On the Reproduciblity and Robustness of QPP Experiments

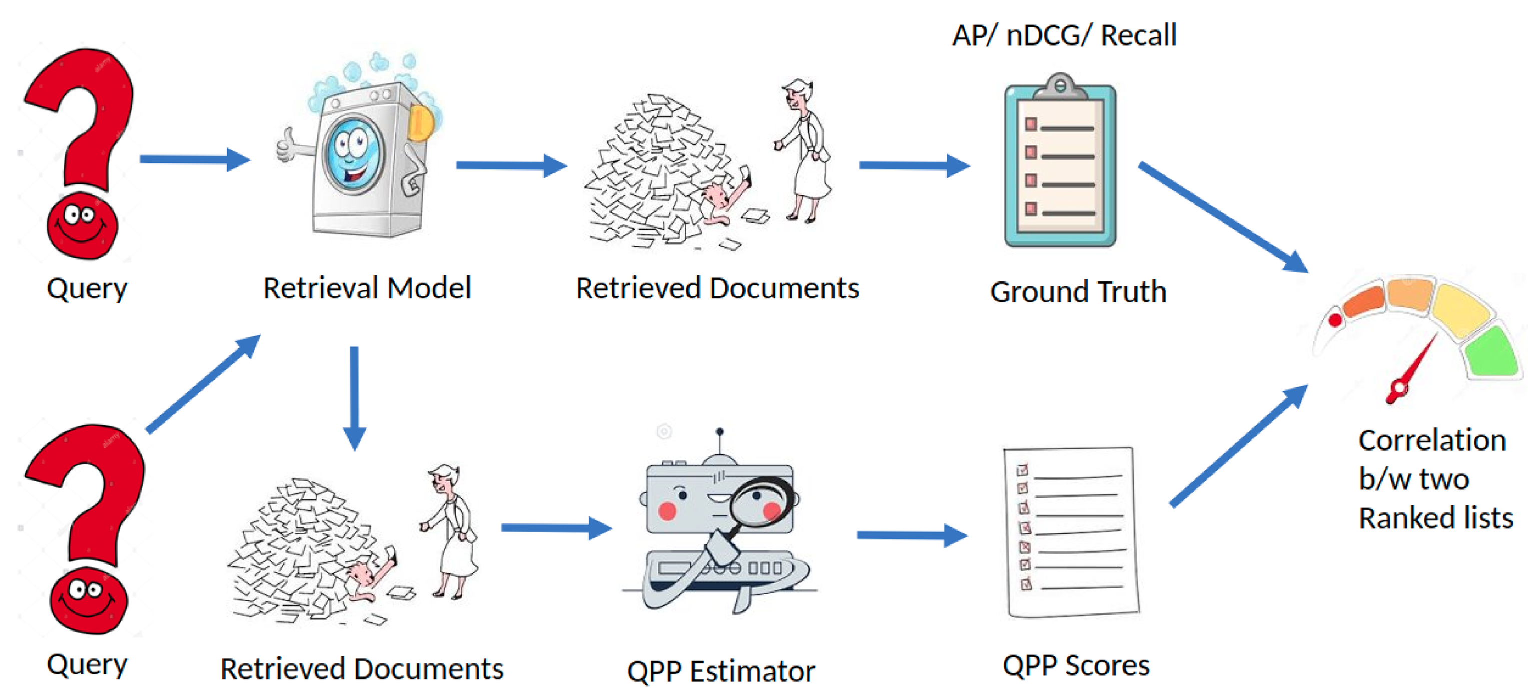
Suchana Datta [1] Debasis Ganguly [2] Mandar Mitra [3] Derek Greene [1]

[1]University College Dublin [2]University of Glasgow [3]Indian Statistical Institute

