Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR)

Jaime Arguello[1], Matt Crane[2], Fernando Diaz[3],
Jimmy Lin[4], and Andrew Trotman[5]

[1] University of North Carolina [2] University of Otago [3] Microsoft Research
[4] University of Waterloo [5] eBay Inc.

Abstract

The SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) took place on Thursday, August 13, 2015 in Santiago, Chile. The goal of the workshop was two fold. The first to provide a venue for the publication and presentation of negative results. The second was to provide a venue through which the authors of open source search engines could compare performance of indexing and searching on the same collections and on the same machines - encouraging the sharing of ideas and discoveries in a like-to-like environment. In total three papers were presented and seven systems participated.

1 Introduction

Many, if not most, published research papers in Information Retrieval (IR) describe the following process: the authors identify an opportunity to improve on a particular IR task, implement an experimental system, and compare its performance against one or more baselines (or a control condition, in the case of a user study). The quality of the research is judged based on the magnitude of the improvement and whether the methodological choices suggest external validity and generalizability, for example, whether the experimental setup is "realistic" or whether the baseline methods reflect the state of the art.

Unfortunately, research demonstrating the *failure* to reproduce or generalize previous results does not have a similar publication venue. This sort of result—often referred to as a 'negative result'—serves to control the quality of published research in a scientific discipline and to better understand the limits of previously published methods. Publication venues for such research exist in fields such as ecology,[1] biomedicine,[2] pharmacy,[3], and social science.[4]

The SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) provided a venue for publication and discussion of IR research that failed

[1]`http://jnr-eeb.org/index.php/jnr`
[2]`http://www.jnrbm.com/`
[3]`http://www.pnrjournal.com/`
[4]`http://jspurc.org/intro2.htm`

to reproduce a previously published result under the same or similar experimental conditions (e.g., same test collection and system configuration) and research that demonstrated the failure to generalize an existing approach to a new domain. To this end, we developed a set of categories covering different ways in which a result may fail to reproduce or generalize, and circulated a call for papers in these categories.

2 Scope

We expected papers in this workshop to focus on different scenarios in which previous results might fail to reproduce. Specifically, we invited submissions from the following four categories: repeatability of published experiments, reproducibility of published experiments on comparable data, generalizability of published results to comparable tasks, and inexplicability of unpublished experiments. We provide more details about these categories below.

2.1 Repeatability

Although IR experiments vary in subtle ways that may influence the precise values of, for example, evaluation numbers, we expect hypothesis tests to be robust to these subtle variations. A submission in this category was to demonstrate a failure to repeat a published result under approximately the same conditions in which the previously published experiments occurred. Hypothetical examples included papers making claims such as:

(a) "published mean average precision (MAP) improvements on TREC8 for BM25 with Rocchio pseudo-relevance feedback are not reproducible."

Papers in this area serve to control the quality of results in IR research.

2.2 Reproducibility

IR experiments are often conducted on specific corpora, sets of queries, and relevance judgements. In many cases, these experiments can be conducted on other comparable corpora, queries, or relevance judgements. A submission in this category was expected to fail to reproduce a published result on a comparable dataset. Hypothetical examples include papers making claims such as:

(a) "published MAP improvements on TREC8 for BM25 with Rocchio pseudo-relevance feedback are not reproducible on Reuters, a comparable news corpus and queries."

(b) "published production interleaving improvements on Bingle, a portal web search engine, for ranking with LTRx are not reproducible on Yandu, a comparable production environment."

Papers in this area serve to test the sensitivity of results in IR research to experimental conditions.

2.3 Generalizability

Many IR strategies have demonstrated effectiveness across different comparable task definitions (e.g., 'BM25 is an effective term weighting scheme for different text ranking tasks'). A submission in this category was expect to fail to reproduce a published result on a comparable task. Hypothetical examples include papers making claims such as:

(a) "published MAP improvements on TREC8 for BM25 with Rocchio pseudo-relevance feedback do not generalize to the TREC Entity Track."

(b) "published production interleaving improvements on Bingle, a portal web search engine, for ranking with LTRx do not generalize to Twitbook post search, a comparable production search task."

