Topic Difficulty: Collection and Query Formulation Effects

J. SHANE CULPEPPER, RMIT University, Australia
GUGLIELMO FAGGIOLI, University of Padua, Italy
NICOLA FERRO, University of Padua, Italy
OREN KURLAND, Technion, Israel Institute of Technology, Israel

Several recent studies have explored the interaction effects between topics, systems, corpora, and components when measuring retrieval effectiveness. However, all of these previous studies assume that a topic or information need is represented by a single query. In reality, users routinely reformulate queries to satisfy an information need. In recent years there has been renewed interest in the notion of "query variations" which are essentially multiple user formulations for an information need. Like many retrieval models, some queries are highly effective while others are not. This is often an artifact of the collection being searched which might be more or less sensitive to word choice. Users rarely have perfect knowledge about the underlying collection, and so finding queries that work is often a trial-and-error process. In this work, we explore the fundamental problem of system interaction effects between collections, ranking models, and queries. To answer this important question we formalize the analysis using ANalysis Of VAriance (ANOVA) models to measure multiple components effects across collections and topics by nesting multiple query variations within each topic. Our findings show that query formulations have a comparable effect size to the topic factor itself, which is known to be the factor with the greatest effect size in prior ANOVA studies. Both topic and formulation have a substantially larger effect size than any other factor, including the ranking algorithms and, surprisingly, even query expansion. This finding reinforces the importance of further research in understanding the role of query rewriting in IR related tasks.

CCS Concepts: • **Information systems → Evaluation of retrieval results**; **Retrieval effectiveness**.

Additional Key Words and Phrases: topic difficulty; query formulation; effect size; retrieval effectiveness

ACM Reference Format:
J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2021. Topic Difficulty: Collection and Query Formulation Effects. *ACM Transactions on Information Systems* XX, YY, Article ZZZ (February 2021), 36 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

The interplay between simple keyword queries and large document collections has challenged researchers in *Information Retrieval (IR)* for more than half a century. Document retrieval ranking models are now both complex and highly effective. However, poorly performing queries, sometimes referred to as *tail queries*, continue to surprise and challenge industrial and academic researchers. Some queries are highly effective, while others perform poorly, and changing the ranking models to compensate for *difficult* queries can have negative effects on the performance of queries that were performing well previously. This notion of *query difficulty* has received a great deal of attention

Authors' addresses: J. Shane Culpepper, RMIT University, Melbourne, Australia, shane.culpepper@rmit.edu.au; Guglielmo Faggioli, University of Padua, Padua, Italy, guglielmo.faggioli@phd.unipd.it; Nicola Ferro, University of Padua, Padua, Italy, ferro@dei.unipd.it; Oren Kurland, Technion, Israel Institute of Technology, Haifa, Israel, kurland@ie.technion.ac.il.

1046-8188/2021/2-ARTZZZ $15.00
https://doi.org/0000001.0000001

over the years. For example, NIST ran the Robust Track in 2004 and 2005 to reexamine sets of queries which had performed poorly across all systems evaluated in the Ad hoc track [73, 74]. It is clear that certain queries challenge even the best performing systems.

A series of recent papers have begun exploring the relationship between query diversity and information needs [6, 7]. In experimental settings where an information need is clearly defined, a comprehensive analysis of query formulation is possible. While the idea that information needs can and should be expressed differently is not a new idea, this important caveat can be lost when treating every query independently [9]. The distinction between a topic (information need) and a query can have a profound impact on the effectiveness of retrieval, as well as how IR researchers typically categorize and compare system performance.

In this paper, we reexamine the idea of query difficulty from the topic perspective, where a topic can have many different query formulations, and the retrieval system and the underlying document collection can change. We explore this issue by addressing the following research questions:

RQ1 How does the formulation of an information need impact system performance *within corpora*?
RQ2 How does the formulation of an information need impact system performance *across corpora*?
RQ3 How does topic difficulty vary *across corpora* based on the formulation of an information need?

RQ1 allows us to investigate the effect size of topics and query formulations with respect to systems and their components in order to better understand what contributes to topic difficulty, the magnitude of the effect, and which system components are most affected. This research question builds on an established body of prior work which studies the interaction between topics and systems. Here we extend these approaches to include query formulations. RQ2 extends RQ1 by looking at what happens across corpora and allows us to also explore corpora-specific topic / query formulations jointly with system components. Replicates are available when queries can be used on multiple collections, allowing the interactions between query formulations and system components to be computed using ANOVA for the first time. Without the addition of multiple corpora, this interaction cannot be observed experimentally. Finally, RQ3 examines the topic difficulty across multiple corpora.

In order to address RQ1 and RQ2, we develop a set of *ANalysis Of VAriance (ANOVA)* models which allow us to break down the overall system performance into topic, query formulation, system, and corpora effects; we call this a *macro-level* model. We also break down the system effect by component and show how each of these interact with topics, query variations, and corpora; we refer to this level of granularity a *micro-level* model. A *Grid of Points (GoP)*, i.e. a set of systems induced using all the combinations of targeted components – stop lists, stemmers, IR models, and *Query Expansion (QE)* in our case, is used as the data input into our models.

The *across corpora* ANOVA models also allow us to partially answer RQ3 as they enable us to quantify changes in topic difficulty using multiple corpora. In order to investigate this research question more deeply, we measure variance in arbitrarily ranked topics across corpora. The key idea is that the likelihood of observing arbitrary rank orderings of topics by effectiveness is analogous to topic difficulty being an *intrinsic* property. That is, high volatility in topic ordering suggests that topic hardness is not absolute. Rather it is an artifact of system / corpora interaction.

We conduct a thorough experimental investigation using three related *Text REtrieval Conference (TREC)* collections – Robust 2004, CORE 2017, and CORE 2018 – and a GoP consisting of 224 IR systems, each ran on all three collections, producing a total of 672 runs. A total of 2 stop lists, 4 stemmers, 7 IR models, and 4 QE techniques were used for our GoP.

Overall, we demonstrate and emphasize the fundamental difference between the notions of query difficulty and topic difficulty. More specifically, the idea of a single query being difficult is an artifact

of collection design, but it appears that topic difficulty can reliably be circumvented through careful query reformulation. This is a promising step in a fundamentally important problem in IR — that of robust system effectiveness.

The paper is organized as follows: Section 2 introduces related work; Section 4 describes our approach to answer the above research questions; Section 5 discusses the experimental setup and the experimental findings; finally, Section 6 draws some conclusions and outlooks for future work.

2 RELATED WORK

Four key bodies of prior work are most relevant to our own. These include query difficulty, query representation, query performance prediction, and system component analysis. We also provide some background on ANOVA.

2.1 Topic difficulty

Query performance, query hardness, query quality and query ambiguity all explore aspects of topic difficulty. The difficulty of a topic can be based on system performance [5, 22, 24, 56, 71, 81] or human perception [36]. Topic difficulty affects systems and users. It can also play an important role in user agreement during relevance assessment [25, 67] and *Query Performance Prediction (QPP)* [21]. It is important to emphasize that most prior work uses the terminology query difficulty and topic difficulty interchangeably, which is not problematic when a single query represents a topic. This is the most common scenario in IR, but not true in this work and other recent work on query variants. Topic difficulty is generally defined as the average effectiveness of a set of systems for a topic, more specifically the "average" average precision (AAP) originally [51]. A similar approach was also used by Carterette et al. [22] to classify topics into "easy", "medium" and "hard" for the 2009 million query track.

Topic difficulty has also proven to be an important factor in IR evaluation. For example, in Mizzaro [50] the query difficulty is used to develop a "Normalized" version of the well-known *Average Precision (AP)* measure, that penalizes and rewards systems whether they perform well or poorly on easy or hard topics. Note that as in other studies in this field, Mizzaro [50] use the concept of topic difficulty and query difficulty interchangeably. Additionally, they do not provide any generalized concept of "topic difficulty"; it is defined based on the performance achieved by the majority of the systems using a specific formulation of the topic, and on a specific corpus. A deeper analysis of the relationship between query versus topic difficulty and IR evaluation can be found in [61]. The ability to predict the difficulty of a topic provides an opportunity to adapt the system to the input query [3, 53]. Pehcevski et al. [53] develop a topic difficulty classifier that uses textual features of the topic (such as the length of the title formulation or the narrative representing the topic) to predict the topic difficulty. This classifier is then used to select the parameters used by the subsequent entity-ranking model. As in other work, topic difficulty is defined based on the AP achieved by a set of systems on a specific corpus. There are additional studies on estimating query difficulty based on linguistic features of the collection [37, 43].

2.2 Query Representation

It has been known for decades that short keyword queries often result in vocabulary mismatches, as the terms used in the query must exactly match the terms found in relevant documents in the collection being searched. The vast majority of statistical retrieval models have an implicit term independence assumption which means that, small reformulations of a query can have a significant impact on retrieval performance. Stemming and lemmatization can help alleviate this problem for commonly used terms, but not eliminate it entirely.

The classic approach to address the vocabulary mismatch challenge is *query expansion* [18, 44, 57, 60], which is the process of finding terms in the collection that the user did not initially select, but that are presumably relevant to the information need; true relevance or pseudo-relevance feedback is the most common way to operationalize query expansion in practice [79]. This is analogous to typical user behavior where a user reformulates a query for a search engine by adding additional terms to the original query when the retrieved results are not satisfactory.

There are several early studies on how users independently express an information need as a query. These "query variations" were explored in the context of TREC retrieval experiments and were shown to be highly effective when combined through fusion [9, 10, 69]. This line of work has seen a revival of interest in recent years as more powerful retrieval models and hardware provide new opportunities to leverage multiple sources of information simultaneously [7, 11–14]. For example, fusing the result lists retrieved in response to query variations was shown to be of much merit [7, 11]. Using multiple query variations can also improve query expansion as recently demonstrated by Lu et al. [46]. Interestingly, a recent study shows that, on average, query variations automatically selected from a query log of a commercial search engine can be as effective as human created variations [45].

2.3 Query Performance Prediction

The query performance prediction (QPP) task is defined as "estimating retrieval effectiveness without relevance judgments" [21]. In practice, all previous work on QPP focused on the task of estimating *topic* difficulty, with respect to a given fixed retrieval method, where each topic was represented using a *single* query [58, 83]. However, recent work shows that the relative prediction quality of existing predictors can significantly change when varying the queries used to represent the information needs [28, 83]. Indeed, topic difficulty and query difficulty are in essence two different concepts which should be carefully distinguished given that a topic (information need) can be represented using various queries, and retrieval effectiveness is clearly sensitive to the query terms selected by the user. We re-visit this point below.

2.4 System Component Analysis

A body of related work has explored component analysis to measure the effects of complex systems on retrieval effectiveness, including factors such as topic composition, collection, and system components.

