Modelling IR System Performance Towards Predictive Evaluation

Guglielmo Faggioli[1]

[1]*University of Padova, Padova, Italy*

Abstract

Evaluation in Information Retrieval (IR) is prominently an empirical discipline – experimental collections allow us to assess the performance of the systems to compare them. This has several advantages: it allows for fair and reproducible comparisons while limiting the cost of the evaluation to developing an offline – reusable – collection. Nevertheless, it also presents some limitations: we cannot generalize our findings to previously unseen collections, data, or tasks. In this work, we lay out the cornerstone for developing a predictive performance evaluation framework for IR performance that moves from using collections as a testbed to using them as evidence to predict how a system will perform in unseen scenarios. We start from two well-known aspects of the IR evaluation and prediction, namely linear modelling of the performance (i.e., ANOVA) and Query Performance Prediction (QPP). We then identify which research directions can help realize a holistic system to predict the performance of an IR system. In particular, our research aims at investigating three main fields: i) Causal inference – a framework to study causal relations between variables in our experimental scenario; ii) General Linear Model (GLM), a generalization of the linear modelling framework that allows for relaxing some of the assumptions underlying linear modelling; iii) distributional representation of the system performance.

1. Motivation

IR evaluation is often anchored to the empirical analysis: almost every evaluation framework requires evaluation collections, which often follow the well-known Cranfield paradigm. In the Cranfield paradigm, an experimental collection contains three elements: a corpus of documents, a set of topics, and a set of relevance judgements for each topic. The development of a test collection is an expensive process. It requires hiring experts to judge the relevance of documents for each topic: the cost rapidly increases with the specificity of the domain considered. In specific areas, such as the medical one, it can also have a high ethical price: asking medical practitioners to invest time in annotating collection might be worthy only if there is a great advantage in doing that. There are also emergency periods where the development of good collections might become more complicated. Consider, for example, the crisis related to the COVID-19: it was vital to rapidly develop reliable datasets to help researchers find information to better study the disease. Experimental collections in IR are often static entities. For example, the pool of judged documents often derives from systems used during the original evaluation campaign. Works as [1] have shown that this does not invalidate the collection itself. Nevertheless, what if the definition of relevance changes, for example, due to newly gained knowledge? Are the collections still usable? Finally, the majority of the experimental collections represent the

CEUR Workshop Proceedings (CEUR-WS.org)

judgement as a single value: different users might perceive different degrees of relevance for the same item. It is clear that, in IR offline evaluation, we are inherently bounded by the development of the evaluation collections in terms of speed and quality. User studies partially solve the problem. However, even with user studies, we are still bounded to the documents considered during the experimental phase (i.e., the user study) and the system(s) taken into account. A third alternative is represented by crowd-sourcing. The growing interest in this direction has enabled the development of new evaluation paradigms. Nevertheless, such approaches introduce new and possibly more complex obstacles: for example, how much are crowd users reliable? Finally, user studies and crowd-sourcing experiments are often non-reproducible, preventing a fair comparison among different systems. The main limitation in evaluating IR Systems stems from its empirical nature. With the proposed research line, we plan to set the cornerstone for the mathematical prediction of performance in the IR evaluation. If we could predict a system's performance before deployment, we could reduce the need for experimental collections. Therefore in this work, we propose a set of techniques that can allow us to stop using the experimental collections as a testbed but rather use them as training data to generalize the performance prediction on previously unseen collections. The proposed research will rely on two main pillars: General Linear Model (GLM) and Causal Inference (CI). GLM will allow us to move further from the current linear modelling of the performance. CI describes causal relations between the experimental setup and the performance. Such property will, in turn, enable the prediction of the system behaviour in previously unseen scenarios. Additionally, we plan to study which features are the most prominent in determining the quality of a system. Furthermore, we plan to exploit distributional estimations of the system performance rather than point estimation. This allows modelling the performance of a system and better describes the real-world scenario. The remainder of this manuscript is organized as follows. Section 2 describes the main related theoretical background to the proposed framework. Section 3 contains the current state of the work. Section 4 indicates the main study fields that can provide the theoretical tools to develop the performance prediction model mentioned earlier. Finally, in Section 5 we detail the future works and draw the conclusions.

