Bibliometric Data Fusion for Biomedical Information Retrieval

ACM/IEEE Joint Conference on Digital Libraries
2023, Santa Fe, New Mexico, USA

Timo Breuer[1] **Christin Katharina Kreutz**[1] **Philipp Schaer**[1] **Dirk Tunger**[2]

[1]TH Köln - University of Applied Sciences, Germany

[2]Forschungszentrum Jülich, Germany

June 27, 2023

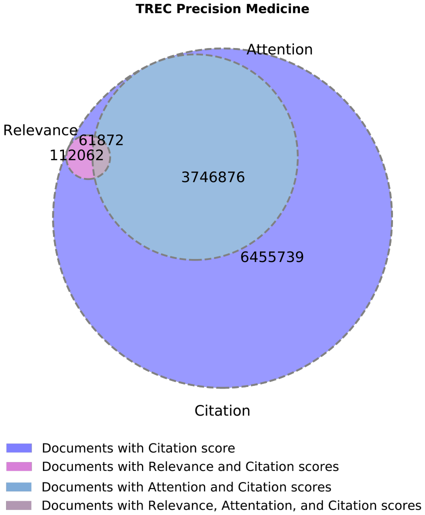
Motivation

Bibliometric measures are implicit relevance signals

Correlation between bibliometrics and relevance labels of IR test collections [1]

How to exploit these relevance signals for document retrieval?

[1] *Relevance assessments, bibliometrics, and altmetrics: a quantitative study on PubMed and arXiv*, Breuer, Schaer, and Tunger, Scientometrics 2022



Methodology

Retrieve a baseline ranking

Fuse the ranking list with additional bibliometric signals

Evaluate the re-ranked result list

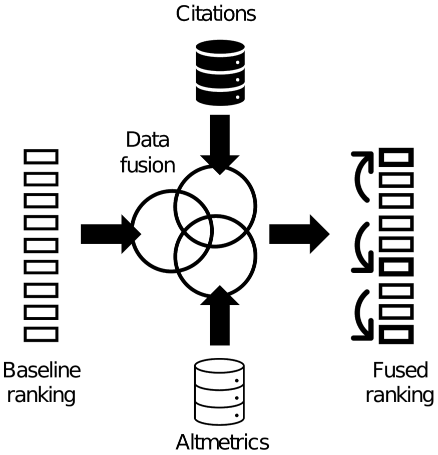


Figure: Bibliometric data fusion based on polyrepresentation.

Research Questions

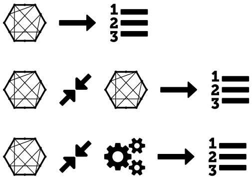
RQ1 *To what extent can bibliometric relevance signals be used as ranking criteria for biomedical information retrieval?*

RQ2 *Can bibliometric-enhanced data fusion methods improve the overall retrieval performance?*

Research Questions

RQ1 *To what extent can bibliometric relevance signals be used as ranking criteria for biomedical information retrieval?*

RQ2 *Can bibliometric-enhanced data fusion methods improve the overall retrieval performance?*



i) Rankings based on bibliometric measures

ii) Rankings based on fused bibliometrics

iii) Bibliometric data fusion with retrieval systems

Polyrepresentation

"Cognitively and functionally different representations of information objects may be used in information retrieval to enhance quality of results." [2]

➜ Enhance biomedical retrieval systems with bibliometric metadata like citations, altmetrics, etc.

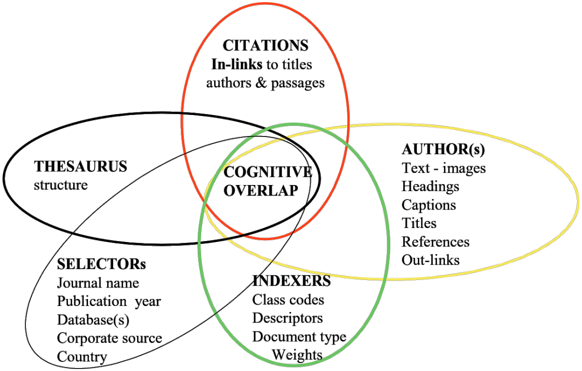


Figure: Principle of polyrepresentation (reproduced from [2]).

Data Fusion

Combine multiple rankings for better retrieval effectiveness than the best single ranking.

