	1.0000	0.9048	0.9048	0.8095
AP@100	BM25 (1.0, 1.0)				0.9048	0.9048	0.9048	0.9048
nDCG@100					1.0000	1.0000	0.9048	0.9048
R@100					0.9048	0.9048	1.0000	1.0000
P@10					0.8095	**0.7143**	**0.7143**	1.0000
AP@100	BM25 (0.3, 0.7)					1.0000	1.0000	1.0000
nDCG@100						1.0000	0.9048	0.9048
R@100						1.0000	0.9048	0.9048
P@10						0.9048	0.9048	0.8095
AP@100	LMDir (100)						1.0000	1.0000
nDCG@100							0.9048	0.9048
R@100							0.9048	0.9048
P@10							0.8095	**0.7143**
AP@100	LMDir (500)							1.0000
nDCG@100								1.0000
R@100								1.0000
P@10								**0.7143**

- Relative ranks of QPP systems are quite stable across IR models.
- LMJM leads to more instability in the QPP outcomes.
- Relative ranks of QPP systems are more stable with Kendall's τ.

Ongoing work (Submitted to WSDM'23)

Adaptive Pseudo-relevance Feedback

- A supervised approach to QPP works quite well!
- QPP prediction can let us decide whether to apply PRF or not.



Concluding Remarks

Concluding Remarks

- Summary:
  - QPP experiments are very sensitive to the metric used for evaluation, and the IR model which derives a ranked list.
  - A purely data-driven early interaction-based model improves QPP.
  - Transformer-based approach further improves results; however, training and inference slower compared to 2DCNN.
  - Supervised QPP also useful for adaptive relevance feedback.
- Future Directions:
  - QPP provides an estimate about the quality of a model's prediction.
  - Can be used for recommender systems - because the output is a ranked list of top-k items; Query $\mapsto$ Context of a user.
  - Can also be used for other prediction systems; $\theta : \vec{x} \mapsto \mathbb{Z}$ could be transformed to $\phi : \vec{x}, \vec{z} \mapsto \mathbb{R}$.
  - QPP can be extended to *sessions of queries*.

Thank you!

For any questions you may have, please e-mail me at:

Debasis.Ganguly@glasgow.ac.uk