Papers in this area serve to test the sensitivity of results in IR research to task definitions.

2.4 Inexplicability

Finally, in some cases, IR research involves testing hypotheses that we expect to be positive, based on prior work in IR or related disciplines. We would also like to test the ability to generalize to tasks that are either vaguely comparable to, or completely different from, previously-studied tasks. A submission in this category was expected to fail to obtain improvements using well-established principles/methods or well-motivated approaches. Hypothetical examples include papers making claims such as:

(a) "pseudo-relevance does not improve performance on image retrieval."

(b) "incorporating social signals does not improve production portal web search."

Papers in this area serve to test the sensitivity of results in IR research to task definitions and help understand the limits in applying straightforward techniques in novel domains.

2.5 The Open-Source IR Reproducibility Challenge

The goal of the challenge was to generate reproducible baselines by inviting authors of open source search engines to index and search on the same collections of data and on physically the same computer (an Amazon EC2 instance). Producing baselines is more challenging than it appears. To provide two examples: Mühleisen et al. [5] reported large differences in effectiveness across four systems that all purport to implement BM25. Trotman et al. [6] pointed out that BM25 and query likelihood with Dirichlet smoothing can actually refer to at least half a dozen different variants; in some cases, differences in effectiveness are statistically significant. Given this state of affairs, how can we confidently report comparisons to "baselines" when the baselines are ill-defined? Indeed, Armstrong et al. [1] point to the issue of weak baselines as the reason why *ad hoc* retrieval techniques haven't really been improving.

3 Proceedings

3.1 Invited Talk

Ellen Voorhees began her talk with a history of IR starting with the 1959 International Conference on Scientific Information and ending with the inception of TREC in 1992. She went on to discuss reproducibility and how TREC in 2015 introduced "open runs", runs backed by github repositories so that others could download and compare using the same code used by the submitters.

Comparable, she pointed out, is weakly defined. She used as an example performance of systems on ClueWeb09 Category A and Category B, showing that (despite Category B being a subset of Category A), systems behave differently on each collection. Indeed, the science of determining similarity between collections and therefore predicting which system components will work well on a new collection is in its infancy.

Voorhees reminded us of the importance of the RIA workshop and the associated failure analysis. The failure analysis at RIA was able to identify cross-system failures and causes of those failures. She suggested that such an analysis may be the only way to understand some of the general principles we still seek.

She went on to remind us of the weaknesses of the Cranfield methodology. These include the variance in relevance assessments, topical relevance not being utility, static assessments not being able to model changing user needs, and the unknowability of recall. Things are changing as the community adopts standard techniques for building collections, standard metrics, and multiple collections. However much more work is needed. For example, we still know very little about the effect of a single component (e.g. tokenizer) on the overall performance of the system.

3.2 Papers

Only three papers were submitted, all were of sufficient quality for inclusion in the workshop.

In Observed Volatility in Effectiveness Metrics [4], Lu et al. discuss the stability and robustness of various metrics under the condition of increased information. The particular case of increased information they examined was more relevance assessments. In other words, does the relative rank orders of the systems remain constant as the evaluation depth increases. They show that different kinds of data (newswire vs web) exhibit different behaviour on different metrics. They also show that different types of metrics (utility based vs rank biased precision based) exhibit different behaviours. Certainly, their investigation suggests that more care needs to be taken in reporting which systems out-perform which others as this is dependent not only on the data being used and the metric, but also the depth at which the runs are being evaluated.

In Unfolding Off-the-shelf IR Systems for Reproducibility [2], Di Buccio et al. discuss a platform for identifying the hidden parameters in any IR system and measuring the effects on MAP. They raise as an example, how BM25 is implemented in Lucene vs the theoretical model, and the various methods of resolving ties when two documents have the same rsv - both of which affect MAP. They propose building a taxonomy of the components of an IR system (tokenizer, stemmer, ranker, etc.) and under this taxonomy a grid of implementations (not all Porter stemmers are the same). Then, each component can be implemented as part of a pipeline and the performance of various pipelines can be measured; consequently the effect of any one component can be measured.