Reciprocal Rank Fusion [3]:

$$RRF\,\text{score}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}$$

D is the document set,
R is the set of fused rankings,
$r(d)$ is the rank r of document d,
k is a fixed parameter set to 60.

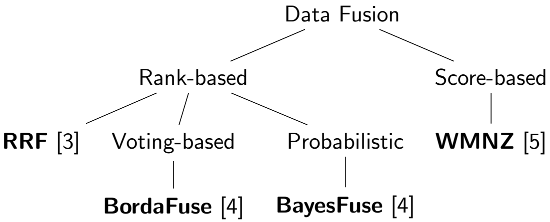


Figure: Overview of the analyzed data fusion methods.

TREC Precision Medicine Abstract Task 2017 to 2019

TREC organized several **biomedical shared task**, e.g., Precision Medicine

Two tasks: Ranking of 1) **medical abstracts** and 2) clinical trials

Information needs / topics based on **patient profiles**

Table: Number of relevance judgements, of teams who submitted and of submitted runs per year of TREC-PM.

Year	Topics	Qrels	Teams	Runs
2017	30	22,642	29	125
2018	50	22,429	24	103
2019	40	18,316	14	62

Retrieval Engines and Approaches of TREC-PM 2017 to 2019

- Most of the rankings are made with a Lucene-based retrieval engine
- Data fusion is a common technique

! Few systems use bibliometric metadata to rank scientific abstracts

Table: Overview of TREC-PM 2017 to 2019.

		2017	2018	2019	$\sum$
	Reports per year	20	20	14	54
Engine	ElasticSearch	5	8	7	20
	Lucene	6	3	2	11
	Terrier	3	3	1	7
	unknown	1	2	2	5
	Solr	2	2	1	5
	Galago	2			2
	Indri	1	1		2
	Whoosh		1	1	2
Approaches	Query expansion	16	14	12	42
	KB + ontologies	17	14	6	37
	Re-ranking	6	7	9	22
	Embeddings	3	5	5	13
	Data fusion	4	5	3	12
	LTR	1	3	5	9
	LLM			3	3
	Citation-based	2			2

Coverage of Bibliometric Metadata

Dataset covers

- **C**itations,
- **A**ltmetrics,
- **P**ublication years,
- **R**esearch levels,
- **I**mpact factors.

Public resource hosted on Zenodo:
`https://doi.org/10.5281/zenodo.5883400`

Table: Coverage of the bibliometrics wrt. judged abstracts

Year	**2017**	**2018**	**2019**
C	14170 (66%)	11214 (55%)	11381 (61%)
A	6134 (29%)	4547 (22%)	5639 (30%)
P	14586 (68%)	11618 (57%)	12221 (66%)
R	14067 (66%)	11239 (55%)	11707 (63%)
I	11449 (53%)	9246 (45%)	9387 (51%)

→ Rankings Based on Bibliometric Measures

→ Rankings Based on Fused Bibliometrics

→ Retrieval Effectiveness of Bibliometric Relevance Signals

- High recall rates comply with our earlier work [1]

- **C**itations, **A**ltmetrics, and **P**ublication years are the most effective bibliometric relevance signals

- ! BM25 outperforms query-agnostic bibliometric rankings

Table: Retrieval effectiveness of bibliometric relevance signals. Superscripts denote significant differences.

	Model	**C**	**A**	**P**	**R**	**I**	**BM25**
2017	Recall	0.7853^{ARI}	0.4162	**0.7972**^{CARI}	0.7608^{AI}	0.6301^{A}	0.4640
	nDCG	0.4992^{ARI}	0.3163	**0.5069**^{ARI}	0.4666^{AI}	0.4162^{A}	0.4423
	AP	**0.1812**^{AI}	0.1020	0.1733^{AI}	0.1546^{A}	0.1399^{A}	0.1636
	P@10	**0.2700**^{R}	0.2400^{R}	0.2033	0.1200	0.2500^{R}	0.4667
	Bpref	**0.1577**	0.1434	0.1541	0.1307	0.1444	0.2714
2018	Recall	0.7916^{ARI}	0.4066	**0.8019**^{CARI}	0.7739^{AI}	0.6438^{A}	0.7828
	nDCG	**0.5728**^{ARI}	0.3651	0.5671^{ARI}	0.5297^{AI}	0.4744^{A}	0.6376
	AP	**0.2905**^{ARI}	0.1765	0.2815^{AI}	0.2591^{AI}	0.2261^{A}	0.3195
	P@10	0.3760^{R}	**0.3860**^{R}	0.3180^{R}	0.2360	0.3420^{R}	0.5680
	Bpref	**0.2896**^{AI}	0.2355	0.2809^{A}	0.2612	0.2506	0.4852
2019	Recall	0.8260^{AI}	0.4732	**0.8849**^{CARI}	0.8435^{AI}	0.6690^{A}	0.7574
	nDCG	0.5754^{ARI}	0.3693	**0.6031**^{ARI}	0.5433^{AI}	0.4818^{A}	0.5870
	AP	0.2756^{ARI}	0.1633	**0.2896**^{ARI}	0.2442^{A}	0.2182^{A}	0.2584
	P@10	**0.3525**^{RI}	0.2850^{R}	0.3075^{R}	0.1925	0.2850^{R}	0.5125
	Bpref	**0.2460**^{R}	0.2064	0.2416	0.2024	0.2283	0.3946

