Online Appendix to Evaluating Elements of Web-based Data Enrichment for Pseudo-Relevance Feedback Retrieval

Timo Breuer, Melanie Pest, and Philipp Schaer

TH Köln (University of Applied Sciences)
`firstname.lastname@th-koeln.de`

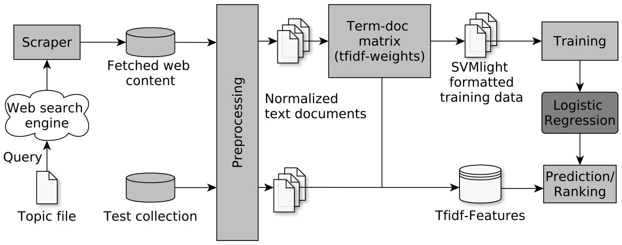

Fig. 1: Visualization of the workflow proposed by Grossman and Cormack [1].

Table 1: Results of baseline and advanced runs derived from Core18. This Table shows results of measures instantiated with AP and corresponds to Table 1 in the main paper.

	`uwmrgx` (baseline run)		`uwmrg` (advanced run)	
Run	AP	RMSE	AP	RMSE
GC [1]	0.2362	0	0.2761	0
`c18_g_td`	0.2472	0.1391	0.2784	0.0836
`c18_g_t`	0.2223	0.1325	0.2668	0.0871
`c18_d_td`	0.2824	0.1570	0.2672	0.0968
`c18_d_t`	0.2622	0.1288	0.2725	0.0975

Table 2: Results of baseline and advanced runs derived from Core18. This Table shows results of measures instantiated with P@10 and corresponds to Table 1 in the main paper.

	`uwmrgx` (baseline run)		`uwmrg` (advanced run)	
Run	P@10	RMSE	P@10	RMSE
GC [1]	0.4360	0	0.5000	0
`c18_g_td`	0.4280	0.2553	0.4660	0.1975
`c18_g_t`	0.3820	0.2553	0.4660	0.1703
`c18_d_td`	0.4780	0.3043	0.4400	0.2182
`c18_d_t`	0.4440	0.2078	0.4680	0.1844

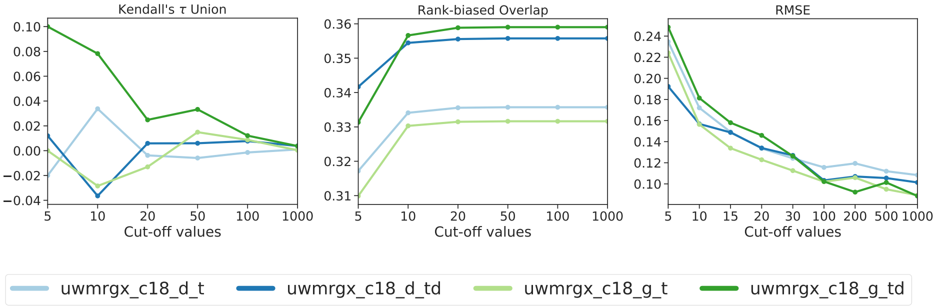

Fig. 2: Kendall's τ Union, Rank-biased Overlap, and the Root-Mean-Square-Error of the advanced run `uwmrg` averaged across the topics of Core18. This Figure complements Figure 2 in the main paper.

References

1. Grossman, M. R., and Cormack, G. V. MRG_UWaterloo Participation in the TREC 2018 Common Core Track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018* (2018), E. M. Voorhees and A. Ellis, Eds., vol. Special Publication 500-331, National Institute of Standards and Technology (NIST).

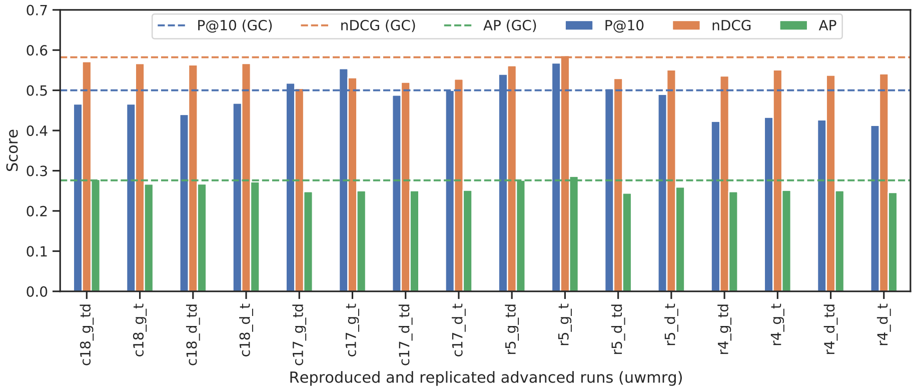

Fig. 3: Absolute scores of advanced runs. This Figure complements Figure 4 in the main paper.

Table 3: Overall effects of different run versions instantiated with P@10 & AP of the search engines (SE) and queries (Q). † and $*$ denote significant differences ($p < 0.05$) between SE and Q, respectively.