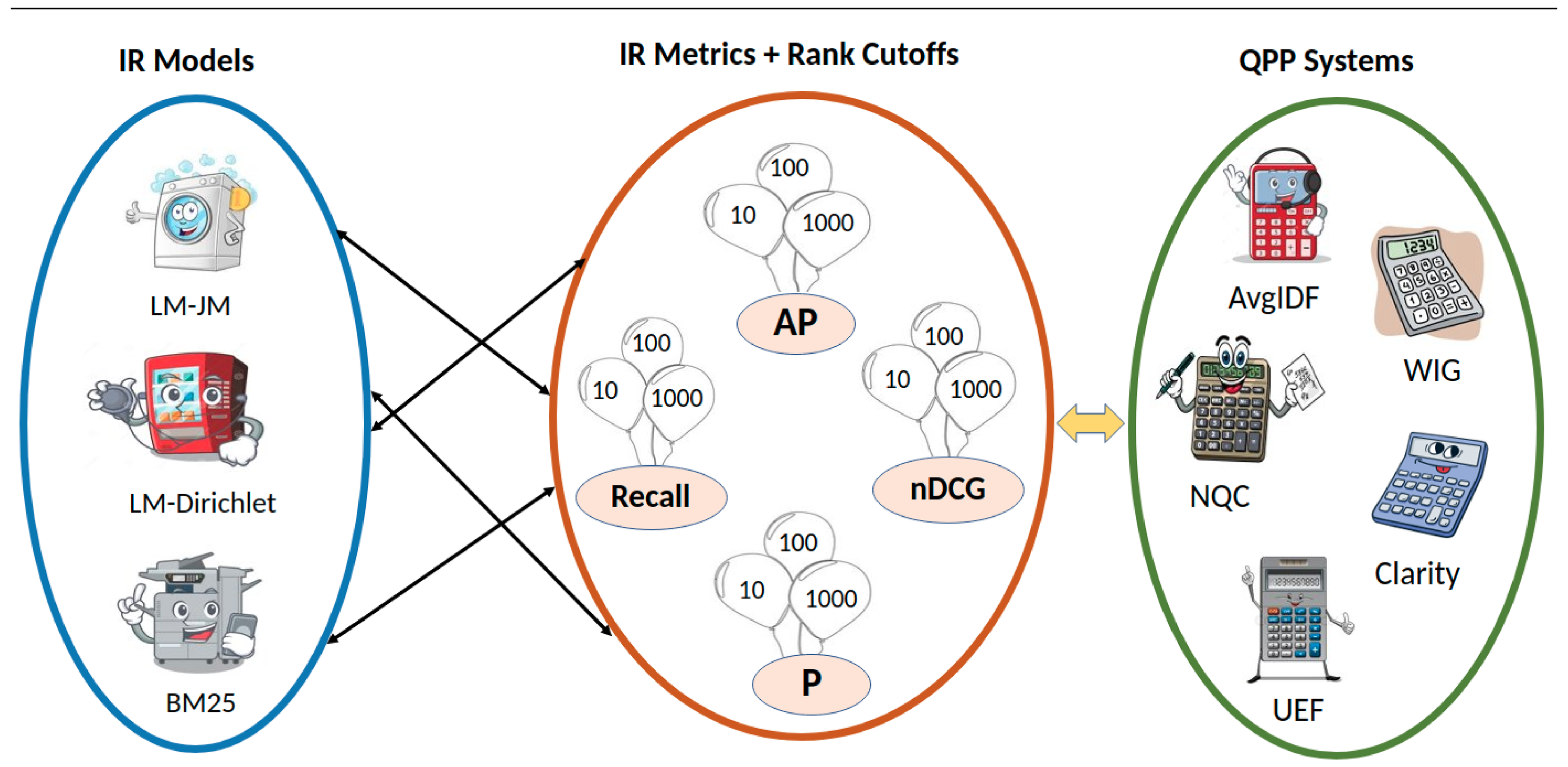
Our work : At a glance

- We analysed the relative stability of QPP outcomes (rank correlations) with respect to changes in the IR models or the IR evaluation metrics.
- We emphasize that variations in QPP results (both in terms of the absolute values themselves and also the relative effectiveness of different QPP systems) can lead to difficulties in reproducing QPP experiment results on standard datasets.

How do we evaluate QPP Estimators?



There are too many combinations!



Research Question - 1

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.

Research Question - 2

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

RQ-1: Variations due to IR Evaluation Metrics

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

		IR Evaluation Metric (θ)				
	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	

AvgIDF

		IR Evaluation Metric (θ)				
	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**	

WIG

RQ-1: Variations due to IR Models

- 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.

		IR Evaluation Metric (θ)				
	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

		IR Evaluation Metric (θ)				
	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

		IR Evaluation Metric (θ)				
	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

		IR Evaluation Metric (θ)				
	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	0.0308	**0.0202**	

(d) UEF(WIG)

RQ-2: Variations due to IR Evaluation Metrics

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

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

RQ-2: Variations due to IR Models

- 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.

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**

Concluding Remarks

The main takeaway of this work is that any future experiment on QPP should emphasize clear specification of the experimental setup to warrant better reproducibility.