In Using Simulation to Analyze the Potential for Reproducibility [3], Carterette & Sabhnani observer that given a p-value from a t-test, and the mean, and sample size, its possible to compute the variance. This is already of interest because low variance is an indication of high reproducibility. However, they go on to show that knowing the mean and variance allows them to perform statistical analyses between systems that have not previously been compared. In other words, across papers by different authors. They perform such an analysis across many papers and show that 70-80% of comparisons within a paper are likely to be reproducible, suggesting that any one comparison is relatively weak. They go on to show how to apply their technique across papers using the same document collection and the same metrics.

3.3 Reproducibility Challenge

The workshop included a reproducibility challenge exercise.[5] The purpose of this was to invite developers of open-source search engines to provide a series of reproducible baselines of their systems in a common environment on Amazon's EC2. The initial aims of the challenge were high, with the ultimate goal of being able to expose some generalizations, such as the general effect of stemming regardless of algorithm.

Developers of each system were invited to create scripts that would index and search across several test collections from TREC[6] and CLEF.[7] The TREC collection considered in these experiments is the .GOV2 collection with three sets of TREC queries: 701–750, 751–800, and 801–850. Table 3.3 summarizes the main details of the CLEF collections in 12 different European and non-European languages, considered in the experiments; all the data can be freely downloaded by means of the DIRECT[8] system.

For the TREC collection there were a total of 7 systems that provided scripts: ATIRE, Galago, Indri, JASS, Lucene, MG4J, and Terrier. The scripts were free to index and search with varying parameters. As a result, a total of 13 different indexes were generated, and 17 sets of search results. For the CLEF collections there were 3 systems together with their required scripts: Indri, Lucene, Terrier; in the case of Terrier different retrieval models (BM25, Hiemstra LM, PL2, and TFIDF) were experimented in conjunction with different configurations for stop lists[9] and stemmers.[10]

The scripts were run by an individual who was not involved in producing that script, and from a clean checkout of the repository. Any issues were reported to the authors of the scripts for correction, at which point the procedure was repeated. Statistics were gathered from the systems, ranging from indexing time, index size, search speed, to effectiveness (measured by MAP@1000 by `trec_eval`). Systems were free to use multiple threads for indexing, but were constrained to one thread for searching.

The time to index and index size for .GOV2 is shown in Table 2. There is can be seen that there was large variation in indexing time, ranging from 46 minutes and 26.5 hours. Likewise there was large variation in index size. Unsurprisingly, indexes that included positional information were larger than indexes that did not.

Time to search .GOV2 is shown in Table 4 and the precision scores are shown in Table 3. There was large variation in mean time per query, but only minimal difference in MAP@1000. Even systems that purported to implement the same ranking function showed small differences in MAP@1000, even with the same values for parameters. The "Terrier & DPH + Bo1 QE" run of Terrier had statistically significantly better MAP@1000 than all other runs. Both Lucene runs were statistically significantly better than Terrier's BM25

[5] `github.com/lintool/IR-Reproduciblity`
[6] `http://trec.nist.gov/`
[7] `http://www.clef-initiative.eu/`
[8] `http://direct.dei.unipd.it/`
[9] `http://members.unine.ch/jacques.savoy/clef/index.html`
[10] `https://github.com/snowballstem`

