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

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

[1] University of North Carolina at Chapel Hill [2] Microsoft Research
[3] University of Maryland, College Park [4] eBay Inc.

jarguello@unc.edu, fdiaz@microsoft.com, jimmylin@umd.edu

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and Software—Performance evaluation

Keywords: evaluation; negative results

1. MOTIVATION

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) aims to provide a venue for publication and discussion of IR research that fails to reproduce a previously published result under the same or similar experimental conditions (e.g., same test collection and system configuration) and research that demonstrates the failure to generalize an existing approach to a new domain. To this end, we have developed a set of categories covering different ways in which a result may fail to reproduce or generalize as well as a set of reviewing principles unique to this style of research. We believe the RIGOR Workshop provides a forum in which to present and discuss an under-represented aspect of IR research.

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

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767858.

2. SCOPE

We see papers in this workshop focusing on different scenarios in which previous results might fail to reproduce. Specifically, we invite 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 demonstrates a failure to repeat a published result under approximately the same conditions in which the previously published experiments occurred. Examples include 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 judgments. In many cases, these experiments can be conducted on other comparable corpora, queries, or relevance judgments. A submission in this category fails to reproduce a published result on a comparable dataset. 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 fails to reproduce a published result on a comparable task. 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 the ability to generalize to tasks that are either vaguely comparable to or completely different from previously-studied tasks. A submission in this category fails to obtain improvements using well-established principles/methods or well-motivated approaches. 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.