2. Background and Related Work

ANOVA And Performance Modelling An essential requirement for evaluation approaches is assessing whether systems differ statistically significantly. ANalysis Of VAriance (ANOVA) [2] is a well-known statistical technique to determine whether there is a statistically significant difference among different categorical factor levels on a continuous dependent response. It consists of modelling the experiment through a linear model, under the form *Data=Model+Error*. The *Data* is a continuous variable, such as a performance measure. The *Model* embeds the effect of the experimental conditions – e.g., the systems considered. The *Error* describes the portion of the *Data* which cannot be explained by the *Model*. ANOVA allows to break down the variance observed on the response variable on the different factors included in the model. Several works use ANOVA to model the performance of the IR systems [3, 4, 5, 6]. ANOVA allows computing the effect size of the different factors. Observing a high effect size for a specific factor means that it has a great impact on performance. Besides, ANOVA allows computing the effect size

of the interaction between factors. Replicates are required to compute interaction factors: we need multiple observations for the same experimental setup. This is not typically the scenario in IR where, given a collection, a system and a topic, we can compute a single data point [6]. To solve such limitation, works as [3, 7], randomly split the collection and compute the point performance estimation on each shard of the collection. Typically, ANOVA is associated with a posthoc procedure. Once we have the statistical evidence for considering different levels of a factor, we need to understand which pair of levels are different. Tukey's HSD test is one of the most commonly used [8].

Query Performance Prediction A first effort toward the performance prediction of IR systems is represented by the QPP models. Traditionally QPP models are divided into two macro-categories: pre- and post-retrieval QPP approaches. Systems belonging to the former category try to predict the performance of a query on a corpus without considering the system. They typically are based on features like the TF-IDF of the query terms [9] or the semantical complexity of the query [10]. The second class of QPP approaches requires to retrieve the documents using the query. Post-retrieval predictors are further divided into score distribution-based [11], language model-based [12], and robustness-based [13] approaches. QPP widely differs from our idea of predicting the performance of a system for two main reasons *i)* features considered, *ii)* expected output. Features exploited by QPP are often linked to the content of the documents and the queries or their retrieval score. In this sense, they are agnostic on the system used to retrieve the documents. Secondly, QPP do not directly predict the performance of a query. Several works [14, 15] have shown how QPP tend to perform inadequately when used to predict a retrieval metric. The value of the prediction for a query from a QPP model can only be interpreted concerning the predictions for other queries. Nevertheless, QPP poorly perform when they are required to sort queries representing the same information need [16, 17].

3. State-of-the-art

3.1. Analysis of the Prediction Features

One key aspect to enable search engine performance prediction is individuating exploitable features in the prediction phase. In [18], the authors developed a series of experiments that allow them to study the impact of different factors on the topic difficulty. [18] differs from previous works, such as [3, 4, 5, 6, 7] due to the introduction of multiple topic formulations. Furthermore, it also considers multiple collections to generalize and compare the results. The authors in [18] study the behaviour of the different query formulations under different setups by computing a Grid of Points (GoP) of performance as defined in [19] considering several components - stemmers, models, and query expansion approaches. This work allowed the authors to reach meaningful conclusions on the concept of topic difficulty. They observe how the "difficulty" is not an intrinsic property of the topics since it only relates to the triple *formulation of the topic*, *system*, *collection*. Furthermore, they observe that it is almost always possible to induce any ranking of the topics based on their performance by solely varying the components of the triple mentioned above. Such a finding highlights the pivotal role played by the multiple formulations and corpora in determining the performance of a system. Multiple topic formulations and

corpora should be considered to correctly model and predict the system performance. The interaction between the topics and the systems was already known to have a large size[6, 3, 7] while Culpepper et al. [18] illustrate how the interaction between query formulations and systems is also influential. The ideal predictor should thus be able to model such interactions. Similarly, Faggioli and Marchesin [20] investigate which features correlate with the semantic complexity of a query. In [20], authors consider that some query characteristics might influence the difference between the human interpretation of a query and its representation internal to the machine – often called semantic gap. Furthermore, the authors observe that features such as the number of synonyms for query words or the number of concepts associated with each token correlate positively with the "semantic hardness" of the query.