Tague-Sutcliffe and Blustein [72] adopted a two-way ANOVA model to decompose the overall performance into a topic and system effect; they also hypothesized that the topic*system interaction should be an important factor, but were unable to estimate the effect size as they did not have a sufficient number of replicates available in their experiments. Banks et al. [8] provided an overview of methods available to analyze the performance of IR systems, and reexamined the model of Tague-Sutcliffe and Blustein in order to compare and contrast each of the methods. Banks et al. also performed an indirect estimation of the size of the topic*system interaction effect, providing further evidence that it has a large effect size.

Bodoff and Li [15] used multiple relevance judgments to induce the replicates needed to estimate the topic*system interaction effect. They relied on *Generalizability Theory* [16, 68], a generalization of ANOVA, to estimate the topic*system, topic*assessor, and assessor*system interaction with the goal of improving model accuracy when estimating differences between systems. Similarly, Robertson and Kanoulas [59] used simulated data to show that an ANOVA model including the topic*system interaction is better equipped to detect significant differences between systems, and reaffirm that topic*system interactions consistently have large effect sizes.

Bailey et al. [6] presented an ANOVA model comprised of topic, system, and query variation factors and found that query variations also have a large effect size and can be even larger than topic*system effect sizes. However, query variations and topics were treated as separate factors, which may be at odds with independence assumptions made by the standard ANOVA model.

Ferro and Silvello [34, 35] decomposed system factors into component effects and their respective interactions using a *Grid of Points (GoP)* [30], which is the set of all system configurations possible when permuting every system component combination being targeted. Their work found that stop lists have a medium-size effect, stemmers have a small-size effect and IR models have a small to medium-size effect. Among all of the possible interactions, only the stop list*IR model interaction was found to be significant, with a medium effect size. Ferro and Silvello used Terrier[1] to generate the GoP, and more recently Ferro [29] used Lucene[2] to generate a GoP using (almost) identical components as Ferro and Silvello in order to perform similar experiments. Ferro found that the stemmer has a large effect size while the stop list has a small effect size, and only the stemmer*IR model interaction was found to be significant, with a small effect size. This suggests that the subtle implementation differences between the two systems may influence findings in empirical studies that measure system component effects.

Sanderson et al. [65] and Jones et al. [39] studied how sub-corpora or *shards* of a given collection impact IR effectiveness and how collection size and the choice of documents influenced the way that evaluation exercises using a single test collection might influence comparisons between retrieval systems. Both of these studies emphasized the impact of sub-corpora/shards on system performance but they did not rely on a comprehensive ANOVA model that integrated all of the possible factors together. Ferro and Sanderson [32] developed such an ANOVA model and found that the shard factor has a medium-size effect but the shard*system interaction was not significant.

In similar work, Voorhees et al. [77] used sharding to produce the replicates necessary for estimating topic*system interaction effect sizes. Ferro et al. [31] developed the first exhaustive model of topics, systems, shards, as well as all their interactions, and ran extensive experiments using several sharding schemes, including a selection of randomized and deterministic methods. They found that the topic factor is a large-size effect, the system factor is a small to medium-size effect, the shard factor is a medium to large-size effect (roughly half of the topic factor), the topic*system interaction is a large-size effect (roughly one-third of the topic factor), the system*shard interaction is a small-size effect and the topic*shard interaction is a large-size effect – and may be as large as the topic effect in certain scenarios. Ferro and Sanderson [33] went on to provide a formal demonstration of why topic*shard interactions are crucial when determining which systems are significantly different from others. Faggioli and Ferro [27] compare and analyze how various ANOVA approaches behave under different conditions. Finally, Zampieri et al. [82] used ANOVA to model topics, systems and collections (and not sub-corpora or shards as in previous work) and found results consistent with those mentioned above.

3 BACKGROUND ON ANOVA

A *General Linear Mixed Model (GLMM)* [47, 63] models variations of a dependent variable ("Data") w.r.t. a controlled variation of independent variables ("Model"), in addition to a residual uncontrolled variation ("Error"): $Data = Model + Error$. The most basic example of GLMM is a simple linear regression, where $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. The dependent variable Y_i, representing the score of the i-th subject, is explained (predicted) in terms of an intercept β_0 and an independent variable X_i

[1] http://terrier.org/
[2] https://lucene.apache.org/

(predictor) times the regression coefficient β_1, the slope of the regression line, plus a residual error ε_i, not explained by the model, which follows a zero-mean Gaussian distribution.

ANalysis Of VAriance (ANOVA) when viewed as a *General Linear Mixed Model (GLMM)* attempts to explain data (the dependent variable scores) in terms of the experimental conditions (the model) and an error component. Typically, ANOVA is used to determine under which experimental condition dependent variable score means differ and what proportion of the variance observed for a dependent variable can be attributed to differences between specific experimental groups or conditions, as defined by the independent variable(s) being modeled. An ANOVA can be regarded as a type of regression analysis using only categorical predictors.

The regression model described above is expressed in ANOVA terms as $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, where Y_{ij} is the i-th dependent variable subject score in the j-th experimental condition, the parameter μ is the grand mean of the experimental condition population means underlying all of the dependent variable scores of the subjects, the parameter α_j is the effect of the j-th experimental condition and the random variable ε_{ij} is the error, which reflects any variance caused by an undefined source. The above regression model corresponds to the ANOVA version once you add as many X_{ij} predictors and as many levels as there are in the experimental condition α_j.

For a given model, the ANOVA table summarizes the outcomes of the ANOVA test indicating, for each factor, the *Sum of Squares (SS)*, the *Degrees of Freedom (DF)*, the *Mean Squares (MS)*, the F statistics, and the p-value of that factor, which allows us to determine the significance of that factor. For a detailed description on how to estimate GLMM model parameters and assess their statistical significance using ANOVA, refer to Maxwell and Delaney [47], Rutherford [63].

Independent variables can be *fixed effects* – i.e., they have precisely defined levels, and inferences about its effect apply only to those levels or *random effects* – i.e., they describe a random and independently drawn set of levels that represent variation for a clearly defined wider population. The latter case is a more sophisticated model which, in the estimation of the variance attributed to the different factors, also accounts for the additional randomness due to sampling of effect levels. The experimental design determines how to compute the model and how to estimate the parameters. In particular, it is possible to have an *independent measures* design where different subjects participate under different experimental conditions (factors) or a *repeated measures* design, where each subject participates in all of the experimental conditions (factors). A final distinction is between *crossed/factorial* designs, where every level of one factor is measured in combination with every level of the other factors, and *nested* designs, where levels of a factor are grouped within each level of another nesting factor.

3.1 Effect Size

Since the F statistic tends to increase and the p-value tends to decrease as the sample size increases, the *effect size* of a factor is used, and quantifies the magnitude of the variance observed in the model using an unbiased estimator [52, 64]:

$$\hat{\omega}^2_{\langle fact \rangle} = \frac{\mathrm{df}_{fact}(\mathrm{F}_{fact} - 1)}{\mathrm{df}_{fact}(\mathrm{F}_{fact} - 1) + N} \tag{1}$$

where F_{fact} is the F-statistic and df_{fact} are the degrees of freedom for the factor while N is the total number of samples. In this way, we are able to assess not only if a factor is significant but its contribution as well. The common rule of thumb [63] when classifying $\hat{\omega}^2_{\langle fact \rangle}$ effect size is: 0.14 and above is a *large-size effect*, 0.06–0.14 is a *medium-size effect*, and 0.01–0.06 is a *small-size effect*. Note, $\hat{\omega}^2_{\langle fact \rangle}$ can be negative; in such cases it has no contribution.

3.2 Assumptions

ANOVA is based on the following assumptions [42]:

- Normality of the error terms;
- Equal variance (homoskedasticity) of the error terms;
- Independence of the error terms, i.e. they are a random sample.

ANOVA is known to be quite robust to violations of the first two assumptions. Ito [38, p. 205] observes that the F-test is remarkably insensitive to non-normality. In commonly occurring cases where the group sample sizes are equal, it is insensitive to the heterogeneity of variance across groups. Similarly, Mendenhall and Sincich [49] note that, for relatively large samples (e.g. 20 or more observations per factor) ANOVA is robust to violations of the normality assumption and that it is also robust to differences in variances when using a balanced design. On the other hand, violation of the third assumption can severely impact the F-test and any subsequent conclusions. This issue is discussed in detail, for example, by Scariano and Davenport [66].

IR performance scores are known to violate the first two ANOVA assumptions [23, 72]. Tague-Sutcliffe and Blustein [72] noted that performance scores did not satisfy the homoskedasticity assumption and applied a transformation, which is typically used in the case of ratio data, consisting of taking the arcsine of the square root of the original scores. However, they only observed small differences in the analysis conducted on the transformed data and ultimately decided to continue using the untransformed scores, which are easier to interpret. Carterette [23] observed that both of the first two assumptions are commonly violated and that performance scores are typically bounded between [0,1]. However, Carterette concluded that ANOVA is nevertheless resilient to the kind of violations of normality encountered in IR performance scores and that also the violations of homoskedasticity have a limited impact, which supports the previous findings of Tague-Sutcliffe and Blustein [72].

In our case, our model uses topics, queries, component-wise system configurations, and multiple corpora which can be easily combined to induce the necessary sample size required to ensure that the model is robust w.r.t. violations of the normality assumption. We have also adopted a balanced design where group sample sizes are equal (discussed further in the next section), limiting the impact of any violations of homoskedasticity. Finally, when considering any possible violations of the independence assumption in topics, query variations can be regarded as independent samples from a universe of possible queries representing an information need as the queries were gathered independently using several hundred test subjects. We discuss the relationship between queries and topics and show how it can be more reliably modeled in the next section.

4 APPROACH

4.1 ANOVA Analysis

In this work we investigate factor effects – namely topics, query variations, systems and their components, and corpora – as well as the respective interactions between each. We adopt a *repeated cross-measure* design since each subject – topics in our case – is tested for every experimental condition combination – systems, their components, and corpora. In addition, we also treat query formulations as a *nested factor* within topics, since each query formulation corresponds only to a specific information need and are not shared across topics.

