The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction

Ferro, Nicola; Fuhr, Norbert; Grefenstette, Gregory; Konstan, Joseph A.; Castells, Pablo; Daly, Elizabeth M.; Declerck, Thierry; Ekstrand, Michael D.; Tintarev, Nava; More Authors

DOI
10.1145/3274784.3274789

Publication date
2018

Document Version
Accepted author manuscript

Published in
ACM SIGIR Forum

Citation (APA)
Ferro, N., Fuhr, N., Grefenstette, G., Konstan, J. A., Castells, P., Daly, E. M., Declerck, T., Ekstrand, M. D., Tintarev, N., & More Authors (2018). The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction. *ACM SIGIR Forum*, *52*(1), 91-101. https://doi.org/10.1145/3274784.3274789

WORKSHOP REPORT

The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction

Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan, Pablo Castells,
Elizabeth M. Daly, Thierry Declerck, Michael D. Ekstrand, Werner Geyer,
Julio Gonzalo, Tsvi Kuflik, Krister Lindn, Bernardo Magnini, Jian-Yun Nie,
Raffaele Perego, Bracha Shapira, Ian Soboroff, Nava Tintarev, Karin Verspoor,
Martijn C. Willemsen, Justin Zobel

Abstract

This paper reports the findings of the Dagstuhl Perspectives Workshop 17442 on performance modeling and prediction in the domains of Information Retrieval, Natural language Processing and Recommender Systems. We present a framework for further research, which identifies five major problem areas: understanding measures, performance analysis, making underlying assumptions explicit, identifying application features determining performance, and the development of prediction models describing the relationship between assumptions, features and resulting performance.

1 Introduction

The IR field has always had a strong evaluation focus on the effectiveness of models and methods. However, for many years, the validity of experiments has been regarded as being implicitly given by following established experimentation procedures.

Internal validity refers to the fact that the conclusions derived from an experiment are actually supported by the data. This topic has gained some attention recently, by putting more emphasis on the reproducibility of the experiments [7] as well as on proper statistical methods [12].

External validity is the extent to which the results of a study can be generalized. In IR, researchers usually apply their methods to a number of test collections. However, even if the proposed approach yields performance improvements in each case, it is unclear for which other applications this finding will also be true.

In Fall 2017, researchers from the areas of IR, natural language processing and recommender systems met for a week-long workshop in Dagstuhl, Germany, for developing a research agenda for the issue of performance modeling and prediction. The outcome of this workshop are described in a Dagstuhl Manifesto [11]. In the remainder of this paper, we summarize the major findings of this report from an IR point of view.

2 Prediction Problems in IR

Current approaches to IR evaluation mean that predictability can be poor, in particular:

- Assumptions or simplifications made for experimental purposes may be of unknown or unquantified validity; they may be implicit.
- Test collections tend to be specific, and to have assumed use-cases; they are rarely as heterogeneous as ordinary search.
- Test environments rarely explore cases such as poorly specified queries, or the different uses of repeated queries.
- Researchers typically rely on point estimates for the performance measures, instead of giving confidence intervals.
- Highly correlated measures (e.g. MAP vs. nDCG are often reported as if they were independent; while, on the other hand, measures reflecting different quality aspects (e.g. precision vs. recall) are averaged (e.g. in the F-measure), thus obscuring their explanatory power.
- Current analysis tools are focused on sensitivity (differences between systems) rather than reliability (consistency over queries).
- Summary statistics are used to demonstrate differences, but the differences remain unexplained. Averages are reported without analysis of changes in individual queries.
- Due to the gap between offline and online evaluation, offline predictions of changes in user satisfaction continue to be poor because the mapping from metrics to user perceptions and experiences is not well understood.