3.2. Query Performance Prediction Evaluation

Besides being able to predict the system performance correctly, it is necessary to assess whether the prediction model is working correctly. In this sense, the work Faggioli et al. [21] explores the domain of QPP evaluation. More in detail, the authors of [21] started from the need for ground evaluation techniques in the QPP domain. They observe that traditional evaluation techniques based solely on correlations between predicted and observe query performance tend to discriminate poorly between predictive models. The main limitation is that such correlation measures are aggregation over multiple queries. Not being able to discriminate between the complexity of different queries allows modelling only partially the QPP performance. Therefore, in [21] the authors develop a new measure, based on Spearman's footrule [22], capable of modelling both the effect of the single queries and the effect of the different predictive models. Such an approach discriminates better the quality of the system: the query complexity is modelled and does not act as a confounder. We plan to exploit the knowledge gained through [21] to develop the needed evaluation tools to assess the quality of the performance prediction models.

3.3. Analysis of the Statistical Approaches

An often under-looked problem is the replicability of statistical evaluation techniques. Such a problem prevents the development of sound comparative analyses assessing the performance of multiple systems. As a consequence, it also prevents the development of high-quality prediction models. In [23] the authors tackle the above-mentioned problem. In particular, they consider multiple ANOVA [3, 7] approaches and determine which one is more stable. All the ANOVA procedures relied on the concept of *sharding*. The sharding consists of splitting a single collection into different shards. Shards provide the required replicates to compute the interaction effect between the topics and the systems. [3] relied on a bootstrap-based implementation of ANOVA which resamples the residuals to compute multiple ANOVA models. [7] used the standard ANOVA approach. A second aspect investigated in work mentioned above is the effect of different posthoc multiple comparison procedures. For example, [3] used the lenient Benjamini-Hochberg (BH) FDR controlling procedure [24], while [7] adopted the more strict Honestly Significant Difference (HSD) Tukey procedure [8]. Additionally, in [23] the authors test also the BH procedure using the traditional ANOVA. We plan to use the insights derived from [23] to

develop a consistent and ground prediction model for the system performance. In particular, we expect our model to be capable of working with different posthoc multiple comparisons procedures. This would allow the practitioner to adapt the prediction framework to their specific need. Nevertheless, we expect our model to be robust enough to provide consistent results, independently from the idiosyncrasies of the collections or statistical procedures adopted.

4. Approach

4.1. Causal Inference

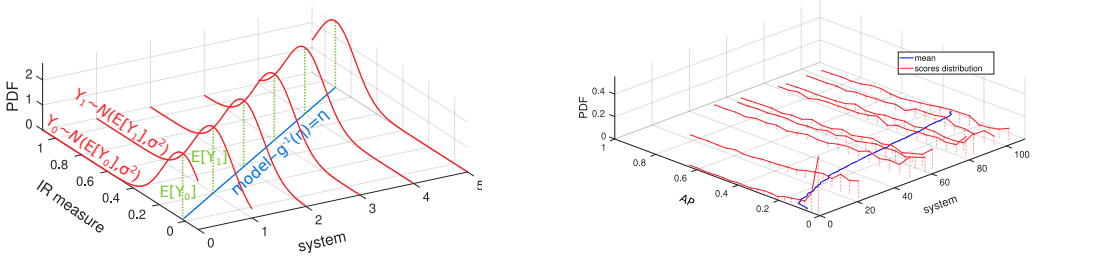
An approach to the data modelling that has recently gained interest is the Causal Inference (CI) [25, 26]. CI is based on two main components, Structural Causal Models (SCM) and the inference itself. A SCM is a Directed Acyclic Graph (DAG) composed of a set of vertices, one for each variable (causes and effects), and a set of edges that describe the causal relationship between the variables. The nodes of the graphs can represent either *exogenous* or *endogenous* variables. The formers represent variables on which we do not have control and that we cannot measure. The latter are those variables observed and measured during the data collection phase. Edges represent the functions that allow determining a node's value, given its parents' values. The topology of the graph allows inferring whether two variables in the model should be independent, conditionally to others. Note that the SCM is typically obtained through the *a priori* knowledge of the experimental setting. Using the SCM and its properties, the researcher can test the correctness of her hypotheses on how variables are linked by cause-effect relations. The most important element of the CI is the *do*-calculus. Another important tool in the CI framework is the *counterfactual reasoning*. A counterfactual describes a hypothetical situation for which we do not have any data. It corresponds to a *what-if* scenario. By looking at the post-intervention distributions, we can estimate the distribution that the effect would follow if the causes were different. In this sense, it enables predicting the effect under different scenarios. Traditionally CI has proven to be helpful mainly with observational data. This is not the typical setting in the offline evaluation of IR systems where the data are mainly experimental.

Advantages CI can grant our predictive system a solid theoretical ground. Furthermore, through CI, we can understand how features contribute to obtaining the overall result.