As many different factors are under consideration, we consider two different ANOVA models which are the *single-corpus* model and the *across-corpora* model, where the former is applied on a single corpus of documents and used to address RQ1, while the latter is applied to multiple corpora in order to address RQ2 and RQ3. For each of these, we distinguish between a *macro-level ANOVA* model, which groups related factors, i.e. topics and query variations, system component

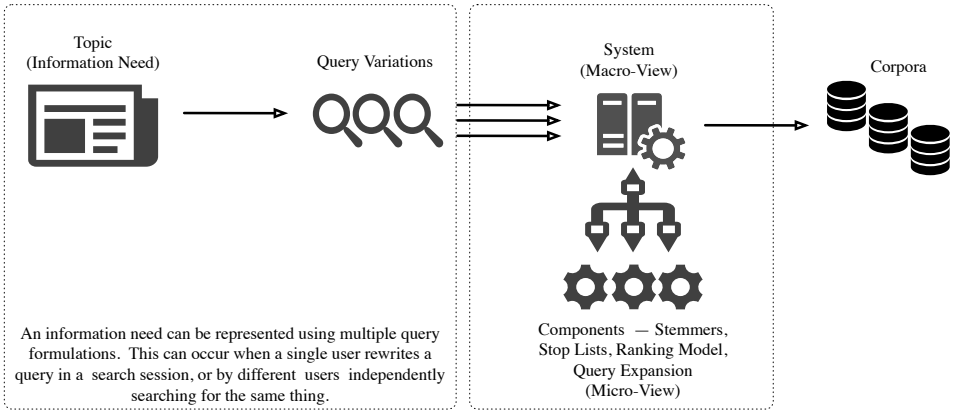
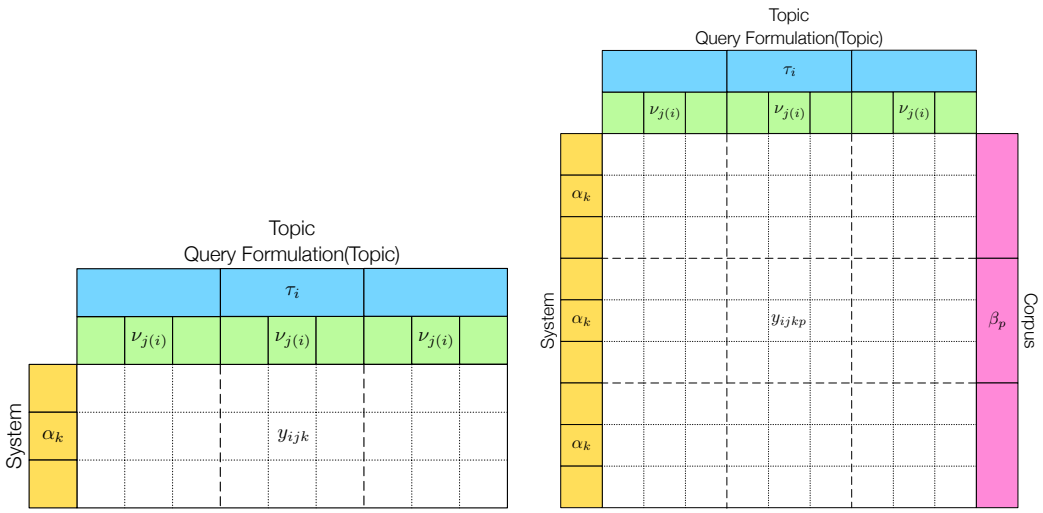


Fig. 1. Factor groupings used by the ANOVA models proposed in this work. The two high-level groupings are macro-level and micro-level. The key distinction between the two groupings is that for the micro-level analysis, each system is decomposed into all possible combinations of component factors such as a stemmer, stop list, ranker, and query expansion model.



(a) Single-corpus model (MD1$_{ma}$). (b) Across-corpora model (MD2$_{ma}$).

Fig. 2. Macro-level ANOVA model design.

configurations, and corpora, and a *micro-level ANOVA* model, which is a full break-down of all system factors into their respective contributions – stop list, stemmer, IR model, and query expansion. Figure 1 shows the relationship between the macro-level and micro-level models used in this work.

```
<num> Number: 656
<title> lead poisoning children
<desc>
How are young children being protected against lead poisoning from
paint and water pipes?
<narr>
Documents describing the extent of the problem, including suits
against manufacturers and product recalls, are relevant. Descriptions
of future plans for lead poisoning abatement projects are also
relevant. Worker problems with lead are not relevant. Other poison
hazards for children are not relevant.
```

Fig. 3. Topic 656 as defined in the TREC 2004 Robust test collection.

RQ1: Single-corpus ANOVA

The macro-level ANOVA model shown in Figure 2a is used to investigate RQ1 for a single corpus:

$$y_{ijk} = \mu + \tau_i + \nu_{j(i)} + \alpha_k + (\tau\alpha)_{ik} + \varepsilon_{ijk} \qquad (\text{MD1}_{ma})$$

where: y_{ijk} is the score of the i-th topic and j-th query formulation for the k-th system; μ is the *grand mean*; τ_i is the effect of the i-th topic; $\nu_{j(i)}$ is the effect of the $j(i)$-th query formulation; α_k is the effect of the k-th system; $(\tau\alpha)_{ik}$ is the interaction between topics and systems; finally, ε_{ijk} is the error margin for the model in predicting y_{ijk}. Note that the query formulation factor $\nu_{j(i)}$ is nested within the i-th topic since query formulations are specific to each topic. A nested factor conceptually means that each query formulation can only exist as a "subcomponent" of a topic, which is the formal description of the information that a searcher intends to retrieve. For example, Figure 3 shows the full description of topic 656 from the TREC 2004 Robust test collection as described by Voorhees [73]. Note that NIST by default often provides a query for the topic (the title) and this is included as one of the many possible query formulations which are nested as a factor of this topic. Since every query formulation for this topic is *not* independent, it cannot be treated as a separate factor as ANOVA factors by default are always assumed to be independent.

So for example, the j-th formulation "child protection laws for lead poisoning" and "lead poisoning children" are both a possible query formulation for Topic 656, the latter being title formulation provided by NIST. Considering them as nested factors allows to correctly model the variance since each formulation contributes only to the variance of a single topic. Nesting does not compare the j-th formulation of a specific topic against the j-th formulation of another one, which is analogous to the effect captured between two topics. This has the added benefit of reducing computational costs while still modeling the topic effect itself.

Model (MD1_{ma}) extends the "classical" two-way ANOVA models of Banks et al. [8], Tague-Sutcliffe and Blustein [72] to study topic and system factors, as our initial goal is to add a query variation factor. Our adaptation also extends the model of Bailey et al. [6] so that we are able observe and measure the topic*system interactions that are produced when using query variants as replicates for a topic, which is possible when query formulation are nested factors of topics.

As discussed previously, to decompose the component-wise contribution of the system factor α_k, we must apply a GoP. The following micro-level ANOVA model addresses RQ1 by breaking down the component factors for a single corpus:

$$y_{ijqrst} = \mu + \tau_i + \nu_{j(i)} + \gamma_q + \delta_r + \zeta_s + \kappa_t + (\tau\gamma)_{iq} + (\tau\delta)_{ir} + (\tau\zeta)_{is} + (\tau\kappa)_{it} + \varepsilon_{ijqrst} \qquad (\text{MD1}_{mi})$$

where, with respect to model ($MD1_{ma}$), the system factor α_k is replaced by its component-wise decomposition: γ_q is the effect of the q-th stop list; δ_r is the effect of the r-th stemmer; ζ_s is the effect of the s-th IR model; κ_t is the effect of the t-th query expansion; $(\tau\gamma)_{iq}$ is the interaction between topics and stop lists; $(\tau\delta)_{ir}$ is the interaction between topics and stemmers; $(\tau\zeta)_{is}$ is the interaction between topics and IR models; $(\tau\kappa)_{it}$ is the interaction between topics and query expansion.

Model ($MD1_{mi}$) extends the model proposed by Ferro and Silvello [34, 35] to decompose the component effects. The resulting model has a nested query formulation factor as well as capturing traditional interactions between topics and components. The model also supports query expansion models as a factor, which was not explored previously by Ferro and Silvello.

RQ2 and RQ3: Across-corpora ANOVA

A macro-level ANOVA model as shown in Figure 2b, is used to investigate RQ2 and RQ3 on multiple corpora:

$$y_{ijkp} = \mu + \tau_i + \nu_{j(i)} + \alpha_k + \beta_p + (\tau\alpha)_{ik} + (\tau\beta)_{ip} + (\alpha\nu)_{kj(i)} + (\alpha\beta)_{kp} + (\beta\nu)_{pj(i)} + (\tau\alpha\beta_p)_{ikp} + \varepsilon_{ijkp} \quad (MD2_{ma})$$

where, with respect to model ($MD1_{ma}$), it adds: β_p is the effect of the p-th corpus; $(\tau\beta)_{ip}$ is the interaction between topics and corpora; $(\alpha\nu)_{kj(i)}$ is the interaction between systems and query formulations; $(\alpha\beta)_{kp}$ is the interaction between systems and corpora and; $(\beta\nu)_{pj(i)}$ is the interaction between corpora and query formulations; $(\tau\alpha\beta_p)_{ikp}$ is the interaction between topics, systems, and corpora.

Model ($MD2_{ma}$) is a combination of several models that have been used recently [31–33, 77, 82], and also extends the model by Zampieri et al. [82] in order to cover all of the new interactions that are created when nesting query formulations. The models of Ferro et al. [31], Ferro and Sanderson [32, 33], Voorhees et al. [77] were also extended so that the query formulations can be included in addition to all of the resulting cross-factor interactions.

A micro-level ANOVA model is used to address RQ2 and RQ3 by breaking down the system factor component-wise:

$$\begin{aligned} y_{ijpqrst} =& \mu + \tau_i + \nu_{j(i)} + \beta_p + \gamma_q + \delta_r + \zeta_s + \kappa_t + (\tau\beta)_{ip} + (\beta\nu)_{pj(i)} + (\tau\gamma)_{iq} + (\tau\delta)_{ir} + \\ & (\tau\zeta)_{is} + (\tau\kappa)_{it} + (\beta\gamma)_{pq} + (\beta\delta)_{pr} + (\beta\zeta)_{ps} + (\beta\kappa)_{pt} + (\gamma\nu)_{qj(i)} + (\delta\nu)_{rj(i)} + \\ & (\zeta\nu)_{sj(i)} + (\kappa\nu)_{tj(i)} + (\tau\beta\gamma)_{ipq} + (\tau\beta\delta)_{ipr} + (\tau\beta\zeta)_{ips} + (\tau\beta\kappa)_{ipt} + (\beta\gamma\nu)_{pqj(i)} + \\ & (\beta\delta\nu)_{prj(i)} + (\beta\zeta\nu)_{psj(i)} + (\beta\kappa\nu)_{ptj(i)} + \varepsilon_{ijpqrst} \end{aligned} \quad (MD2_{mi})$$

where, with respect to model ($MD2_{ma}$) and ($MD1_{mi}$): $(\beta\gamma)_{pq}$ is the interaction between corpora and stop lists; $(\beta\delta)_{pr}$ is the interaction between corpora and stemmers; $(\beta\zeta)_{ps}$ is the interaction between corpora and IR models; $(\beta\kappa)_{pt}$ is the interaction between corpora and query expansion; $(\gamma\nu)_{qj(i)}$ is the interaction between stop lists and query formulations; $(\delta\nu)_{rj(i)}$ is the interaction between stemmers and query formulations; $(\zeta\nu)_{sj(i)}$ is the interaction between IR models and query formulations; $(\kappa\nu)_{tj(i)}$ is the interaction between query expansion and query formulations; $(\tau\beta\gamma)_{ipq}$ is the interaction between topics, corpora, and stop lists; $(\tau\beta\delta)_{ipr}$ is the interaction between topics, corpora, and stemmers; $(\tau\beta\zeta)_{ips}$ is the interaction between topics, corpora, and IR models; $(\tau\beta\kappa)_{ipt}$ is the interaction between topics, corpora, and query expansion; $(\beta\gamma\nu)_{pqj(i)}$ is the interaction between corpora, stop lists, and query formulations; $(\beta\delta\nu)_{prj(i)}$ is the interaction between corpora, stemmers, and query formulations; $(\beta\zeta\nu)_{psj(i)}$ is the interaction between corpora, IR models, and query formulations; $(\beta\kappa\nu)_{ptj(i)}$ is the interaction between corpora, query expansion.