3 Performance Prediction Framework

In order to predict performance, a framework model was developed at the workshop (see figure 1). First we have to choose the performance criteria and define corresponding *measures*. When performing experiments with different test collections and observing the system's output and the measured performance, we will carry out a *performance analysis*. For that, we will look at violations of the *assumptions* underlying the method applied. Also, characteristics of the data and tasks will have an important effect on the outcome. Finally, we aim at developing a *performance prediction model* that takes these factors into account.

Once these tasks are well understood we can begin to try and predict performance in an unseen situation if enough of the above still hold.

3.1 Measures

The definition of a metric relies on several alternatives and decisions, which happen before the actual measurement takes place, also to avoid any post-hoc bias.

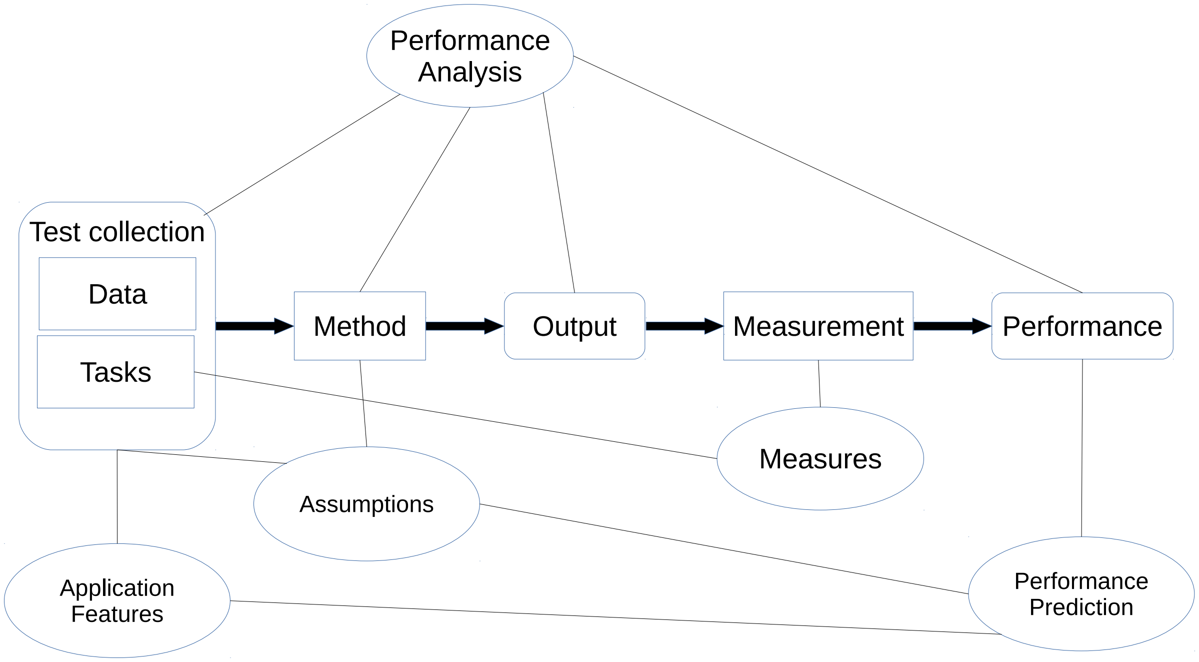

Figure 1: An overarching performance prediction framework

We first have to choose the criteria that reflect the goals of our evaluation, for instance relevance, diversity, or novelty. However, as said before, the performance of a system is not only a matter of goals but also of the "utility" delivered to users. Therefore, we need to identify/choose a prototypical user behavior; for example, when ranking is involved, a stopping point, after which no more recommended items or retrieved documents are considered, introduces a clear separation between seen and unseen items, where only the former influence the measurement outcome. Instead of a deterministic behavior, we might also assume a stochastic model for this aspect as done, for example, by [3, 4, 15]. We also have to define the user's preferences concerning the items seen, like e.g. the total number of useful items, or the ratio between useful and useless items.