Retrieval Effectiveness of Bibliometric Relevance Signals

- High recall rates comply with our earlier work [1]

- **C**itations, **A**ltmetrics, and **P**ublication years are the most effective bibliometric relevance signals

! BM25 outperforms query-agnostic bibliometric rankings

Table: Retrieval effectiveness of bibliometric relevance signals. Superscripts denote significant differences.

	Model	C	A	P	R	I	BM25
2017	Recall	0.7853^{ARI}	0.4162	$\mathbf{0.7972}^{CARI}$	0.7608^{AI}	0.6301^{A}	0.4640
	nDCG	0.4992^{ARI}	0.3163	$\mathbf{0.5069}^{ARI}$	0.4666^{AI}	0.4162^{A}	0.4423
	AP	$\mathbf{0.1812}^{AI}$	0.1020	0.1733^{AI}	0.1546^{A}	0.1399^{A}	0.1636
	P@10	$\mathbf{0.2700}^{R}$	0.2400^{R}	0.2033	0.1200	0.2500^{R}	0.4667
	Bpref	**0.1577**	0.1434	0.1541	0.1307	0.1444	0.2714
2018	Recall	0.7916^{ARI}	0.4066	$\mathbf{0.8019}^{CARI}$	0.7739^{AI}	0.6438^{A}	0.7828
	nDCG	$\mathbf{0.5728}^{ARI}$	0.3651	0.5671^{ARI}	0.5297^{AI}	0.4744^{A}	0.6376
	AP	$\mathbf{0.2905}^{ARI}$	0.1765	0.2815^{AI}	0.2591^{AI}	0.2261^{A}	0.3195
	P@10	0.3760^{R}	$\mathbf{0.3860}^{R}$	0.3180^{R}	0.2360	0.3420^{R}	0.5680
	Bpref	$\mathbf{0.2896}^{AI}$	0.2355	0.2809^{A}	0.2612	0.2506	0.4852
2019	Recall	0.8260^{AI}	0.4732	$\mathbf{0.8849}^{CARI}$	0.8435^{AI}	0.6690^{A}	0.7574
	nDCG	0.5754^{ARI}	0.3693	$\mathbf{0.6031}^{ARI}$	0.5433^{AI}	0.4818^{A}	0.5870
	AP	0.2756^{ARI}	0.1633	$\mathbf{0.2896}^{ARI}$	0.2442^{A}	0.2182^{A}	0.2584
	P@10	$\mathbf{0.3525}^{RI}$	0.2850^{R}	0.3075^{R}	0.1925	0.2850^{R}	0.5125
	Bpref	$\mathbf{0.2460}^{R}$	0.2064	0.2416	0.2024	0.2283	0.3946

→ Retrieval Effectiveness of Bibliometric Relevance Signals

- High recall rates comply with our earlier work [1]
- **C**itations, **A**ltmetrics, and **P**ublication years are the most effective bibliometric relevance signals

! BM25 outperforms query-agnostic bibliometric rankings

Table: Retrieval effectiveness of bibliometric relevance signals. Superscripts denote significant differences.