Model (MD2$_{mi}$) extends the models proposed by Ferro and Silvello [34, 35], Zampieri et al. [82] to account for topics, query formulations and corpora interactions between all of the system components.

4.2 RQ3: Topic Difficulty via the Lenses of Topic Ranking

In order to gain a deeper insight into the notion of topic difficulty and further investigate RQ3, we also consider the following question: to what extent can we find a system α_k whose performance on a corpus β_p results in an arbitrarily chosen ranking of topics τ_i? If we (often) succeed in finding a system that induces the desired ranking of topics, it means that the topic itself cannot be thought of as always easy or difficult, since it can appear at any position with respect to other topics within the ranking.

That is, the key insight here is that if "topic difficulty" in a system effectiveness sense is fixed, the estimator will converge towards a stable topic-wise ordering for all system configurations. However, if the estimator diverges from a fixed ordering as more samples are examined, topic difficulty is not idempotent. The more volatile the orderings, the more likely a (system, corpus) pair can be found, and topics can arbitrarily be hard or easy.

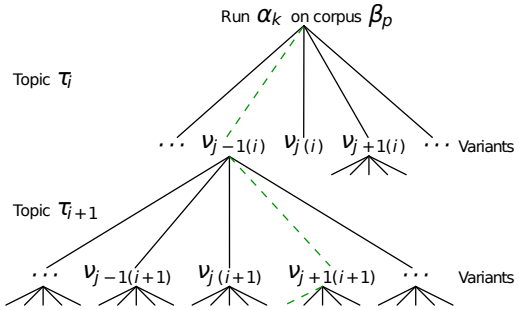
More formally, let C be the number of corpora, T the number of topics, V the number of query formulations per topic, and S be the number of systems. A random permutation $[\tau_1, \tau_2, \ldots, \tau_T]$ of topics is then selected and set as the *expected* rank ordering of the topics in the set. Then, all possible permutations within the test data are inspected to see if a match for the target rank ordering of topics exists. More specifically, for each (system, corpus) pair and all query formulations $v_{j(i)}$ available for each topic, the query formulations $[v_{j(1)}, v_{j'(2)}, \ldots, v_{j''(T)}]$ are selected when the performance of system α_k on corpus β_p induces the requested ranking of topics.

As shown in Figure 4a on the left, for each system α_k on a corpus β_p, there are V^T possible combinations of query reformulations and zero or more of them may induce the expected topic ordering. Therefore, in the worst case, a total of $C \cdot S \cdot V^T$ topic-wise rankings must be inspected in order to determine if exact or partial match solution exists. This process is then repeated for P random permutations, which is not computationally tractable in practice for even moderately sized test sets.

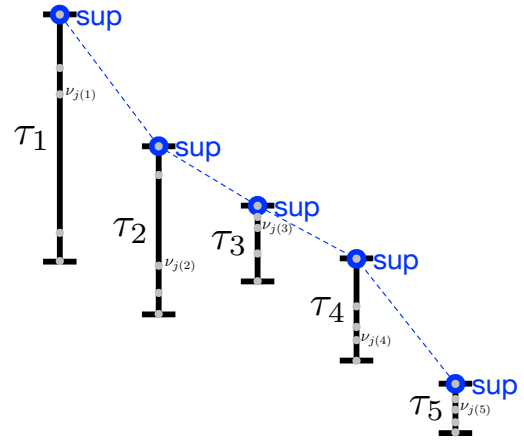
Therefore, we propose a greedy algorithm with quartic complexity $O(C \cdot S \cdot T \cdot V)$ in the worst case. The pseudo-code for our greedy algorithm is shown in Figure 5. When at least one ranking of topics exists which exactly matches the expected ordering, the algorithm is guaranteed to find a solution. When no such ranking exists, the algorithm finds a sub-optimal ranking that is "close enough" to the one requested, but a better (although not exactly matching) ranking may still exist. So in this sense, our algorithm provides a lower bound for the case of topic rankings which do not exactly match the requested one, and is suitable for our purposes since exact matches provide empirical evidence for our hypothesis that not true topic ordering can actually exist, and therefore solutions further away from exact are conservative estimates.

Figures 4b-4d demonstrate the algorithm in action. Assume that the random permutation calls for the following topic ranking: $\tau_1 \geq \tau_2 \geq \tau_3 \geq \tau_4 \geq \tau_5$. So, a (system, corpus) pair is fixed and, in Figure 4, each "bar" corresponds to a topic, where the bar represents the range of performance of the query formulations for that topic, and each gray dot on the bar is the true performance – AP in our case – of system α_k on corpus β_p for formulation $v_{j(i)}$.

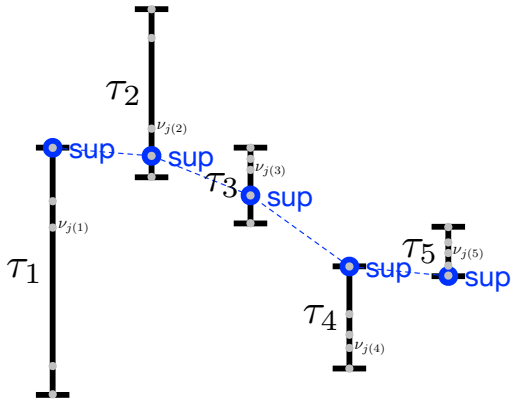
Next, the algorithm attempts to find the requested ordering by iterating over topics from this targeted ordering. The basic idea is that the maximum of a topic must be greater than or equal to the minimum of the next topic. This holds in the easiest cases, as the one shown in Figure 4b, and can even lead to topic orders which never change in certain corner cases. For example, a poor query



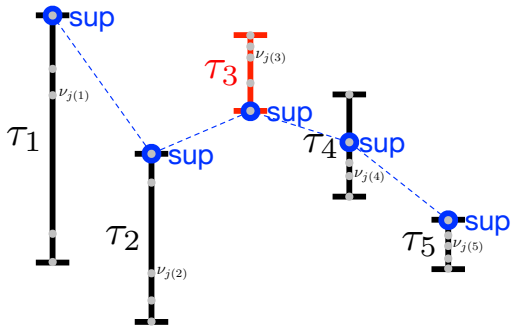
(a) Exploration space.



(b) An exact matching ranking.



(c) Another exact match ranking.



(d) A partial match ranking: all the formulations for topic τ_3 have higher AP than for topic τ_2.

Fig. 4. Ranking topics by their performance.

formulation for an "easy" topic might still perform worse than the best known query formulation for a "difficult" topic, suggesting that it is not hard to show that a topic is either easy or difficult depending on the goal. However, as we will show in the experimental section, the patterns observed in the collections available tend to be much more complex.

Moreover, the simple max-min strategy described above really just ensures a relative ordering among topics, but not an overall ordering starting from τ_1. Indeed, it is also possible for $\tau_1 \geq \tau_i$, $\tau_i \geq \tau_{i+1}$ and $\tau_{i+1} \geq \tau_1$ to occur. Therefore, in each iteration, the maximum allowed value (*sup* in Figure 4) is updated by choosing *sup* as the maximum of the performance of the formulations of the next topic which are less than or equal to the current *sup*. This choice accommodates more complex cases, such as the one shown in Figure 4c, which still induces the desired ordering.

Finally, the requested ranking *may be impossible*, as shown by the red τ_3 in Figure 4d. In this case, we must choose as new *sup* which is the minimum performance of the formulations of the non-compliant topic, and the algorithm attempts the ranking selection process again.

Fig. 5. Pseudocode for the greedy search algorithm.

Data: $\mathcal{T}$: list of topics; $\mathcal{B}$: set of corpora; $\mathcal{A}$: set of systems; $\mathcal{P}$: sample of all possible permutations of topics; $\mathcal{V} = \{\mathcal{V}_{(\tau_i)} \forall \tau_i \in \mathcal{T}\}$ where $\mathcal{V}_{(\tau)}$ is a set of query reformulations for topic τ; **AP** tensor containing AP scores for each triple (corpus, system, query);

$globalCorr \leftarrow 0$;
for $\pi \in \mathcal{P}$ **do**
  $bestCorr \leftarrow -1$;
  for $\beta \in \mathcal{B}, \alpha \in \mathcal{A}$ **do**
    $\tau \leftarrow \pi\ [1]$;
    $sup \leftarrow \text{MAX}\ (\textbf{AP}\ [\beta, \alpha, \mathcal{V}_{(\tau)}])$;
    `/* The list sortedAP contains the topic AP score mapping of the query reformulations that induce the ordering with the highest correlation` π`. */`
    $sortedAP \leftarrow [sup]$;
    for $\tau \in \pi\ [2 : end]$ **do**
      if $\exists v_{i(\tau)}$ *s.t.* **AP** $[\beta, \alpha, v_{i(\tau)}] \leq$ sup **then**
        $sup \leftarrow \text{MAX}\ (\textbf{AP}\ [\beta, \alpha, \{v_{i(\tau)} \forall v_{i(\tau)} \in \mathcal{V}_{(\tau)}$ s.t. $v_{i(\tau)} \leq sup\}])$
      else
        $sup \leftarrow \text{MIN}\ (\textbf{AP}\ [\beta, \alpha, \mathcal{V}_{(\tau)}])$;
      $sup \oplus sortedAP$;
    $bestCorr \leftarrow \text{MAX}\ (bestCorr, \text{KENDALL}\ (\pi, sortedAP))$;
  $globalCorr \oplus \frac{bestCorr}{|\mathcal{P}|}$

Since we can have both exact matches and partial matches, we compute Kendall's tau correlation [40] between the ranking of topics requested by the given permutation and the one we have found for a given (system, corpus) pair. Kendall's tau is 1 when we find an exact match and less than 1 otherwise. Finally, for each permutation, we record the maximum Kendall's tau across all of the (system, corpus) pairs to indicate the extent to which we have been able to find the request ranking of topics.

Note that we have adopted the use of Kendall's tau correlation coefficient, which weights the same a swap at any rank position, and not a more top-heavy correlation coefficient, like AP correlation [80], because our goal is to study the extent to which topics can be "arbitrarily" easy or difficult, and thus we are equally interested in swaps at any position in the ranking.

5 EXPERIMENTS

5.1 Experimental Setup

Data and Methods. We used the following collections: TREC Robust 2004 Ad Hoc [73], TREC Common CORE 2017 [1], and TREC Common CORE 2018 [2] for our experiments. The Robust Ad Hoc track used Disk 4 and 5 of the TIPSTER corpus minus the Congressional Record sub-collection and contains approximately 528K documents; the TREC 2017 Common CORE track used the New

Table 1. Summary statistics of the collections used. The column 'Shared' contains statistics on topics which overlap in all three collections.