Challenges The main challenge to the application of CI is the type of data: CI is typically applied in an observational scenario, while in IR, we have experimental data.

4.2. Generalized Linear Models

The most popular techniques to model IR performance include t-test, and ANOVA [27, 28]. These techniques are both examples of Linear Models. Linear models based on four assumptions: *1)* independence of the samples used to fit the model; *2)* normality of the response; *3)* homoscedasticity of the data; *4)* linearity of the relation between the experimental conditions and the response. These assumptions allow for the mathematical computation of the model itself, and the more the data deviate from the ideal conditions, the lower the quality of the fitted



(a) The "ideal" linear model.

(b) Empirical IR data from Robust 04 collection [30].

Figure 1: A comparison between how the data should look like to fully exploit the potential of the linear modeling, against the empirical distribution of the data in a real IR collection.

model. Previous work [29, 5] argues that IR data poorly comply with assumptions underling linear models. Figure 1 further highlights such limitation – Figure 1a shows how data should distribute in order to fully exploit the potential of linear models, while 1b exhibits how data form Robust 04 [30] actually behave: the qualitative difference in terms of shape and distributions is evident. The GLM framework can help address these limitations of the current evaluation approach. In particular, GLMs allow using different distributions for the response other than the normal one. Furthermore, they allow for non-linear relations between the response and the experimental conditions. We recently[1] investigated how GLMs can be used in the IR. In particular, we observe how selecting the appropriate link in a GLM – the function that maps the explanatory variables to the response – helps increase the model's fitting to the data, achieving better performance in terms of discriminative capabilities of the models. Therefore, we plan to explore further the GLMs domain, focusing on the distribution considered, to improve the fitting of our statistical tools to the IR data.

Advantages GLM techniques by relaxing the assumptions of linearity allow for increasing the fitness of our evaluation models to the data.

Challenges Empirical IR data is subject to several constraints (e.g., often IR measures are defined between 0 and 1) – it is thus necessary to identify the proper modelling approach to apply GLMs fully.

4.3. Distributional Representation of the User Satisfaction

Performance estimations in IR are point observations: given a system, a set of documents, and a query, we can compute a single value of the most common IR measures, such as mAP or nDCG. This can appear as a limitation for correctly assessing the performance of a system. We can imagine that multiple users will experience different degrees of satisfaction. A line of work [31, 32, 33, 34] explore how to model the system's performance as a distribution rather than a single value for each topic. We plan to follow a similar line of thought: the performance predictor that we plan to develop should model the user's satisfaction through a probability

[1]The work is currently under review process

distribution. This would partially overcome some of the limitations of the current state-of-the-art in the QPP field, where we can obtain a single prediction for each pair query system. Note that this reasoning can be easily extended to predicting a single data point - as currently done. We can force the resulting distribution to have specific properties. For example, we can impose expectation over the probability distribution to correspond to the measure we aim to predict.

Advantages Distributions adhere better than point estimations to the real world. Multiple users will perceive the system quality differently. Similarly, the system will perform differently on different data collections. Being able to model such variability is a desirable feature.

Challenges One of the primary purposes of the evaluation is to infer when a system is "better" than another. This is easily done with performance distributions over the topics by comparing the means of the distributions. When we consider multiple distributions for each system, the definition of "better" become much looser. It is necessary to find a way to combine multiple distributions to compare pairs of systems. In this sense, a possible approach is meta-analysis [32]. Furthermore, we artificially inflate our sample size by using distributions instead of point estimations. This means that if we use directly such distributions as performance estimations, we underestimate the standard error of the original performance distribution. How to correctly exploit the information derived from the distributional representation of the performance without wrongfully deflating the standard error is still an open issue.

4.4. Putting the Pieces Together

We are now in the position of describing the envisioned holistic predictive framework for the performance of the IR systems. Figure 2 illustrates the envisioned system. Labelled elements are those investigated in our research. As described in Subsection 3.1 we have understood that such a framework needs to consider multiple signals. Ideally, it should include both aspects related to multiple collections and data sources but also multiple formulations of the same topic (*O1* in Figure 2). Limiting ourselves to a single collection would produce poorly generalizable results. Considering only a single formulation for each information need would represent poorly the real world, where users typically express information needs in multiple ways.