	Model	C	A	P	R	I	BM25
2017	Recall	0.7853^{ARI}	0.4162	**0.7972**^{CARI}	0.7608^{AI}	0.6301^{A}	0.4640
	nDCG	0.4992^{ARI}	0.3163	**0.5069**^{ARI}	0.4666^{AI}	0.4162^{A}	0.4423
	AP	**0.1812**^{AI}	0.1020	0.1733^{AI}	0.1546^{A}	0.1399^{A}	0.1636
	P@10	**0.2700**^{R}	0.2400^{R}	0.2033	0.1200	0.2500^{R}	0.4667
	Bpref	**0.1577**	0.1434	0.1541	0.1307	0.1444	0.2714
2018	Recall	0.7916^{ARI}	0.4066	**0.8019**^{CARI}	0.7739^{AI}	0.6438^{A}	0.7828
	nDCG	**0.5728**^{ARI}	0.3651	0.5671^{ARI}	0.5297^{AI}	0.4744^{A}	0.6376
	AP	**0.2905**^{ARI}	0.1765	0.2815^{AI}	0.2591^{AI}	0.2261^{A}	0.3195
	P@10	0.3760^{R}	**0.3860**^{R}	0.3180^{R}	0.2360	0.3420^{R}	0.5680
	Bpref	**0.2896**^{AI}	0.2355	0.2809^{A}	0.2612	0.2506	0.4852
2019	Recall	0.8260^{AI}	0.4732	**0.8849**^{CARI}	0.8435^{AI}	0.6690^{A}	0.7574
	nDCG	0.5754^{ARI}	0.3693	**0.6031**^{ARI}	0.5433^{AI}	0.4818^{A}	0.5870
	AP	0.2756^{ARI}	0.1633	**0.2896**^{ARI}	0.2442^{A}	0.2182^{A}	0.2584
	P@10	**0.3525**^{RI}	0.2850^{R}	0.3075^{R}	0.1925	0.2850^{R}	0.5125
	Bpref	**0.2460**^{R}	0.2064	0.2416	0.2024	0.2283	0.3946

Retrieval Effectiveness of Bibliometric Relevance Signals

- High recall rates comply with our earlier work [1]
- **C**itations, **A**ltmetrics, and **P**ublication years are the most effective bibliometric relevance signals

! BM25 outperforms query-agnostic bibliometric rankings

Table: Retrieval effectiveness of bibliometric relevance signals. Superscripts denote significant differences.

	Model	**C**	**A**	**P**	**R**	**I**	**BM25**
2017	Recall	0.7853^{ARI}	0.4162	**0.7972**^{CARI}	0.7608^{AI}	0.6301^{A}	0.4640
	nDCG	0.4992^{ARI}	0.3163	**0.5069**^{ARI}	0.4666^{AI}	0.4162^{A}	0.4423
	AP	**0.1812**^{AI}	0.1020	0.1733^{AI}	0.1546^{A}	0.1399^{A}	0.1636
	P@10	**0.2700**^{R}	0.2400^{R}	0.2033	0.1200	0.2500^{R}	0.4667
	Bpref	**0.1577**	0.1434	0.1541	0.1307	0.1444	0.2714
2018	Recall	0.7916^{ARI}	0.4066	**0.8019**^{CARI}	0.7739^{AI}	0.6438^{A}	0.7828
	nDCG	**0.5728**^{ARI}	0.3651	0.5671^{ARI}	0.5297^{AI}	0.4744^{A}	0.6376
	AP	**0.2905**^{ARI}	0.1765	0.2815^{AI}	0.2591^{AI}	0.2261^{A}	0.3195
	P@10	0.3760^{R}	**0.3860**^{R}	0.3180^{R}	0.2360	0.3420^{R}	0.5680
	Bpref	**0.2896**^{AI}	0.2355	0.2809^{A}	0.2612	0.2506	0.4852
2019	Recall	0.8260^{AI}	0.4732	**0.8849**^{CARI}	0.8435^{AI}	0.6690^{A}	0.7574
	nDCG	0.5754^{ARI}	0.3693	**0.6031**^{ARI}	0.5433^{AI}	0.4818^{A}	0.5870
	AP	0.2756^{ARI}	0.1633	**0.2896**^{ARI}	0.2442^{A}	0.2182^{A}	0.2584
	P@10	**0.3525**^{RI}	0.2850^{R}	0.3075^{R}	0.1925	0.2850^{R}	0.5125
	Bpref	**0.2460**^{R}	0.2064	0.2416	0.2024	0.2283	0.3946