	CORE 2017	**CORE 2018**	**Robust 2004**	**Shared**
# of documents	1,855.658	595,037	528,155	-
# of topics	50	50	250	25
total # of formulations per corpus	1286	625	3402	625
avg # of formulations per topic	25.72	25.0	13.61	25.0
min # of formulations per topic	20	20	9	20
max # of formulations per topic	53	40	53	40
avg # of words per formulation	4.8	4.7	5.3	4.7
min # of words per formulation	1	1	1	1
max # of words per formulation.	15	11	17	11

York Times Annotated Corpus which contains over 1.8 million articles; finally, the TREC Common CORE 2018 track used the Washington Post corpus, roughly containing 600K news articles.

A large seed set of human curated query formulations originally developed using the TREC Robust 2004 Ad Hoc search collection were used in our experiments [11][3]. These were further enriched with query reformulations extracted from a Bing search log and mapped to the original 249 topic descriptions [45].

In 2017 and 2018, TREC ran the CORE track which reused many of the original topics from the 2004 Robust exercise. There are 50 overlapping topics in CORE 2017 and 25 overlapping topics in 2018. To provide comparable results across multiple corpora, we use only the subset of topics which overlap in all three of the corpora. Thus, in the following experiments, we considered 625 query formulations for the 25 overlapping topics and not all of the 3,402 variations available for the 249 topics in Robust 2004. Note that one of the original Robust topics has no relevant documents in QREL judgments created by NIST, and is therefore omitted from consideration. All of the collections use graded relevance judgments, with the 3 grades being: not relevant, relevant, and highly relevant; we mapped to binary relevance judgments by using a lenient approach, i.e. everything above not relevant is considered as relevant since *Average Precision (AP)* is used for all evaluation comparisons in this work [19]. Table 1 provides additional statistical information on the collections used.

As discussed in Section 2, previous work has simulated various collection effects by splitting a collection into shards or sub-collections [31, 32, 82]. The TREC CORE tasks were ran in 2017/2018 and reused topics originally created for the TREC Ad Hoc tasks between 2002 and 2005. Thus, they allow us to compare system performance across multiple corpora using the same set of topics and identical system configurations. This allows us to compute the effect of the collection, without simulating it and ensures that the system component factors can be the same for every collection being used. The underlying collection was composed primarily of news articles in all three TREC campaigns, but changed from the original Newswire collection in 2004 to the New York Times collection in 2017 and then to the Washington Post collection in 2018.

For all experiments, we used a modified version of the Terrier Search engine (version 5.1) to create our GoP. The modifications were required in order to maximize the diversity of components (stemmers and ranking models) available for our experimental setup. Based on a few preliminary runs for multiple system configurations on each collection, we selected 9 retrieval ranking models and 3 query expansion models (plus the no query expansion), which are described in Table 2.

[3]http://culpepper.io/publications/robust-uqv.txt.gz

Table 2. Terrier Retrieval and Query Expansion Models.

Model	Description
BM25	Okapi BM25.
DPH	The parameter-free hyper-geometric divergence from randomness (DFR) model using Popper's normalization.
Hiemstra_LM	Hiemstra's language model.
In_expB2	Inverse expected document frequency model for randomness, the ratio of two Bernoulli's processes for first normalisation, and Normalisation 2 for term frequency normalisation.
Js_KLs	A weighted combination of Jeffreys divergence and Kullback Leibler divergence.
TF_IDF	The TF×IDF weighting function, using Robertson's TF and the IDF of Sparck Jones.
PL2	Poisson estimation for randomness, Laplace succession for first normalisation, and Normalisation 2 for term frequency normalisation .
TF_IDF_DRF	Same as above with a pBiL DFR term dependency model [54] enabled and a window size of 5.
Hiemstra_LM_DRF	Same as above with a pBiL DFR term dependency model [54] enabled and a window size of 5.
BA	The approximation of the binomial distribution using the Kullback-Leibler divergence to induce the weighted query terms during expansion.
Bo1	The Bose-Einstein 1 DFR expansion technique.
KL	Kullback-Leibler divergence based query expansion.

Moreover, runs were built using 4 different stemmers: Krovetz, Porter, S-Stemmer, and Lovins. Finally, we doubled the number of available runs by either keeping or removing the stop words.

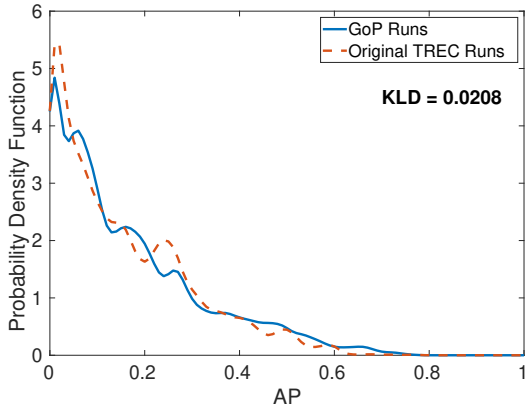
Stemmers, Retrieval Models, and Query Expansion Models have been chosen to maximize the variety in our system configurations, in terms of overall effectiveness and the documents retrieved. The total number of available configurations was 288, which have been applied on each corpus, giving us a total of 864 runs. To aid reproducibility in the future, data and runs are publicly available.[4]

5.2 Validation of the Experimental Setup

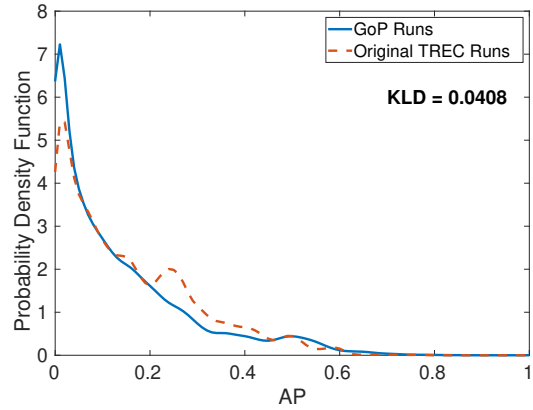
We perform a preliminary inspection of our dataset in order to verify that our GoP has a performance distribution comparable to typical runs submitted to TREC and that query formulations are not skewed or biased in any specific way.

5.2.1 Validation of the Grid of Points. We now investigate how close the performance distribution of the original systems submitted to that TREC track is to the performance distribution of the GoP systems on the same track. To quantify this "closeness" we use the *Kullback-Leibler Divergence (KLD)* [41] between the two performance distributions. In order to compute the KLD, we need the *Probability Density Function (PDF)* of the performance distributions, which we estimate by using a *Kernel Density Estimation (KDE)* [78] approach.

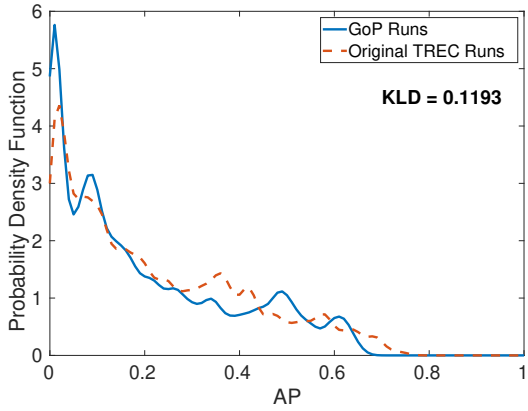
[4]https://github.com/jsc/anova-query_formulations



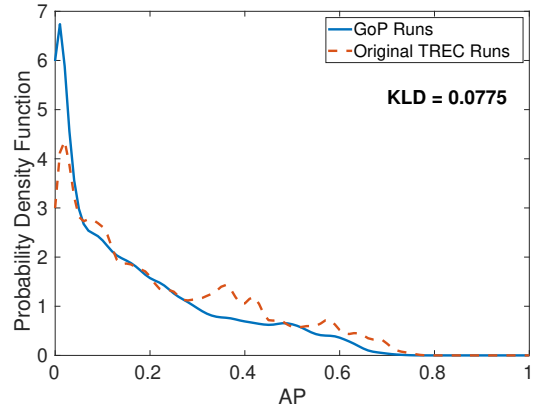
(a) Robust 2004, using only the title query formulation for GoP runs.



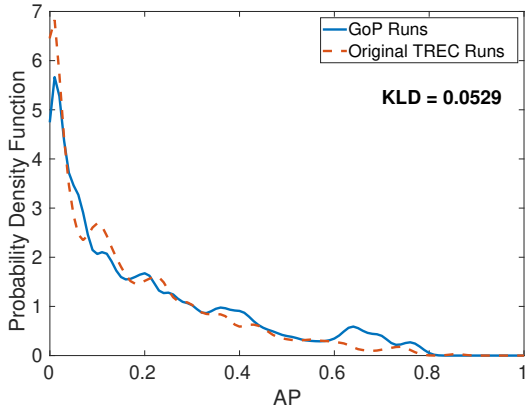
(b) Robust 2004, using all the query formulations for GoP runs.



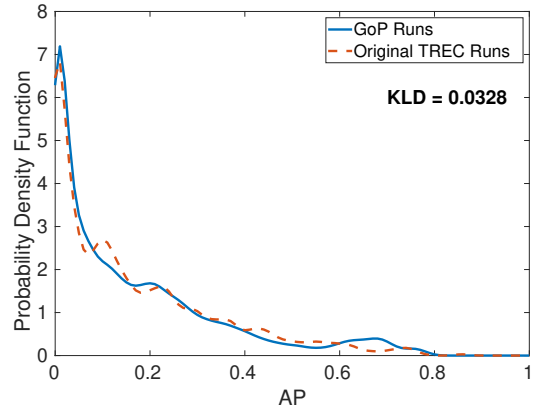
(c) CORE 2017, using only the title query formulation for GoP runs.



(d) CORE 2017, using all the query formulations for GoP runs.



(e) CORE 2018, using only the title query formulation for GoP runs.



(f) CORE 2018, using all the query formulations for GoP runs.

Fig. 6. Comparison between AP score distribution of GoP runs and the original TREC runs. A small divergence between the original scores distributions and the scores achieved using the GoP can be observed in all six plots.

Given a vector X of m elements, the KDE estimation of the PDF is given by

$$\hat{f}_X(x) = \frac{1}{mb} \sum_{i=1}^{m} K\left(\frac{x - X_i}{b}\right) \tag{2}$$

where X_i is the i-th component of the array, b is the *bandwidth* or *window width* and is greater than 0; $K(\cdot)$ is the *kernel* satisfying $\int_{-\infty}^{+\infty} K(x)dx = 1$. In this work, we use a Gaussian kernel with bandwidth $b = 0.015$.