As detailed in Subsections 3.2 and 3.3, the measurement of the performance and their prediction cannot be untied from a thorough statistical analysis. In this sense, we plan to include a meta-evaluation component that allows assessing how the predictive model is working (*O2* and *O3* in Figure 2). Subsections 4.2 and 4.1 have shown which tools and framework we want to adopt to develop the prediction engine. Subsection 4.1 has described the role that the CI can play in the envisioned framework. Through Structural Causal Modeling, we plan to develop a formal representation of our variables and factors. Using the *do*-calculus and counterfactual reasoning, we plan to predict the effect that each component of the system has on the explained variable (*O4* in Figure 2). Finally, Subsection 4.2 has detailed the fact that currently used linear models do not have the required complexity to explain the IR data. More in detail, we plan to embed aspects related to the non-linearity of the response variable in our models (*O5* in Figure 2). This should overcome the limitations of linear models and allow multiplicative modelling of the performance. We observed in Subsection 4.3 how the predictive model should not be limited to

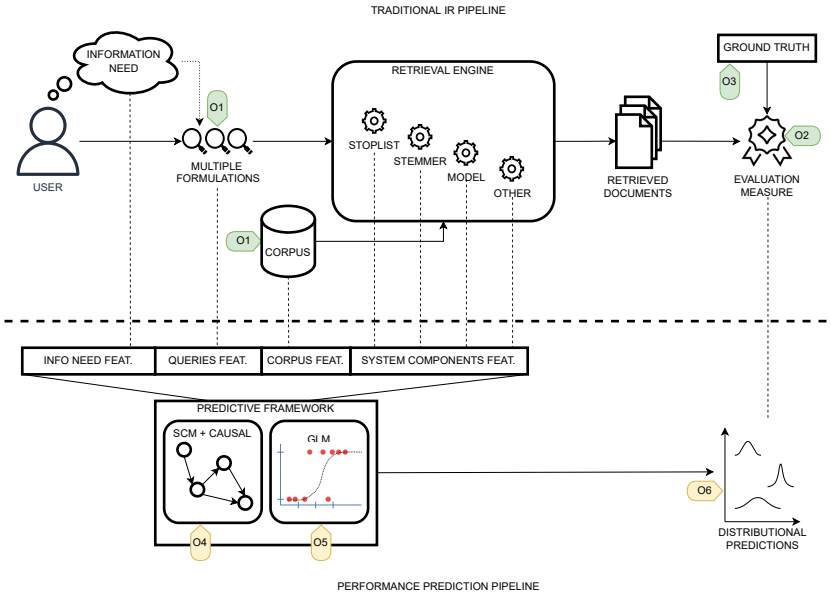


Figure 2: Architecture of the envisioned predictive framework. Labels indicate relevant aspects for this study. In green, elements already considered in previous works. In yellow, components that will be addressed in future works.

a single outcome for each pair system-topic. However, it should be able to model the complexity of reality through probability distributions. Predicting a single point estimation, such as the AP, in this framework would correspond to computing the expectation over the probability distribution (*O6* in Figure 2).

5. Conclusion and Future Work

IR evaluation is, at the current time, a strictly experimental discipline. It is not possible to predict the behaviour of a system prior to its deploying. Moreover, results are linked solely to the evaluation collections considered. With this work, we propose to overcome the current limitations in IR through a predictive framework capable of predicting, and thus generalizing, the system performance. Such a framework needs to rely on multiple signals and features. For example, we plan to consider multiple formulations of topics and several collections to fit our models. Additionally, the proposed framework should include components that go beyond the traditional linear modelling of the performance. We intend to achieve such property through GLM and Causal Inference. We plan to exploit distributional representations of the performance to better model the real world users experience. Future works will investigate how to adapt the above-mentioned methods to the IR setting. We will study the possibility of applying such approaches to performance prediction both the in the offline and online scenarios.