Fusion of Bibliometric Relevance Signals

2017

	Recall	nDCG	AP	P@10	Bpref
CA	0.8088	0.5013	0.1661	0.1867	0.1364
CP	0.7972	0.4819	0.1492	0.1400	0.1197
CR	0.7967	0.4823	0.1492	0.1567	0.1206
CI	0.7931	0.4815	0.1492	0.1600	0.1218
AP	0.8157	0.5067	0.1673	0.1900	0.1361
AR	0.7989	0.4997	0.1648	0.2100	0.1359
AI	0.7328	0.4653	0.1513	0.1933	0.1352
PR	0.7972	0.4829	0.1484	0.1500	0.1205
PI	0.7972	0.4844	0.1495	0.1567	0.1218
RI	0.7616	0.4685	0.1442	0.1567	0.1218
CAP	0.8157	0.4999	0.1625	0.1867	0.1298
CAR	0.8155	0.5006	0.1628	0.2033	0.1300
CAI	0.8133	0.4993	0.1624	0.1833	0.1312
CPR	0.7972	0.4825	0.1491	0.1567	0.1203
CPI	0.7972	0.4833	0.1497	0.1600	0.1212
CRI	0.7967	0.4830	0.1496	0.1600	0.1218
APR	0.8157	0.5025	0.1626	0.2100	0.1292
API	0.8157	0.5008	0.1625	0.1933	0.1305
ARI	0.7995	0.4934	0.1597	0.1933	0.1309
PRI	0.7972	0.4841	0.1493	0.1567	0.1217
CAPR	0.8157	0.4971	0.1595	0.1833	0.1264
CAPI	0.8157	0.4968	0.1598	0.1733	0.1280
CARI	0.8155	0.4968	0.1598	0.1800	0.1284
CPRI	0.7972	0.4830	0.1493	0.1600	0.1215
APRI	0.8157	0.4976	0.1594	0.1767	0.1278
CAPRI	0.8157	0.4954	0.1580	0.1733	0.1263

2018

	Recall	nDCG	AP	P@10	Bpref
CA	0.8133	0.5728	0.2767	0.3240	0.2692
CP	0.8019	0.5402	0.2363	0.2640	0.2191
CR	0.8012	0.5401	0.2359	0.2700	0.2186
CI	0.7972	0.5381	0.2369	0.2660	0.2238
AP	0.8194	0.5748	0.2749	0.3220	0.2653
AR	0.8108	0.5703	0.2722	0.3280	0.2640
AI	0.7426	0.5346	0.2541	0.3300	0.2618
PR	0.8019	0.5370	0.2309	0.2620	0.2121
PI	0.8019	0.5382	0.2337	0.2660	0.2182
RI	0.7744	0.5233	0.2260	0.2660	0.2165
CAP	0.8194	0.5695	0.2664	0.3220	0.2507
CAR	0.8194	0.5697	0.2663	0.3280	0.2498
CAI	0.8166	0.5680	0.2669	0.3240	0.2541
CPR	0.8019	0.5398	0.2346	0.2700	0.2168
CPI	0.8019	0.5398	0.2360	0.2660	0.2210
CRI	0.8012	0.5394	0.2357	0.2660	0.2210
APR	0.8194	0.5677	0.2625	0.3240	0.2449
API	0.8194	0.5680	0.2641	0.3220	0.2502
ARI	0.8108	0.5631	0.2612	0.3220	0.2492
PRI	0.8019	0.5375	0.2324	0.2660	0.2170
CAPR	0.8194	0.5656	0.2592	0.3260	0.2400
CAPI	0.8194	0.5651	0.2601	0.3060	0.2447
CARI	0.8194	0.5651	0.2599	0.3080	0.2446
CPRI	0.8019	0.5391	0.2344	0.2660	0.2187
APRI	0.8194	0.5634	0.2570	0.3020	0.2409
CAPRI	0.8194	0.5619	0.2549	0.3000	0.2377