Given two m element vectors X and Y, the KLD between the PDFs is given by

$$D_{KL}\left(X\middle\|Y\right) = \sum_{x} \ln\left(\frac{\hat{f}_X(x)}{\hat{f}_Y(x)}\right) \hat{f}_X(x) \tag{3}$$

Note that D_{KL} is not symmetric and so, in general, $D_{KL}\left(X\middle\|Y\right) \neq D_{KL}\left(Y\middle\|X\right)$.

As initially proposed by Burnham and Anderson [20], $D_{KL} \in [0, +\infty)$ denotes the information lost when Y is used to approximate X; in our context, it denotes the information lost when the GoP systems are used to "approximate" an original set of systems submitted to a TREC track. Therefore, 0 means that there is no loss of information and, in our context, that the original systems and the GoP ones are considered the same; $+\infty$ means that there is a full loss of information and, in our context, that the original systems and the GoP ones have no similarity.

Note that the TREC runs may have used the title, the description, and/or the narrative of a topic, as well as manual formulations for that topic, but generally the title is the most commonly used field. Therefore, we have chosen to compare TREC runs to our GoP runs using only the title of a topic, as shown in Figure 6 on the left. Moreover, to validate that the other query formulations do not introduce any specific bias, in Figure 6 on the right, we compare the original TREC runs with respect to our GoP using all the query formulations and verify that the distributions have a similar composition.

In all of the comparisons in Figure 6, the estimated distributions are very similar and KLD is small. This indicates that our data mimics the behavior of the original TREC runs. Moreover, in the case when all the query formulations are used, the distributions are still very close, suggesting that the query formulation did not distort the outcomes.

5.2.2 Validation of the Query Formulations. In the boxplot [48] shown in Figures 7, 8, we consider the performance of each query reformulation across all GoP systems; for each topic, there is a box corresponding to each corpus; the performance of the title query is highlighted with a diamond. In Figure 7 on the left we use, as aggregation statistics, the average of AP, also referred to as *Average Average Precision (AAP)* by Mizzaro and Robertson [51]; in Figure 8 on the right we use the median of AP as aggregation statistics.

From Figures 7, 8, we can see that there is no topic for which all the reformulations perform similarly on all corpora. Indeed, the performance distribution of the topics vary widely when using the query formulations and are even less predictable when changing the underlying corpus. However, the topics 442 and 690 do have similar distributions across all corpora, suggesting that their easiness or difficulty is more stable than the others. Overall, this provides some visual intuition of how unstable topic effectiveness is when the corpora and the query formulation varies.

In general, the query formulations appear to be high quality. More concretely, in both figures, the red diamond indicates either the AAP or the median AP achieved by the title formulation for the specific TREC topic over all systems, and we can observe that even though the title formulation is often in the top quartile of possible outcomes, there are many cases where the title formulation performs poorly compared to many of the reformulations, and may even be the worst performing one. It is also interesting to observe that, even though a query formulation might

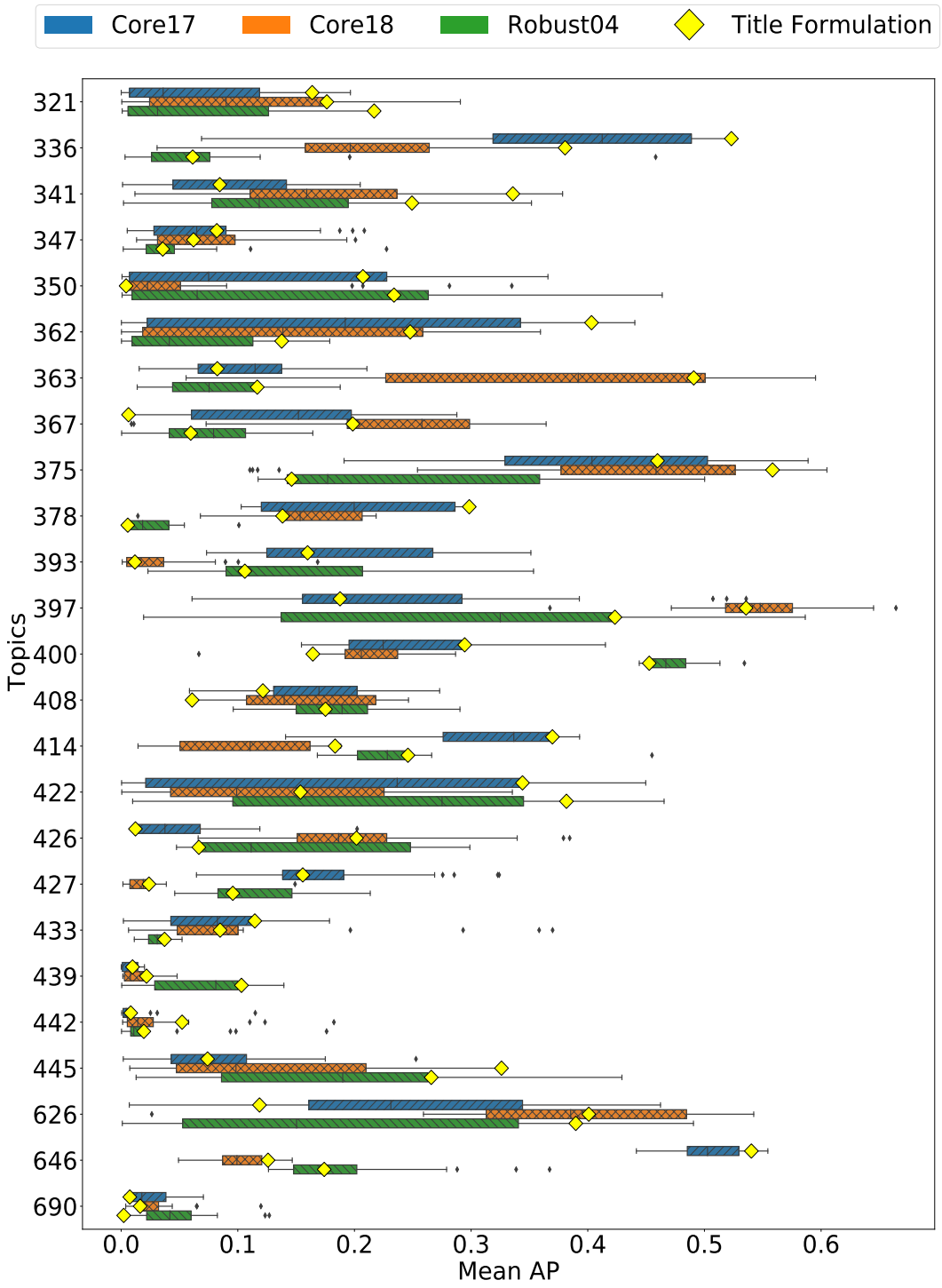


Fig. 7. Distribution of the Average AP of the different query formulations over all GoP systems. For each topic, there is a box corresponding to each corpus; the performance of the title query is denoted with a diamond.

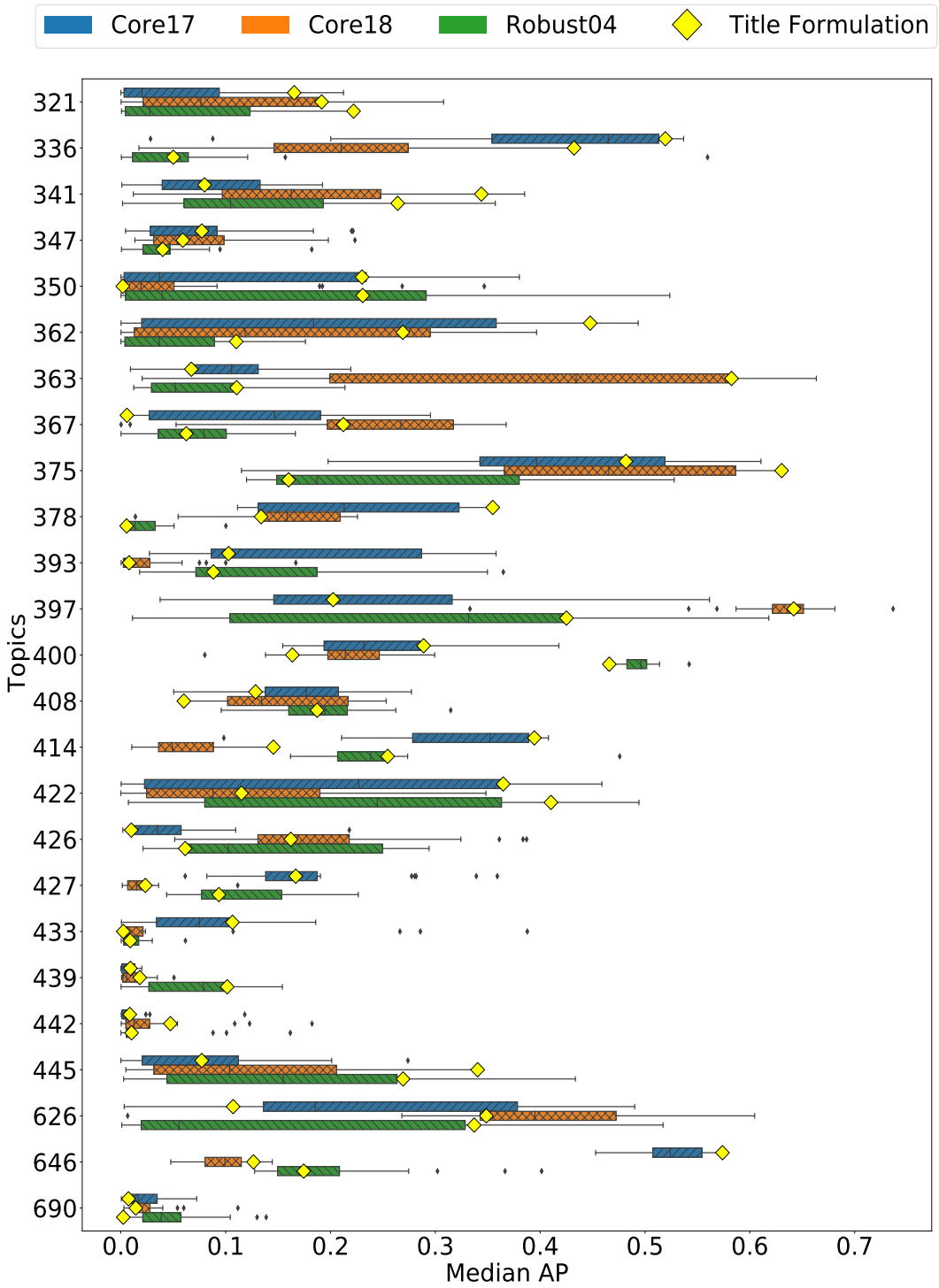


Fig. 8. Distribution of the Median AP of the different query formulations over all the GoP systems. For each topic, there is a box corresponding to each corpus; the performance of the title query is denoted with a diamond.