Finally, we have to choose an aggregation method like arithmetic or geometric mean, where the former focuses on absolute differences, and the latter on relative changes, paying attention that the aggregation method is admissible when considering the scale properties – ordinal, interval, ratio scales – of a measure [5, 12]

Overall, we must be aware that each metric represents a specific user standpoint, and often the standpoint may be context-dependent. Thus, an evaluation focusing on a single low-level metric will either ignore many user standpoints, or represent an intransparent mixture of different standpoints.

3.2 Performance Analysis

Reporting of averages and average improvements is often unhelpful, and is uninformative in terms of explaining what system elements contributed to success, what data and queries the method is applicable to, and for which data and queries the method fails. That is, instead of focusing on statistically significant differences in the average from control to treatment, we need to move to understanding the changes in specific tasks and task types, and to understanding the contributions of individual system components.

In this context, researchers also should pay attention to the problems of multiple testing and sequential testing: When performing multiple significance tests on the same data set, they must adjust the significance level accordingly, using e.g. Bonferroni's method[1] [12]. Even more problematic is the sequential testing case, where the same data is used by other researchers, who have learned from previous results on the same test collection, and then perform significance tests for their new method(s), not considering the large number of tests carried out before. As shown in [2], this usually leads to totally random results. As a consequence, statistically meaningful results cannot be expected from heavily re-used test collections. A similar statement might also hold for multiple qualitative analyses on the same data set. Thus, re-use of a test collection is problematic, which leads to the need for more (and more diverse) collections.

Another important viewing angle is the consideration of measures representing different user standpoints: instead of focusing on universal performance, more emphasis should be put on performance differences wrt. different metrics. E.g., in retrieval, many users will look at the top ranking documents only (e.g. in Web search), while others are aiming at locating all potentially relevant documents (e.g. patent search). Thus, instead of looking at overall performance only, it is more interesting to identify methods that support specific user standpoints.

Classic failure analysis inspects individual tasks where performance is significantly altered, but other data interrogation methods, such as systematic addition of noise, can illustrate the robustness and vulnerabilities of the system that is under investigation.

3.3 Documenting and Understanding Assumptions

3.3.1 Role of assumptions

Any method or model is based on certain assumptions, some of which are explicit; usually, there is an even larger number of implicit assumptions. The performance of a method in an application depends mainly on the extent to which these assumptions are true in this setting. Thus, we have to solve three problems:

- Identify the underlying assumptions (make implicit ones explicit).
- Devise methods for determining if or to what extent these assumptions are fulfilled in an application.
- Develop a model that tells us how the violation of an assumption affects performance.

Only when we have answers to these three questions, we are able to make reasonable predictions.

3.3.2 Assumption categories

Assumptions come into play at many points in the design and evaluation of recommender, IR, and NLP systems. At each point, there are at least two broad categories of assumptions: *fundamental* assumptions and *convenience* assumptions. These two categories are transversal to different kinds of

[1] https://en.wikipedia.org/wiki/Multiple_comparisons_problem

assumptions which we can distinguish according to the role they play in data, algorithms/techniques, evaluation protocols and metrics, and their implications on system performance and the validity of research findings. Overall, they determine a taxonomy which we can use to systematically check and make them explicit.

Convenience assumptions are simplifications (or approximations) intended to make problems tractable, reduce their complexity and/or enable evolving some starting point theoretical expression (e.g. a probability) into a computable form (counting things and doing math upon numbers). Examples include the mutual feature independence assumption in Naive Bayes (of which pairwise word independence in text IR can be seen as a particular case), whereby joint probabilities are decomposed into products of simpler distributions; user, time and context independence as a means to eliminate variables from IR and recommendation problems; or document relevance independence assumption, which enables the definition of simple and easy to compute metrics such as precision. Convenience assumptions may be violated, and yet the algorithm or the metric may still work reasonably. On the other hand, performance differences between collections may be traced back to the violation of certain assumptions.