2019

	Recall	nDCG	AP	P@10	Bpref
CA	0.8882	0.5939	0.2766	0.3225	0.2315
CP	0.8849	0.5780	0.2619	0.2525	0.2151
CR	0.8806	0.5756	0.2601	0.2575	0.2130
CI	0.8620	0.5607	0.2489	0.2225	0.2052
AP	0.9148	0.6104	0.2856	0.3175	0.2333
AR	0.9004	0.6023	0.2803	0.3175	0.2322
AI	0.8141	0.5523	0.2481	0.3075	0.2213
PR	0.8849	0.5784	0.2613	0.2600	0.2139
PI	0.8849	0.5731	0.2559	0.2300	0.2065
RI	0.8438	0.5526	0.2437	0.2300	0.2041
CAP	0.9148	0.6043	0.2797	0.3200	0.2256
CAR	0.9124	0.6031	0.2787	0.3175	0.2245
CAI	0.9026	0.5944	0.2712	0.2950	0.2189
CPR	0.8849	0.5778	0.2615	0.2575	0.2138
CPI	0.8849	0.5740	0.2578	0.2225	0.2081
CRI	0.8807	0.5719	0.2564	0.2225	0.2070
APR	0.9148	0.6053	0.2793	0.3150	0.2248
API	0.9148	0.6025	0.2756	0.3075	0.2196
ARI	0.9004	0.5945	0.2705	0.3050	0.2184
PRI	0.8849	0.5737	0.2572	0.2300	0.2079
CAPR	0.9148	0.6016	0.2765	0.2925	0.2218
CAPI	0.9148	0.5988	0.2735	0.2825	0.2176
CARI	0.9124	0.5975	0.2726	0.2825	0.2170
CPRI	0.8849	0.5744	0.2582	0.2225	0.2086
APRI	0.9148	0.6000	0.2736	0.2900	0.2175
CAPRI	0.9148	0.5972	0.2723	0.2625	0.2162

Figure: Retrieval effectiveness of fused rankings based on bibliometric relevance signals.

Bibliometric Data Fusion with Retrieval Systems

Improvements of TREC-PM 2017 Abstract Task

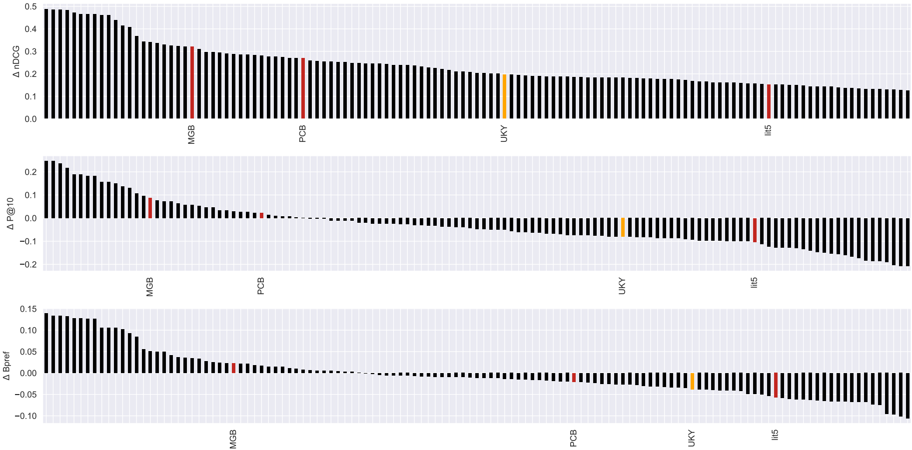


Figure: Rank fusion-based improvements over the baseline runs for the TREC-PM Abstract task for 2017.

Improvements of TREC-PM 2018 and 2019

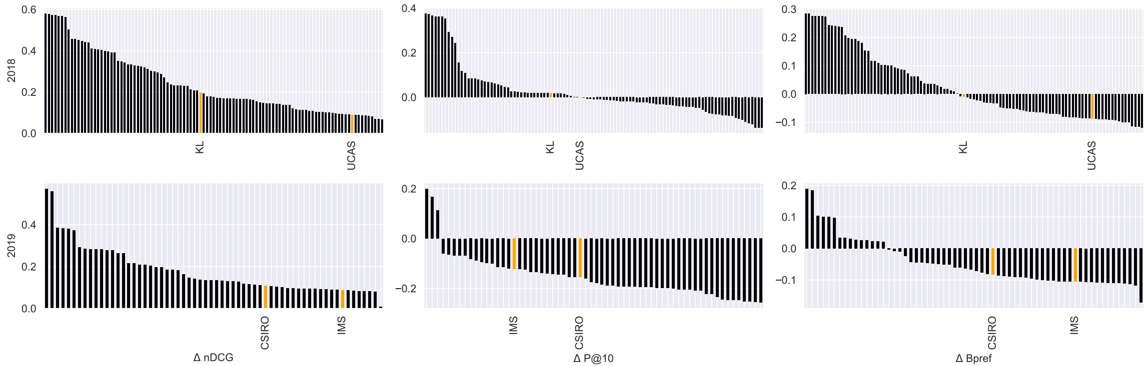


Figure: Rank fusion-based improvements over the baseline runs for the TREC-PM Abstract task for 2018 and 2019.