Language	Corpora	Docs	Topics	Topic IDs
Bulgarian (bg)	SEGA 2002 STANDART 2002	69,195	149	251-291; 293-325; 351-375; 401-450
Dutch (nl)	ALGEMEEN 1994 & 1995 NRC 1994 & 1995	190,604	156	41-159; 161-165; 167-190; 192-193; 195-200
Finnish (fi)	AMULEHTI 1994 & 1995	55,344	120	92; 94-95; 98; 100; 102-103; 105-107; 109; 111; 114-116; 118-119; 122-126; 128; 130-132; 137-140; 142-143; 147-159; 161-166; 168; 170-174; 176-181; 183-185; 187; 190; 192-193; 196-205; 207-230; 232-239; 241-246; 248-250;
French (fr)	LEMONDE 1994 & 1995 ATS 1994 & 1995	177,452	99	251-331; 333-350
German (de)	FRANKFURTER 1994 SDA 1994 SPIEGEL 1994 & 1995	225,371	155	41-43; 45-143; 145; 147-169; 171-190; 192-200
Hungarian (hu)	MAGYAR 2002	49,530	148	251-325; 351-369; 371-375; 401-450
Italian (it)	AGZ 1994 & 1995 LASTAMPA 1994	157,558	90	41-42; 44-49; 60-63; 65-69; 80-99; 110-119; 130-145; 147-149; 161-168; 171; 173-174; 176-179; 190; 192-200
Persian (fa)	HAMSHAHRI	166,774	100	551-650
Portoguese (pt)	FOLHA 1994 & 1995 PUBLICO 1994 & 1995	210,734	100	251-350
Russian (ru)	IZVESTIA 1995	16,716	62	143; 147-149; 151; 153-155; 157; 163-164; 168-169; 172; 176-181; 183; 187; 192-193; 197-203; 207; 209-216; 218; 220-221; 224-228; 230-235; 237-239; 241-242; 244-245; 250
Spanish (es)	EFE 1994 & 1995	454,045	97	41-49; 60; 62-69; 80-99; 110-119; 130-149; 160-168; 170-179; 190-200
Swedish (sv)	TT 1994 & 1995	142,819	103	91-109; 111-159; 161-166; 168-190; 192-193; 195-197; 199-200

Table 1: Benchmarked CLEF collections. ISO 639:1 two letters code within brackets.

System	Type	Size	Time
ATIRE	Count	12 GB	46m
ATIRE	Count + Quantized	15 GB	56m
Galago	Count	15 GB	6h 32m
Galago	Positions	48 GB	26h 33m
Indri	Positions	92 GB	6h 40m
JASS	ATIRE Quantized	21 GB	58m
Lucene	Count	12 GB	1h 25m
Lucene	Positions	40 GB	1h 35m
MG4J	Count	8 GB	1h 25m
MG4J	Positions	37 GB	2h 06m
Terrier	Count	10 GB	8h 04m
Terrier	Count (inc direct)	19 GB	18h 16m
Terrier	Positions	36 GB	9h 37m

Table 2: .GOV2 indexing results.

			Topics			
System	Model	Index	701–750	751–800	801–850	Combined
ATIRE	BM25	Count	0.2616	0.3106	0.2978	0.2902
ATIRE	Quantized BM25	Count + Quantized	0.2603	0.3108	0.2974	0.2897
Galago	QL	Count	0.2776	0.2937	0.2845	0.2853
Galago	SDM	Positions	0.2726	0.2911	0.3161	0.2934
Indri	QL	Positions	0.2597	0.3179	0.2830	0.2870
Indri	SDM	Positions	0.2621	0.3086	0.3165	0.2960
JASS	1B Postings	Count	0.2603	0.3109	0.2972	0.2897
JASS	2.5M Postings	Count	0.2579	0.3053	0.2959	0.2866
Lucene	BM25	Count	0.2684	0.3347	0.3050	0.3029
Lucene	BM25	Positions	0.2684	0.3347	0.3050	0.3029
MG4J	BM25	Count	0.2640	0.3336	0.2999	0.2994
MG4J	Model B	Count	0.2469	0.3207	0.3003	0.2896
MG4J	Model B+	Positions	0.2322	0.3179	0.3257	0.2923
Terrier	BM25	Count	0.2432	0.3039	0.2614	0.2697
Terrier	DPH	Count	0.2768	0.3311	0.2899	0.2994
Terrier	DPH + Bo1 QE	Count (inc direct)	0.3037	0.3742	0.3480	0.3422
Terrier	DPH + Prox SD	Positions	0.2750	0.3297	0.2897	0.2983

Table 3: .GOV2 MAP@1000 scores.