	P@10		Overall Effects		AP		Overall Effects	
Run	uwmrgx	uwmrg	DRI	ER	uwmrgx	uwmrg	DRI	ER
GC [1]	0.4360	0.5000	0	1	0.2362	0.2761	0	1
c18_g_td	0.4280	0.4660	0.0580	0.5938	$0.2472^{\dagger}$	0.2784	0.0427	0.7822
c18_g_t	$0.3820^{\dagger}$	0.4660	-0.0731	1.3125	$0.2223^{\dagger}$	0.2668	-0.0314	1.1164
c18_d_td	0.4780	0.4400	0.2263	-0.5938	$0.2824^{\dagger}$	0.2672	0.2230	-0.3820
c18_d_t	$0.4440^{\dagger}$	0.4680	0.0927	0.3750	$0.2622^{\dagger}$	0.2725	0.1296	0.2588
c17_g_td	0.4620	0.5180	0.0256	0.8750	0.2097	0.2481	-0.0140	0.9616
c17_g_t	$0.4100^{\dagger}$	0.5540	-0.2044	2.2500	$0.1863^{\dagger}$	0.2502	-0.1739	1.6011
c17_d_td	0.4360^{*}	0.4880	0.0275	0.8125	0.2262	0.2504	0.0620	0.6063
c17_d_t	$0.5200^{\dagger*}$	0.5000	0.1853	-0.3125	$0.2474^{\dagger}$	0.2515	0.1526	0.1018
r5_g_td	0.4520	0.5400	-0.0479	1.3750	0.2256	$0.2765^{\dagger}$	-0.0566	1.2753
r5_g_t	0.4620	$0.5680^{\dagger}$	-0.0826	1.6562	0.2122	$0.2861^{\dagger}$	-0.1796	1.8532
r5_d_td	0.4380	0.5040	-0.0039	1.0312	0.2304	$0.2443^{\dagger}$	0.1085	0.3489
r5_d_t	0.4620	$0.4900^{\dagger}$	0.0862	0.4375	0.2299	$0.2595^{\dagger}$	0.0404	0.7407
r4_g_td	0.4201^{*}	0.4229	0.1401	0.0439	0.2349^{*}	0.2480	0.1132	0.3283
r4_g_t	$0.3787^{\dagger*}$	0.4329	0.0036	0.8471	$0.2045^{\dagger*}$	0.2513	-0.0598	1.1727
r4_d_td	0.4269^{*}	0.4265	0.1477	-0.0063	0.2394	0.2504	0.1230	0.2760
r4_d_t	$0.4040^{\dagger*}$	0.4129	0.1249	0.1381	$0.2287^{\dagger}$	0.2459	0.0939	0.4306

Table 4: Average retrieval performance (ARP) of P@10, nDCG, AP and the corresponding p-values of (un-)paired t-tests between the reference runs and our reimplementations. A paired t-test is conducted if the run is derived from the same test collection (Core18) as the reference run GC [1], and an unpaired t-test is conducted if the run is derived from another test collection (Core17, Robust04, Robust05). Significant differences ($p < 0.05$) are marked in bold.

		P@10		nDCG		AP	
	Run	ARP	p	ARP	p	ARP	p
uwmrgx	GC [1]	0.4360	1.0000	0.5306	1.0000	0.2362	1.0000
	c18_d_t	0.4440	0.7886	0.5458	0.4440	0.2622	0.1550
	c18_d_td	0.4780	0.3341	0.5735	0.0700	**0.2824**	**0.0358**
	c18_g_t	0.3820	0.0871	0.5024	0.2444	0.2223	0.4653
	c18_g_td	0.4280	0.8272	0.5325	0.9270	0.2472	0.5800
	c17_d_t	0.5200	0.2166	0.5223	0.8494	0.2474	0.7635
	c17_d_td	0.4360	1.0000	0.4870	0.3251	0.2262	0.7930
	c17_g_t	0.4100	0.6940	**0.4404**	**0.0399**	0.1863	0.1659
	c17_g_td	0.4620	0.6945	0.4836	0.2622	0.2097	0.4599
	r5_d_t	0.4620	0.6872	0.5175	0.7495	0.2299	0.8642
	r5_d_td	0.4380	0.9754	0.5134	0.6902	0.2304	0.8735
	r5_g_t	0.4620	0.6860	0.5003	0.4637	0.2122	0.4958
	r5_g_td	0.4520	0.8104	0.5088	0.6109	0.2256	0.7686
	r4_d_t	0.4040	0.4765	0.5171	0.6812	0.2287	0.8035
	r4_d_td	0.4269	0.8457	0.5317	0.9735	0.2394	0.9132
	r4_g_t	0.3787	0.2132	0.4886	0.1861	0.2045	0.2682
	r4_g_td	0.4201	0.7233	0.5266	0.8990	0.2349	0.9670
uwmrg	GC [1]	0.5000	1.0000	0.5822	1.0000	0.2761	1.0000
	c18_d_t	0.4680	0.2233	0.5668	0.3205	0.2725	0.7994
	c18_d_td	0.4400	0.0508	0.5633	0.1915	0.2672	0.5222
	c18_g_t	0.4660	0.1601	0.5666	0.2209	0.2668	0.4605
	c18_g_td	0.4660	0.2270	0.5713	0.3926	0.2784	0.8450
	c17_d_t	0.5000	1.0000	0.5279	0.1852	0.2515	0.5363
	c17_d_td	0.4880	0.8630	0.5201	0.1543	0.2504	0.5259
	c17_g_t	0.5540	0.4357	0.5313	0.2007	0.2502	0.5103
	c17_g_td	0.5180	0.7988	0.5047	0.0949	0.2481	0.4942
	r5_d_t	0.4900	0.8760	0.5509	0.4243	0.2595	0.6657
	r5_d_td	0.5040	0.9507	0.5295	0.2040	0.2443	0.4132
	r5_g_t	0.5680	0.3031	0.5865	0.9097	0.2861	0.7956
	r5_g_td	0.5400	0.5475	0.5613	0.6086	0.2765	0.9915
	r4_d_t	0.4129	0.0596	0.5411	0.2101	0.2459	0.3270
	r4_d_td	0.4265	0.1274	0.5376	0.1943	0.2504	0.4176
	r4_g_t	0.4329	0.1509	0.5509	0.3154	0.2513	0.4199
	r4_g_td	0.4229	0.1031	0.5357	0.1690	0.2480	0.3725