What Kind of Systems Improve?

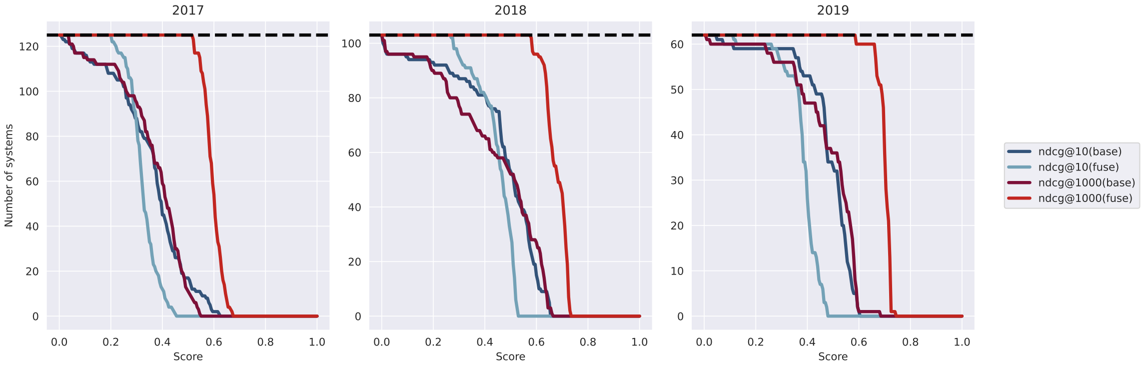


Figure: Number of systems vs. retrieval effectiveness before (dark) and after (light) bibliometric data fusion for nDCG@10 (blue) and nDCG@1000 (red) for TREC-PM. The dashed line corresponds to the total number of systems.

Do the Results Generalize?

- Almost all retrieval systems significantly improve in terms of nDCG and AP.
- Tradeoffs between recall-based improvements and lowered precision.
- ! Results generalize with all data fusion algorithms and TREC-PM datasets.

Table: Bibliometric Data Fusion based on RRF.

Year	2017	2018	2019
Number of systems	125	103	62
(Signif.*) improvements (nDCG)	125 / 125*	103 / 103*	62 / 61*
Average improvement (nDCG)	0.2378	0.2384	0.1815
Overall change (nDCG)	0.2378	0.2384	0.1787
(Signif.*) improvements (AP)	125 / 123*	103 / 103*	62 / 55*
Average improvement (AP)	0.1173	0.1849	0.1237
Overall change (AP)	0.1163	0.1849	0.1161
(Signif.*) improvements (P@10)	37 / 18*	46 / 19*	3 / 3*
Average improvement (P@10)	0.1589	0.2221	0.16
Overall change (P@10)	-0.0299	0.0223	-0.1518
(Signif.*) improvements (Bpref)	46 / 17*	47 / 36*	15 / 6*
Average improvement (Bpref)	0.1047	0.1668	0.1294
Overall change (Bpref)	-0.0033	0.0244	-0.0453

Do the Results Generalize?

- Almost all retrieval systems significantly improve in terms of nDCG and AP.
- Tradeoffs between recall-based improvements and lowered precision.

! Results generalize with all data fusion algorithms and TREC-PM datasets.

Table: Bibliometric Data Fusion based on RRF.

Year	2017	2018	2019
Number of systems	125	103	62
(Signif.*) improvements (nDCG)	125 / 125*	103 / 103*	62 / 61*
Average improvement (nDCG)	0.2378	0.2384	0.1815
Overall change (nDCG)	0.2378	0.2384	0.1787
(Signif.*) improvements (AP)	125 / 123*	103 / 103*	62 / 55*
Average improvement (AP)	0.1173	0.1849	0.1237
Overall change (AP)	0.1163	0.1849	0.1161
(Signif.*) improvements (P@10)	37 / 18*	46 / 19*	3 / 3*
Average improvement (P@10)	0.1589	0.2221	0.16
Overall change (P@10)	-0.0299	0.0223	-0.1518
(Signif.*) improvements (Bpref)	46 / 17*	47 / 36*	15 / 6*
Average improvement (Bpref)	0.1047	0.1668	0.1294
Overall change (Bpref)	-0.0033	0.0244	-0.0453

Do the Results Generalize?