System	Model	Index	Topics 701–750	751–800	801–850	Combined
ATIRE	BM25	Count	131ms	176ms	131ms	146ms
ATIRE	Quantized BM25	Count + Quantized	91ms	93ms	85ms	90ms
Galago	QL	Count	769ms	820ms	661ms	750ms
Galago	SDM	Positions	4134ms	6091ms	3943ms	4723ms
Indri	QL	Positions	1338ms	1715ms	1205ms	1419ms
Indri	SDM	Positions	8146ms	14277ms	7093ms	9839ms
JASS	1B Postings	Count	47ms	50ms	45ms	47ms
JASS	2.5M Postings	Count	26ms	25ms	25ms	26ms
Lucene	BM25	Count	148ms	109ms	141ms	133ms
Lucene	BM25	Positions	119ms	111ms	118ms	116ms
MG4J	BM25	Count	362ms	257ms	267ms	295ms
MG4J	Model B	Count	37ms	48ms	36ms	40ms
MG4J	Model B+	Positions	91ms	90ms	73ms	85ms
Terrier	BM25	Count	357ms	277ms	296ms	310ms
Terrier	DPH	Count	441ms	338ms	369ms	383ms
Terrier	DPH + Bo1 QE	Count (inc. direct)	1633ms	1323ms	1402ms	1452ms
Terrier	DPH + Prox SD	Positions	1250ms	950ms	986ms	1062ms

Table 4: .GOV2 average search time across 3 runs.

based runs. Significance measured as $p < 0.05$ after MCP using Tukeys HSD. Thanks to Ben Carterette for providing the statistical analysis.

The precison scores on the CLEF collections (MAP@1000 calculated by `trec_eval`) are reported in Table 3.3. As with .GOV2, variation in the scores for preportedly the same ranking function can be seen on the CLEF collections.

An overview of the challenge was presented by Matt Crane, then each system was presented by a representative. Andrew Trotman presented ATIRE & JASS, Jimmy Lin presented Lucene on behalf of its authors, Paolo Boldi presented MG4J, and Craig Macdonald presented Terrier. Giorgio Maria Di Nunzio presented the experiments conducted with CLEF data by the University of Padua and University of Montreal research groups.

Following the presentations was a discussion on the results and future of the challenge. This included a debate on the effectiveness results. Multiple issues were then raised, one of which was that the RIGOR results were not identical to previously published results for the search engines, sometimes being higher, sometimes lower. Lower results were seen because the systems were being used "out of the box" without training on the collections. Higher results might have been seen because of system improvements since prior results. Variation in BM25 scores were seen for several reasons, including tuning parameters, subtle differences in interpretation of the equations (e.g. the IDF component), and optimisations. This part of the discussion included a proposal to continue the challenge, but to allow system tuning.

Finally, there was heated debate on the negative consequences of the challenge. This included the necessity to avoid undoing or re-doing all the incredible work of TREC. However, there was concern that by publishing league tables on multiple collections that the challenge would inadvertently create an extra publishing hurdle for those using open source search

System	Model	Stop	Stem	bg	de	es	fa	fi	fr
Terrier	BM25			0.2092	0.2733	0.3627	0.4033	0.3464	–
Terrier	BM25	✓		0.2081	0.2742	0.3656	0.4022	0.3392	–
Terrier	BM25		✓	–	0.3194	0.4347	–	0.4339	–
Terrier	BM25	✓	✓	–	0.3215	0.4356	–	0.4278	–
Terrier	Hiemstra LM			0.1647	0.2520	0.3016	0.3140	0.3125	–
Terrier	Hiemstra LM	✓		0.1640	0.2561	0.3081	0.3193	0.3156	–
Terrier	Hiemstra LM		✓	–	0.2753	0.3673	–	0.3639	–
Terrier	Hiemstra LM	✓	✓	–	0.2801	0.3783	–	0.3636	–
Terrier	PL2			0.2043	0.2625	0.3486	0.4081	0.3316	–
Terrier	PL2	✓		0.2009	0.2658	0.3572	0.4061	0.3388	–
Terrier	PL2		✓	–	0.3080	0.4168	–	0.4222	–
Terrier	PL2	✓	✓	–	0.3102	0.4211	–	0.4152	–
Terrier	TFIDF			0.2071	0.2709	0.3597	0.4050	0.3457	–
Terrier	TFIDF	✓		0.2083	0.2723	0.3658	0.4053	0.3393	–
Terrier	TFIDF		✓	–	0.3185	0.4313	–	0.4354	–
Terrier	TFIDF	✓	✓	–	0.3167	0.4355	–	0.4269	–
Lucene	BM25	✓	✓	–	0.3126	0.4251	0.4158	–	0.3865
Indri	LM Dirichlet	✓	✓	0.2051	0.1365	0.3334	0.3735	–	0.1444