Table 3. A summary of the effect sizes for factors in $MD1_{ma}$ for all three collections. Blue represents the size of the factor, where dark blue is large and light blue is medium. For all three corpora, observe that the majority of the factors have a large size effect. The only medium size factor in two collections (Robust 2004 and CORE 2018) is the System. Furthermore, observe that the Topic*System interaction has a large size effect, which indicates that system configuration and topic performance are correlated, and supports the hypothesis that the "topic difficulty" is linked to the system used and not the query formulation.

	Robust 2004	CORE 2017	CORE 2018
Topic	0.7639	0.8215	0.7834
Formulations (Topic)	0.6941	0.6833	0.6038
System	0.1080	0.2193	0.1445
Topic*System	0.3385	0.3510	0.4386

perform particularly well on a one corpus, often it does not perform equally well on another one. For example, in the case of topic 378 the title formulation is the best performing formulation on the CORE 2017 corpus but it falls in the lowest quartile in the CORE 2018 corpus and is one of the worst formulations in Robust 2004. Given that all three collections are essentially news documents, it is somewhat surprising that the performance is so volatile given that the same components and ranking functions are being used. That is, the search engine is fixed, but the query and documents searched are not.

Now consider the possible outcomes of the greedy algorithm discussed in Section 4.2 and the different cases shown in Figure 4, Figure 4b, which were the easiest case, and where the minimum of a topic is below the maximum of another topic respectively. Figures 7 and 8 clearly show that this rarely happens in our experiments. Instead we typically observe the more complex patterns exhibited in Figure 4c and 4d. As a consequence, the corner case where a really poor query formulation for an "easy" topic performs more poorly than a very good query formulation for a "difficult" topic, which would correspond to a very large bar (the easy topic) whose bottom is below the top of a very narrow bar (the difficult topic), was not observed in these experiments.

5.3 RQ1: Effect Size of Query Formulation within Corpora

5.3.1 Macro-Level ANOVA. Due to the large number permutations and the memory constraints imposed by the underlying ANOVA model, we randomly sample 18 query formulations for each topic in the following analysis.

Table 3 provide a summary of the effect size for each factor, for model ($MD1_{ma}$) using each corpus, where we observe similar performance trends across all three. Table 4, Table 5, and Table 6 contains the complete ANOVA statistics for model $MD1_{ma}$ respectively on Robust 2004, CORE 2017, and CORE 2018.

All factors were found to be statistically significant. Consistently with the previous findings of Tague-Sutcliffe and Blustein [72], the topic factor has a large-size effect size, and it is indeed the largest effect for this configuration. We can also clearly see that query formulations also have a large effect size in our experiments – approaching the topic effect size – suggesting that query formulations strongly influence topic difficulty. In prior work, Bailey et al. [6] also observed that query formulation had an effect in an ANOVA analysis, but their ANOVA used a different nesting of factors than ours and was based on just two systems; this may have had an impact on their reported topic effect, which was a medium-effect size and differs from all other previous literature where topic effect consistently has a high-effect size.

Table 4. Model ($MD1_{ma}$) on track Robust 2004 for AP. The letter in the column "effect size" indicates whether the effect is large (L) or medium (M). On Robust 2004 the effect size of all factors is large, except for the System factor, which has a medium size effect. The topic (information need) and its formulations are the most prominent effect, followed by the interaction between the topic and the system. A large effect for the interaction indicates that some systems are better for specific topics while, for other systems, its the other way around.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	1389.69	24	57.90	17471.22	<1e-6	0.7639	L
Formulations (Topic)	976.17	425	2.30	693.03	<1e-6	0.6941	L
System	52.97	287	0.18	55.69	<1e-6	0.1080	M
Topic*System	242.63	6888	0.04	10.63	<1e-6	0.3385	L
Error	404.26	121975	<1e-2				
Total	3065.72	129599					

Table 5. Model ($MD1_{ma}$) on track CORE 2017 for AP. The letter in the column “effect size” indicates if the effect is large (L). For CORE 2017, all factors have a large size effect. The topic (information need) has an effect that is 3.68 times larger than the system effect. The topic formulation on the other hand has an effect much larger than the system effect. The interaction effect is again very large.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	2073.88	24	86.41	24852.00	<1e-6	0.8215	L
Formulations (Topic)	973.70	425	2.29	658.91	<1e-6	0.6833	L
System	127.59	287	0.44	127.86	<1e-6	0.2193	L
Topic*System	267.65	6888	0.04	11.18	<1e-6	0.3510	L
Error	424.11	121975	<1e-2				
Total	3866.94	129599					

Table 6. Model ($MD1_{ma}$) on track CORE 2018 for AP. The letter in the column "effect size" indicates whether the effect is large (L) or medium (M). On CORE 2018 Effects sizes are close to the one observed for the Robust collection. We have very large effects for the topic and its formulations, a medium-large effect for the system and a large effect for the interaction between the system and the topic. Compared to the Robust, we observe an even larger effect for the interaction.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	2371.78	24	98.82	19532.05	<1e-6	0.7834	L
Formulations (Topic)	1001.38	425	2.36	465.69	<1e-6	0.6038	L
System	112.21	287	0.39	77.28	<1e-6	0.1445	M
Topic*System	547.15	6888	0.08	15.70	<1e-6	0.4386	L
Error	617.14	121975	0.01				
Total	4649.67	129599					

The system factor has a medium to large-size effect and on CORE 2017 almost double the size of Robust 2004 and CORE 2018. This suggests that the interaction between system and corpus can play a role, as will be investigated in subsequent analyses.

Table 7. A summary of effect sizes for factors when using $MD1_{mi}$ on the three collections. The shade of blue indicates the factor size – large being dark blue, and medium or small as lighter shades of blue. The white cells are the factors with significant but negligible effects sizes. The stoplist factor on the Robust 2004 collection is not significant. Observe that the topic and the formulation, as well as all of the micro-components have smaller effects than a system treated as a whole. The two most prominent effects in a system are the Retrieval Model and the Query Expansion Model. It is interesting to note that the Query Expansion Model has different effect sizes that depend on the corpus. Furthermore, the majority of the interactions between the Topic and the various components have medium to large effect sizes. Stopping versus not stopping has a negligible effect, both alone and in interaction with the Topic.

	Robust 2004	CORE 2017	CORE 2018
Topic	0.7560	0.8116	0.7743
Formulation	0.6848	0.6687	0.5910
Stoplist	0.0007	0.0020	0.0013
Stemmer	0.0043	0.0024	0.0060
Model	0.0661	0.0728	0.0197
Query Expansion	0.0232	0.1377	0.1114
Topic*Stoplist	0.0073	0.0026	0.0033
Topic*Stemmer	0.0709	0.0355	0.0673
Topic*Model	0.2158	0.2356	0.1677
Topic*Query Exp.	0.3153	0.1029	0.2958

Another interesting observation in our analysis is the topic*system interaction effect size in Table 3, which is large, and confirms an important supposition of Banks et al. [8], who were only able to provide a rough estimate for the topic*system interaction. While Banks et al. [8] suspected that topic*system interactions should be large, they were not able to actually confirm it due to an insufficient number of replicates in their experimental setup. More recently, Ferro et al. [31] and Voorhees et al. [77], who used collection shards to obtain the necessary replicates required to estimate the topic*system interaction effect size, found that it does indeed have a large-size effect. In our configuration, the replicates necessary to estimate this effect are provided by different query formulations which, to the best of our knowledge, have not been used for this purpose in previous work. This further confirms the prominence of the existence of this effect when using our proposed experimental design.

Overall, query formulation has the second largest effect, with nearly 1.5 times the size of the topic*system interaction which has historically been a point of emphasis in similar performance comparisons. This provides important evidence that query formulation is crucial in retrieval effectiveness, and has deeper implications in rethinking the way many IR experiments currently formalize query / topic difficulty. Topic difficulty for query performance prediction can be viewed at an abstract level as an attempt to predict topic*system interactions, and indeed the quality of these methods are often measured using a Kendall tau of the *ordering* of topics by effectiveness for any given set of topics. This is only possible if certain topics consistently perform better than others, and having a mix of "easy", "medium", and "difficult" tends to provide the most desirable signal. But what if there were *no* difficult topics? Do such scenarios exist given our ability to reformulate queries based on a collection and the surprisingly large factor size observed? We will explore this intriguing question in more detail in Sections 5.4 and 5.5.

5.3.2 Micro-Level ANOVA. In order to better understand the impact of query formulation at the component level, we have also performed a detailed component-wise ANOVA analysis using

model ($MD1_{mi}$), whose results are reported in Table 7. Additional statistics and details we used to create the summary table are also included here in the ANOVA tables for the model $MD1_{mi}$ on Robust 2004 (Table 8), CORE 2017 (Table 9), and CORE 2018 (Table 10). See these tables for additional information.

Table 8. Model ($MD1_{mi}$) on track Robust 2004 for AP. The letter in the column "effect size" indicates whether the effect is large (L), medium (M), small (S) or not significant (-). Between parentheses significant yet almost negligible effects. Among the different factors, only the Topic and the formulations have a large effect. The Retrieval model has a medium effect while both the Stemmer and Query Expansion model have a small size effect. Even though significant, removing or not the stopwords and the stemmer have a very small effect. Although the interaction between topics and components is always significant, we observe variations on the different effect sizes. We observe a large interaction only with the Retrieval model. The interactions between the topic and either the Query Expansion model and the Stemmer have medium size effects.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	1389.69	24	57.90	16729.19	<1e-6	0.7560	L
Form. (Topic)	976.17	425	2.30	663.59	<1e-6	0.6848	L
Stoplist	0.33	1	0.33	93.92	<1e-6	0.0007	(S)
Stemmer	1.94	3	0.65	186.54	<1e-6	0.0043	(S)
Model	31.77	8	3.97	1147.20	<1e-6	0.0661	M
Query Expansion	10.78	3	3.59	1038.29	<1e-6	0.0234	S
Topic*Stoplist	3.38	24	0.14	40.65	<1e-6	0.0073	(S)
Topic*Stemmer	34.50	72	0.48	138.44	<1e-6	0.0709	M
Topic*Model	124.10	192	0.65	186.74	<1e-6	0.2158	L
topic*Query Exp.	47.35	72	0.66	189.98	<1e-6	0.0950	M
Error	445.72	128775	<1e-2				
Total	3065.72	129599					

Table 9. Model ($MD1_{mi}$) on track CORE 2017 for AP. The letter in the column "effect size" indicates whether the effect is large (L), medium (M) or small (S). Between parentheses significant yet almost negligible effects. In the case of CORE 2017, we observe that the factors with large size effects are the topic, the formulations, and the Query Expansion Model. Even though large size, the Query Expansion model has an effect that is 6 time smaller than the topic and 5 times smaller than the formulations. Both stoplists and stemmers are significant, yet have a negligible effect. The Retrieval model has a medium size effect. As previously observed, all the interactions between the topic and the components are significant.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	2073.88	24	86.41	23256.87	<1e-6	0.8116	L
Form. (Topic)	973.70	425	2.29	616.62	<1e-6	0.6687	L
Stoplist	0.97	1	0.97	261.72	<1e-6	0.0020	(S)
Stemmer	1.18	3	0.39	106.14	<1e-6	0.0024	(S)
Model	37.84	8	4.73	1273.00	<1e-6	0.0728	M
Query Expansion	76.96	3	25.65	6902.16	<1e-6	0.1377	M
Topic*Stoplist	1.34	24	0.06	15.02	<1e-6	0.0026	(S)
Topic*Stemmer	17.99	72	0.25	67.24	<1e-6	0.0355	S
Topic*Model	149.14	192	0.78	209.06	<1e-6	0.2356	L
Topic*Query Exp.	55.49	72	0.77	207.42	<1e-6	0.1029	M
Error	478.47	128775	<1e-2				
Total	3866.94	129599					