- Almost all retrieval systems significantly improve in terms of nDCG and AP.
- Tradeoffs between recall-based improvements and lowered precision.

! Results generalize with all data fusion algorithms and TREC-PM datasets.

Table: Bibliometric Data Fusion based on RRF.

Year	2017	2018	2019
Number of systems	125	103	62
(Signif.*) improvements (nDCG)	125 / 125*	103 / 103*	62 / 61*
Average improvement (nDCG)	0.2378	0.2384	0.1815
Overall change (nDCG)	0.2378	0.2384	0.1787
(Signif.*) improvements (AP)	125 / 123*	103 / 103*	62 / 55*
Average improvement (AP)	0.1173	0.1849	0.1237
Overall change (AP)	0.1163	0.1849	0.1161
(Signif.*) improvements (P@10)	37 / 18*	46 / 19*	3 / 3*
Average improvement (P@10)	0.1589	0.2221	0.16
Overall change (P@10)	-0.0299	0.0223	-0.1518
(Signif.*) improvements (Bpref)	46 / 17*	47 / 36*	15 / 6*
Average improvement (Bpref)	0.1047	0.1668	0.1294
Overall change (Bpref)	-0.0033	0.0244	-0.0453

Implications for (Patient) Users

Rank-biased Precision [6]:

$$\mathrm{RBP} = (1 - p) \cdot \sum_{i=1}^{d} r_i \cdot p^{i-1}$$

r_i denotes relevance at rank i,
p is the transition probability to the next document and models the user's patience.
The higher p, the more patient the user.

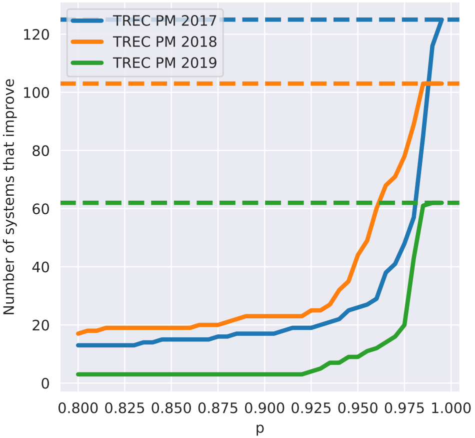


Figure: Number of system improvements vs. user persistence.

Answers to the Research Questions

RQ1: To what extent can bibliometric relevance signals be used as ranking criteria for biomedical information retrieval?

- Bibliometric relevance signals can indicate relevant literature to some extent.
- Bibliometric rankings are not as effective as term-based retrieval methods.
- Fusing bibliometric relevance signals is less effective than using them in isolation.

Answers to the Research Questions

RQ1: To what extent can bibliometric relevance signals be used as ranking criteria for biomedical information retrieval?

- Bibliometric relevance signals can indicate relevant literature to some extent.
- Bibliometric rankings are not as effective as term-based retrieval methods.
- Fusing bibliometric relevance signals is less effective than using them in isolation.

RQ2: Can bibliometric-enhanced data fusion methods improve the overall retrieval performance?

- For all systems of TREC-PM 2017 to 2019, the nDCG and AP scores can be improved.
- Not only weak baselines but also well-performing systems benefit from data fusion.
- The more patient the user, the higher the benefit.

Thank You!

Thank you for your attention.

Questions?




https://github.com/irgroup/jcdl2023-data-fusion

https://ir.web.th-koeln.de

References I

[1] T. Breuer, P. Schaer, and D. Tunger, "Relevance assessments, bibliometrics, and altmetrics: A quantitative study on pubmed and arxiv," *Scientometrics*, vol. 127, no. 5, pp. 2455–2478, 2022.

[2] B. Larsen, P. Ingwersen, and J. Kekäläinen, "The polyrepresentation continuum in IR," in *IIiX*, ACM, 2006, pp. 88–96.

[3] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *SIGIR*, ACM, 2009, pp. 758–759.

[4] J. A. Aslam and M. H. Montague, "Models for metasearch," in *SIGIR*, ACM, 2001, pp. 275–284.

[5] S. Wu and F. Crestani, "Data fusion with estimated weights," in *CIKM*, ACM, 2002, pp. 648–651.

[6] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Trans. Inf. Syst.*, vol. 27, no. 1, 2:1–2:27, 2008.