Convenience assumptions typically represent an opportunity to define new research problems consisting of the elimination of a particular simplifying assumption and dealing with the corresponding complexity. An example is personalized IR, which takes the user variable back into the problem and copes with it; or IR diversity, which removes the document relevance independence assumption; or time-aware or context-aware IR, which do the same with time and context.

By fundamental assumptions we mean hypotheses that algorithms or metrics themselves build upon they are intrinsic to the underlying model. For instance, content-based recommendation assumes item features can partially explain user choices; IR language model algorithms assume language similarity is related to relevance; most text IR models assume term frequency matters; proximity search algorithms assume word order matters too; metrics like precision assume users want to get relevant documents; average search length (rank of first relevant document) assumes users need just one relevant document or item; recommendation diversity metrics may assume people enjoy variety; novelty metrics assume users wish to be surprised; an experimental protocol may assume each and every user has a non-empty set of training (or test) observations. When fundamental assumptions fail to be met, the algorithm or the metric may no longer be effective or valid. Content-based recommendation is as good as random if user choices are unrelated to item features; a novelty metric is irrelevant if users are just willing to stick to familiar experiences; lack of data for a single user may result in an undefined evaluation outcome.

Becoming aware of and understanding fundamental assumptions enables a better and more consistent use of the tool (algorithm, metric, protocol) that builds upon them, and may prevent unintentional misuse. It can also help detect spurious confounders (biases that cause the hypothesis to hold for misleading reasons) and experimental flaws that can easily go unnoticed (e.g. a recommendation algorithms accuracy skyrockets simply because we forgive it refusing to deliver recommendations to certain users; depending on the characteristics of these users, this may result in discriminatory quality of service).

3.3.3 Understanding Violations to Assumptions

A critical aspect in explaining and predicting performance is to understand whether and to what extent the assumptions our methods are based upon have been complied with or violated.

This understanding should happen at both theoretical and experimental level. At theoretical level, among the various assumptions, we should be able to differentiate those that are crucial for a method and whose violations seriously hamper its application from those that are somehow desirable. At experimental level, we should have techniques for assessing each assumption and understand whether and how much it has been violated.

We need to develop commonly agreed *scales* to quantify how much an assumption has been violated. However, given the wide range and diversity in the type of assumptions we have, we should aim at developing assumption checking methods and scales that hold, at best, for families of related assumptions rather than hoping for a single general solution where one-fits-all.

Then, we need to research on the relationship between the severity of departures from assumptions, quantified in the above mentioned scales, and the observed and predicted performances. The final goal is to understand how much resilient are our methods to such violations and how much this impacts on explanation, first, and prediction, after.

Violations of algorithm or technique assumptions are perhaps the easiest to assess: run the algorithm on a data set that violates its assumption(s) and measure its performance and behavior. Violations of evaluation and data assumptions are more challenging, as they undermine the tools by which we measure the behavior of the system in the first place. To assess the impact of these assumptions, we need techniques that allow us to peek behind the curtain and understand the behavior of these components of the experimental process under a range of possible truths, in order to relate their output to our confidence about the relationship of the data and evaluation to the underlying truth and intended task.

An area we can take inspiration from is statistics and the notion of *robustness* in statistical testing, meant as "insensitivity to small deviations from the assumptions" [13]. Robustness is developed both a theoretical level, e.g. by studying it under a null and an alternative hypothesis [14], and defining indicators such as, for example, the breakdown point, i.e. the proportion of incorrect observations an estimator can handle before giving an incorrect result.