References

1. J. Zobel, How reliable are the results of large-scale information retrieval experiments?, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 307–314.
2. A. Rutherford, Introducing ANOVA and ANCOVA: a GLM approach, Sage, 2001.
3. E. M. Voorhees, D. Samarov, I. Soboroff, Using Replicates in Information Retrieval Evaluation, ACM Trans. Inf. Sys. 36 (2017) 12:1–12:21.
4. N. Ferro, G. Silvello, Toward an anatomy of IR system component performances, jasist 69 (2018) 187–200.
5. B. A. Carterette, Multiple testing in statistical analysis of systems-based information retrieval experiments, ACM Transactions on Information Systems (TOIS) 30 (2012) 1–34.
6. D. Banks, P. Over, N.-F. Zhang, Blind men and elephants: Six approaches to trec data, Inf. Retr. 1 (1999) 7–34.
7. N. Ferro, M. Sanderson, Improving the Accuracy of System Performance Estimation by Using Shards, in: Proc. SIGIR, 2019, pp. 805–814.
8. J. W. Tukey, Comparing individual means in the analysis of variance, Biometrics (1949) 99–114.
9. Y. Zhao, F. Scholer, Y. Tsegay, Effective pre-retrieval query performance prediction using similarity and variability evidence, in: European conference on information retrieval, Springer, 2008, pp. 52–64.
10. S. Mizzaro, J. Mothe, K. Roitero, M. Z. Ullah, Query performance prediction and effectiveness evaluation without relevance judgments: Two sides of the same coin, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1233–1236.
11. Y. Tao, S. Wu, Query performance prediction by considering score magnitude and variance together, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014, pp. 1891–1894.
12. S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting query performance, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 299–306.
13. Y. Zhou, W. B. Croft, Ranking robustness: a novel framework to predict query performance, in: Proceedings of the 15th ACM international conference on Information and knowledge management, 2006, pp. 567–574.
14. F. Scholer, S. Garcia, A case for improved evaluation of query difficulty prediction, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 640–641.
15. C. Hauff, L. Azzopardi, D. Hiemstra, The combination and evaluation of query performance prediction methods, in: European Conference on Information Retrieval, Springer, 2009, pp. 301–312.
16. H. Scells, L. Azzopardi, G. Zuccon, B. Koopman, Query variation performance prediction for systematic reviews, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2018, p. 1089–1092.

17. G. M. Di Nunzio, G. Faggioli, A study of a gain based approach for query aspects in recall oriented tasks, Applied Sciences 11 (2021) 9075.
18. J. S. Culpepper, G. Faggioli, N. Ferro, O. Kurland, Topic difficulty: Collection and query formulation effects, ACM Trans. Inf. Syst. 40 (2021). URL: https://doi.org/10.1145/3470563. doi:10.1145/3470563.
19. N. Ferro, D. Harman, CLEF 2009: Grid@CLEF Pilot Track Overview, in: Proc. CLEF, 2010, pp. 552–565.
20. G. Faggioli, S. Marchesin, What makes a query semantically hard?, in: DESIRES, 2021.
21. G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, S. F., An enhanced evaluation framework for query performance prediction, in: Proc. 43nd European Conference on IR Research (ECIR 2021), 2021.
22. C. Spearman, Footrule for measuring correlation, British Journal of Psychology 2 (1906) 89.
23. G. Faggioli, N. Ferro, System effect estimation by sharding: A comparison between anova approaches to detect significant differences, in: Proc. 43nd European Conference on IR Research (ECIR 2021), 2021.
24. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal statistical society: series B (Methodological) 57 (1995) 289–300.
25. J. Pearl, et al., Causal inference in statistics: An overview, Statistics surveys 3 (2009) 96–146.
26. D. Charles, M. Chickering, P. Simard, Counterfactual reasoning and learning systems: The example of computational advertising, Journal of Machine Learning Research 14 (2013).
27. T. Sakai, Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, 2016, p. 5–14.
28. B. Carterette, But Is It Statistically Significant? Statistical Significance in IR Research, 1995-2014, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, 2017, p. 1125–1128.
29. J. M. Tague-Sutcliffe, J. Blustein, A Statistical Analysis of the TREC-3 Data, in: Proc. TREC, 1994, pp. 385–398.
30. E. M. Voorhees, Overview of the trec 2004 robust retrieval track., in: Proc. TREC, 2004.
31. T. Sakai, S. Robertson, I. Newswatch, Modelling a user population for designing information retrieval metrics., in: EVIA@ NTCIR, 2008.
32. M. D. Smucker, C. L. Clarke, Modeling user variance in time-biased gain, in: Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, 2012, pp. 1–10.
33. G. Faggioli, M. Ferrante, N. Ferro, R. Perego, N. Tonellotto, Hierarchical dependence-aware evaluation measures for conversational search, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1935–1939.
34. G. Faggioli, M. Ferrante, N. Ferro, R. Perego, N. Tonellotto, A dependency-aware utterances permutation strategy to improve conversational evaluation, in: Advances in Information Retrieval, 2022, pp. 184–198.