System	Model	Stop	Stem	hu	it	nl	pt	ru	sv
Terrier	BM25			0.2115	0.3233	0.3958	0.3250	0.3666	0.3384
Terrier	BM25	✓		0.2178	0.3182	0.3974	0.3255	0.3449	0.3371
Terrier	BM25		✓	0.3175	0.3619	0.4209	0.3250	0.4740	0.3817
Terrier	BM25	✓	✓	0.3254	0.3591	0.4234	0.3255	0.4753	0.3886
Terrier	Hiemstra LM			0.1642	0.2778	0.3454	0.2738	0.2922	0.3113
Terrier	Hiemstra LM	✓		0.1685	0.2820	0.3523	0.2742	0.2949	0.3160
Terrier	Hiemstra LM		✓	0.2559	0.3061	0.3585	0.2738	0.3891	0.3372
Terrier	Hiemstra LM	✓	✓	0.2656	0.3092	0.3680	0.2742	0.3960	0.3402
Terrier	PL2			0.2060	0.3110	0.3792	0.3183	0.3433	0.3149
Terrier	PL2	✓		0.2091	0.3090	0.3832	0.3184	0.3288	0.3222
Terrier	PL2		✓	0.3040	0.3521	0.4042	0.3183	0.4737	0.3604
Terrier	PL2	✓	✓	0.3179	0.3472	0.4088	0.3184	0.4711	0.3708
Terrier	TFIDF			0.2107	0.3238	0.3946	0.3230	0.3643	0.3344
Terrier	TFIDF	✓		0.2181	0.3205	0.3975	0.3258	0.3403	0.3354
Terrier	TFIDF		✓	0.3105	0.3675	0.4222	0.3230	0.4764	0.3789
Terrier	TFIDF	✓	✓	0.3252	0.3649	0.4253	0.3258	0.4647	0.3869
Lucene	BM25	✓	✓	0.3233	0.3486	0.4172	–	0.4717	0.3775
Indri	LM Dirichlet	✓	✓	0.2381	0.0984	0.2486	–	0.2991	0.3265

Table 5: MAP@1000 scores on the benchmarked CLEF collections. Languages are expressed as ISO 639:1 two letters code. "Stop" indicates if a stop-list was used and "Stem" if a stemmer was used.

engines. Any improvements would, by necessity, have to demonstrate improvements on league table results that are not necessary for a proprietary system on a proprietary collection. This point was argued strenuously by both sides, without conclusion.

Acknowledgments

The experiments on CLEF collections were run by Maria Maistro and Gianmaria Silvello (Terrier) and by Emanuele Di Buccio (Lucene) of the IMS research group of University of Padua and by James Liu (Indri) of the DIRO research group of the University of Montreal. Thanks to Jacques Savoy who provided the language resources (i.e. stop-lists for different languages) for conducting the experiments on CLEF collections.

References

[1] T. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *CIKM 2009*, pages 601–610, 2009.

[2] E. Di Buccio, G.M. Di Nunzio, N. Ferro, D. Harman, M. Maistro, and G. Silvello. Unfolding off-the-shelf IR systems for reproducibility. In *RIGOR 2015*, 2015.

[3] B. Carterette and K. Sabhnani. Using simulation to analyze the potential for reproducibility. In *RIGOR 2015*, 2015.

[4] X. Lu, A. Moffat, and S. Culpepper. Observed volatility in effectiveness metrics. In *RIGOR 2015*, 2015.

[5] H. Mhleisen, T. Samar, J. Lin, and A. de Vries. Old dogs are great at new tricks: Column stores for ir prototyping. In *SIGIR 2014*, pages 863–866, 2014.

[6] A. Trotman, A. Puurula, and B. Burgess. Improvements to BM25 and language models examined. In *ADCS 2014*, pages 58–65, 2014.