Furthermore, simulation and resampling are particularly promising tools for quantifying the importance of assumptions to components of the information processing and evaluation pipeline. Measuring results on different data sets is useful, but only provides a few data points regarding the behavior of a method or evaluation technique, and does not change the relevant variables in a controlled fashion; further, the data set's relationship to underlying ground truth cannot be controlled and may not be known. Simulation and resampling allows a range of possibilities — some within assumptions, some outside — to be tested, and the relationship of data to truth to be controlled, allowing us to precisely characterize the system response to targeted violations of its assumptions. These experiments can take multiple forms, including wholly-synthetic data, resampling of traditional data sets, and sampling of specialized data sets such as ratings collected on complete or uniformly sampled sets of items. As one example, [18] employed simulation to study the robustness of information retrieval evaluations to violations of statistical assumptions about the underlying data sets and their topic distributions.

3.3.4 Increasing awareness in our community

There is a large variance on how assumptions are managed and on the perception itself of their importance. A general recommendation is pushing in any possible way our community to a greater awareness of the need for making assumptions explicit and clear. Inserting in all scientific works a clear statement permitting a precise identification and a deeper understanding of the essential assumptions made and their scope of validity should become a universal practice. To this regard we recall the effort currently conducted in the IR community toward reproducibility of results [6, 7]: after a consciousness campaign last several years, we now have a reproducibility track in the main IR conferences and reproducibility tasks have been just launched in the major evaluation campaigns[2].

The awareness on such an important aspect impacting the validity and reproducibility of results can be disseminated and increased in several ways. A first recommendation is adding an explicit reference to the clarity and completeness of assumptions made in the call for paper and the paper review forms of all conferences. This can have the double effect of educating the reviewers to reserve a particular attention to assumption clarity and, on the other hand, to increase author's awareness on them. Papers claiming results involving assumptions that are not explicitly voiced or understood should not be deemed as solid since no strong conclusion can be drawn from them. As a second step, after a systematization of assumptions and a greater understanding have been reached, the emerging best practices can give origin to commonly accepted requirements to be integrated in the call for papers of specific tracks.

It is significantly harder to test the importance of assumptions in user-facing aspects of the system, such as the presentation of results or the task model, as it is prohibitively expensive to simulate arbitrarily many versions of a system and put them before users. System utility can be remarkably robust to violations of core assumptions — for example, e-commerce vendors obtain great value from collaborative filtering techniques that assume items are functionally interchangeable even when they clearly are not — but rigorous empirical data on this robustness is difficult to obtain. However, measuring hidden factors (see [11]) might help explain why particular versions of a system perform better, directly testing underlying assumptions.

3.4 Application features

One common feature of Natural Language Processing, Information Retrieval and Recommender Systems is the wide space of data and task characteristics that have to be accounted for when designing a system. Adapting existing systems to a new domain, a new data set, or a new task, and then predicting their performance in this new setting is particularly challenging in our research fields because there is always some degree of mismatch between testing and development conditions (either in laboratory or real-world settings) used to create the existing systems and new application area.

As a result, measuring only effect sizes and statistical significance is of little help for predicting out-of-the-lab performance. Even moving between two test collections which apparently share the same features often results in different experimental outcomes. In order to have predictive power, evaluation methodologies need a much higher emphasis on explanatory analysis: why, where and

[2] `http://www.centre-eval.org/`

how systems fail is more relevant than effect sizes on average measures.

In this section we begin by reviewing a few measurable characteristics that make prediction possible but challenging in our research fields, and we then move to advocating explanatory analysis.

3.4.1 Task & Data features

How will an existing method, algorithm or system perform under conditions different from the ones in which it was tested? There are some easily identifiable features related to the data or the task that, if changed, may affect predictability.

With respect to the task, some relevant characteristics are the **language** involved in the task. Will the task be performed using monolingual or multilingual data. Will the output be in a different language from the input (cross-linguistic)? Or is the task language independent? Are there the necessary language resources for the task? Does the task involve some dialect for which these resources have to be adapted? Is the data based on speech, on written text?

Another characteristic of the task is it **dynamicity**: are we dealing with a static collection, or a stream of data? Is the task a one-off, ad-hoc task, or a long standing task, such as filtering a news stream with a static query? Is the task offline, or online, performed with an active user? Does the task change over time as the user performs it?

Task **context** also plays an important role in many situations: Current Web search engines consider already user history, location, time and end device when computing the search result. The same might be true for other types of tasks.

We can characterize the data as **curated**, for example scientific papers, or edited news stories, or as naturalistic, for example, stemming from social media, or transcribed speech. In the latter case, one can sometimes measure the expected error rate, such as the frequency of spelling errors, or transcription errors. Many language processing tools were developed for curated language, without such errors.

Another dimension of data is its **connectedness** or **structure**. Can each data item be considered as a separate item, or are there links between items? For example, web pages link to other web pages. A collection of movies can contain a series of implicitly linked sequels. Users in a social network have both explicit and implicit connections to other users. Each data item can have internal structure (metadata such as timestamps, hand-assigned classification codes, numerical data; or internal structure, such as abstract, body, supplemental material).

With respect to (textual) data, some measurable features are: readability and comprehensibility; domain; users' expertise; how source and target data correlates; verifiability of answers; dependence on assumptions to construct ground truth; richness of features; external validations; existence of corner cases and stress factors; parameterizations that impact performance; quality of domain resources (ontologies, dictionaries, taggers, etc.)

These characteristics, however, are not likely to be sufficiently predictive: even when they are the same in the new application as in laboratory conditions, often components of the systems perform differently. One of the main shortcomings of our experimental methodology is the lack of adequate explanatory methods.

3.4.2 Bias and Scaling

Test collections are often not a representative sample of a larger population. Instead, they have been compiled under certain restrictions (e.g. in IR test collections, rather specific or too general topics are not considered). We need to understand the limitations and bias of our sampling methodology across topics, documents, and systems. Can we determine when differences are due to bias, or when we are sampling from separate distributions?

Another problem is scaling: methods doing well on small test collections might not work on collections orders of magnitude larger, and vice versa.

3.5 Modeling Performance

Trying to explain and model the performance of systems over different datasets and tasks is a preliminary yet indispensable step towards envisioning how to predict the performance of such systems. However, this is often difficult to do due the lack of appropriate analysis techniques and the need for careful experimental designs and protocols, which may be complex and demanding to carry out.

There is therefore a need for further research providing us with the methods for analyzing and decomposing the performance into those of the affecting factors, such as system components, datasets, tasks, and more. These explanatory models will then constitute the basis for developing predictive models.

Performance prediction can take different forms. We commonly wish to make an *ordinal* prediction, of which of two systems will be superior for a kind of task over a class of collections. For a single system, we might aim at an *interval* prediction, giving us a confidence interval for a certain metric; the most simple case would be a prediction for another sample from the same population. While these two approaches target at average performance, we may alternatively wish to estimate risk or *uncertainty*, that is, predict a likelihood of failure.

3.5.1 Performance factor analysis in IR

In the case of IR, over the years, there have been examples of attempts to decompose performance into constituting factors, based on the use of General Linear Mixed Models (GLMM) and ANalysis Of VAriance (ANOVA).

[1, 17] have shown how to break down the performance of an IR system into a Topic and a System effect, finding that the former has a much bigger impact than the latter.

By using a specific experimental design, [9, 10] also broke down the System effect into those of its components – namely stop lists, stemmers, and IR models. They further demonstrated that we are not actually evaluating these components alone, even when we change only one of them and keep all the rest fixed. Rather, we are evaluating whole pipelines where these components are inserted and with which they may have positive (or negative) interaction, boosting (or depressing) their estimated impact.

The difficulty in estimating the Topic*System interaction effects is the lack of replicates for each (topic, system) pair in a standard experimental setting. Therefore, [16] used simulation based on distributions of relevant and not relevant documents to demonstrate the importance of the Topic*System interaction effect. Very recently, [19] exploited random partitions of the document corpus to obtain more replicates of each (topic, system) pair, obtaining an estimation of the Topic*System interaction effect which allowed for improved precision in determining the System effect.

Finally, [8] conducted preliminary studies on the effect of Sub-Corpora and the System*Sub-Corpus, showing their impact and how they can be exploited to improve the estimation of the System effect.

All these GLMMs are not connected yet, meaning that they tackle the problem separately from different viewpoints but there is not yet a single model integrating all these facets. So a first required step toward performance prediction is to unify all these explanatory models into a single one. Then, the next step is to turn these models into predictive ones, e.g. by using some of the features discussed in Section 3.4 to learn how to predict the factors described in these models.

4 Conclusion

Performance prediction in IR is a research problem that has been ignored for many years. In this paper, we have presented a framework for starting research in this area. Some problems might call for substantial resources before they can be addressed. For instance, the analysis of performance-determining application features requires a larger number of more diverse testbeds. Most of the problems, however, just ask for a more analytic approach. Instead of focusing only on performance improvement/system tuning, researchers should aim at improving our understanding of why, how and when the investigated methods work.

References

[1] David Banks, Paul Over, and Nien-Fan Zhang. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval*, 1(1-2):7–34, May 1999.

[2] Ben Carterette. The best published result is random: Sequential testing and its effect on reported effectiveness. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 747–750, New York, NY, USA, 2015. ACM.

[3] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors, *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, pages 621–630. ACM Press, New York, USA, 2009.

[4] M. Ferrante, N. Ferro, and M. Maistro. Injecting User Models and Time into Precision via Markov Chains. In S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, and K. Järvelin, editors, *Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, pages 597–606. ACM Press, New York, USA, 2014.

[5] M. Ferrante, N. Ferro, and S. Pontarollo. Are IR Evaluation Measures on an Interval Scale? In J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz, editors, *Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017)*, pages 67–74. ACM Press, New York, USA, 2017.

[6] N. Ferro. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)*, 8(2):8:1–8:4, February 2017.

[7] N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum*, 50(1):68–82, June 2016.

[8] N. Ferro and M. Sanderson. Sub-corpora Impact on System Effectiveness. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White, editors, *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 901–904. ACM Press, New York, USA, 2017.

[9] N. Ferro and G. Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, and J. Zobel, editors, *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*, pages 25–34. ACM Press, New York, USA, 2016.

[10] N. Ferro and G. Silvello. Toward an Anatomy of IR System Component Performances. *Journal of the American Society for Information Science and Technology (JASIST)*, 69(2):187–200, February 2018.

[11] Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan, Pablo Castells, Elizabeth M. Daly, Thierry Declerck, Michael D. Ekstrand, Werner Geyer, Julio Gonzalo, Tsvi Kuflik, Krister Lindn, Bernardo Magnini, Jian-Yun Nie, Raffaele Perego, Bracha Shapira, Ian Soboroff, Nava Tintarev, Karin Verspoor, Martijn C. Willemsen, and Justin Zobel. Building a predictive science for performance of information retrieval, natural language processing, and recommender systems applications (dagstuhl perspectives workshop 17442). *Dagstuhl Manifestos*, 8, 2018.

[12] N. Fuhr. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum*, 51(3):32–41, December 2017.

[13] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, USA, 2nd edition, 2009.

[14] T. Kariya and B. K. Sinha. *Robustness of Statistical Tests*. Academic Press, USA, 1989.

[15] A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2:1–2:27, 2008.

[16] S. E. Robertson and E. Kanoulas. On Per-topic Variance in IR Evaluation. In W. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 891–900. ACM Press, New York, USA, 2012.

[17] J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 1994.

[18] Julin Urbano. Test collection reliability: a study of bias and robustness to statistical assumptions via stochastic simulation. *Information Retrieval Journal*, 19(3):313–350, December 2015.

[19] E. M. Voorhees, D. Samarov, and I. Soboroff. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)*, 36(2):12:1–12:21, September 2017.