Again, the results are consistent across all the three collections. Stop lists and stemmers have a small-size effect – stop lists were not even significant on Robust 2004 – which was not true in

Table 10. Model ($MD1_{mi}$) on track CORE 2018 using AP. The letter in the column "effect size" indicates whether the effect is large (L), medium (M) or small (S). In the case of CORE 2018, the only factors with large size effects are the topic and the formulations. Both stoplists and stemmers are significant, but have a negligible effect size. The Retrieval model has a small effect on system performance, and Query Expansion has a medium size effect. As previously observed, all of the interactions between the topic and the components are significant, but the interaction between stoplists and topics have a negligible effect sizes, and interaction between the retrieval and query expansion models are large.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	2371.78	24	98.82	18524.14	<1e-6	0.7743	L
Form. (Topic)	1001.38	425	2.36	441.66	<1e-6	0.5910	L
Stoplist	0.93	1	0.93	174.16	<1e-6	0.0013	(S)
Stemmer	4.16	3	1.39	259.65	<1e-6	0.0060	(S)
Model	13.92	8	1.74	326.08	<1e-6	0.0197	S
Query Expansion	86.67	3	28.89	5415.26	<1e-6	0.1114	M
Topic*Stoplist	2.41	24	0.10	18.81	<1e-6	0.0033	(S)
Topic*Stemmer	50.25	72	0.70	130.83	<1e-6	0.0673	M
Topic*Model	140.35	192	0.73	137.02	<1e-6	0.1677	L
Topic*Query Exp.	290.82	72	4.04	757.14	<1e-6	0.2958	L
Error	687.00	128775	0.01				
Total	4649.67	129599					

experiments ran by Ferro and Silvello [34, 35], who both reported a medium-size effect for these factors. We also found, consistent with previous work, that the IR model factor has a small to medium-size effect at the micro-level. These results are aligned with Zampieri et al. [82] who also observed that stemmers and IR models have a small-size effect, albeit two orders of magnitude larger in our configuration.

Query expansion on the other hand has a small-size to medium-size effect and is the largest among all system component factors for CORE 2017 and CORE 2018, which differ from Zampieri et al. [82] who observed a very small small-size effect for this factor. Note that we have incorporated query formulations in our comparison, and the combination of query reformulations and query expansion is the most likely contributor to differences in effect sizes we have observed. We revisit this hypothesis in the next section as we will be in a better position to measure it directly in our final model configuration.

As previously discussed, we were also able to reliably estimate the interaction effect sizes between topics and system components for the first time. In particular, topic*query expansion interaction has a notably large-size effect, followed by the topic*IR model interaction. The interaction with stemmers had a small to medium-size effect while stop lists had a very small-size effect. Note that low IDF terms are dropped when query expansion is enabled as stop words tend to have a negative impact on system effectiveness when they are not removed. In summary, our findings indicate that query expansion and IR models are the components most affected by topics, and this could provide useful hints when debugging and diagnosing which system components to target in order to improve performance.

5.4 RQ2 and RQ3: Effect Size of Query Formulation across Corpora

In this section, we expand our analysis using Models ($MD2_{ma}$) and ($MD2_{mi}$) which can be used to measure cross collection effects, and address RQ2 and RQ3. Note that the computational complexity of these two models in our current configuration is substantial, and therefore the analysis was carried using only the two best performing stemmers – Porter and Krovetz based on our initial analysis. For the same reason, we also limit ourselves to 15 query formulations for each topic.

Table 11. Model ($MD2_{ma}$) for the Robust 2014, CORE 2017, and CORE 2018 tracks and AP. The letter in the column "effect size" indicates whether the effect is large (L), medium (M) or small (S). Observe that the majority of the factors have large/medium effect sizes. The corpus has a medium size effect. The interaction size between the corpus factor and the topic or formulations are of particular interest – both of which are large. This is further empirical evidence that topic difficulty is not a result of th information need: searching for a piece of information in specific corpora or using different formulations on different corpora can result in very different performance. Furthermore, this suggests that we are likely to find a specific formulation for which we achieve better (or worse) performance for any topic on any corpus with less effort than would be required when attempting to achieve similar performance differences by changing only the ranker.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Total	3799.35	125999					
Topic	1544.69	24	64.36	22370.77	0	0.7682	L
Formulation (Topic)	804.82	350	2.30	799.25	0	0.6330	L
System	87.80	143	0.61	213.41	0	0.1579	L
Corpus	33.04	2	16.52	5742.72	0	0.0662	M
System*Topic	216.04	3432	0.06	21.88	0	0.3067	L
System*Form.	240.36	50050	0.00	1.67	0	0.1713	L
System*Corpus	28.57	286	0.10	34.72	0	0.0562	S
Topic*Corpus	960.75	48	20.01	6956.96	0	0.6733	L
Form.*Corpus	527.09	700	0.75	261.72	0	0.5298	L
Topic*System*Corpus	214.42	6864	0.03	10.86	0	0.2946	L
Error	287.99	100100	0.00				
Total	4945.58	161999					

5.4.1 Macro-Level ANOVA. Table 11 shows the results for model ($MD2_{ma}$) using all three corpora – Robust 2004, CORE 2017 and CORE 2018. All the factors are again statistically significant.

We can observe that the topic factor has, as always, a large-size effect even across corpora and that the system factor becomes a moderately large-size effect, being bigger than in the single corpus case (see Table 3); the corpus factor has a medium-size effect. Overall, these results support similar findings reported by Ferro and Sanderson [32] while Zampieri et al. [82] reported that both systems and corpora had a very small-size effects. We also note that the query formulation factor has a remarkably large-size effect, even across corpora, observed here for the first time, suggesting it is a key contributor to topic difficulty.

The topic*system interaction has a noticeably large-size effect but half the size of the query formulation factor alone, and roughly two-thirds of the effect observed in the single corpus case; in addition, the query formulation*system interaction is a medium (almost large) size effect. Overall, this suggests that the multiple corpora further amplify the impact of query formulations, which was already very large. Note that the size of the topic*system interaction reaffirms similar findings from Ferro et al. [31], Zampieri et al. [82].

The topic*corpus interaction also has a large-size effect, the second biggest effect, which is aligned with the findings of Ferro et al. [31], Zampieri et al. [82]. Moreover, both the query formulation*corpus interaction and the topic*system*corpus interaction, observed here for the first time, are clearly important large-size effects.

Finally, we note that the system*corpus interaction is a medium-size effect, in contrast to previous results by Zampieri et al. [82] and Ferro et al. [31] who found it to have a small-size effect. This

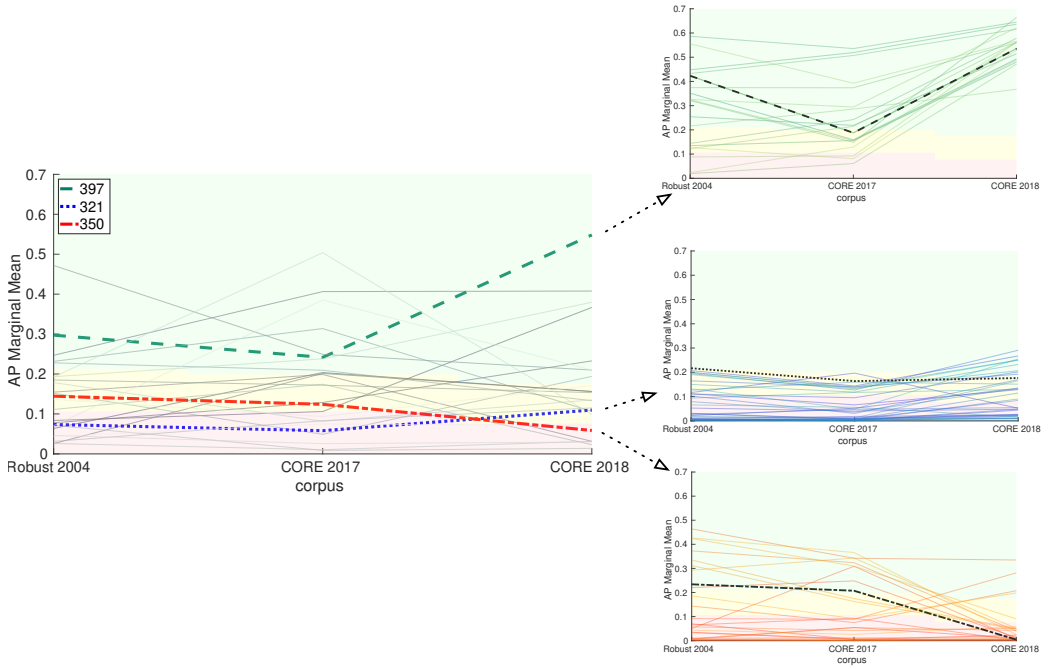


Fig. 9. Interaction effects between topics and corpora (on the left), where each line is a single topic. The Marginal AP is the average over all possible system configurations for either a topic considered as the combination of all its formulations (left), or a single topic formulation (right). Topic 321 in blue ('`women in parliaments`'), is almost always difficult; topic 350 in red ('`health and computer terminals`), almost always medium; and topic 397 in green ('`automobile recalls`'), always easy. On the right, for each of these three topics, the interaction between query formulations and corpora are demonstrated. The black line was the original formulation corresponding to the TREC title query. The red, yellow, and green bands correspond to the hard, medium, and easy query performance ranges.

behavior could be attributed to the presence of query reformulations in our model which increase the variance in performance for systems on different corpora.

Overall, these findings provide further evidence supporting the possibility that difficult topics do not actually exist in any absolute sense. We will further investigate this notion in Section 5.5 where the algorithm described in Section 4.2 leverages the large-size of the above interaction effects to show that it is actually possible to find any desired ranking of topics, providing further evidence that difficulty can not be confidently attributed to a particular topic.

Figure 9 provides a visualization of the volatility of topic difficult as collection and query formulation change. The red, yellow, and green bands in the figure correspond to the hard, medium, and easy query performance ranges, according to the traditional definition proposed in [22], where the 38% of the worst-performing queries have been considered hard, the 30% medium-performing queries have been defined medium, while the upper 30% of queries are the easy ones. On the left of Figure 9 we can see the plot of the topic*corpus interaction factor and we can observe that topics can be easy, medium or hard depending on the corpus. We also highlight specific topics that exhibit consistent effectiveness trends across collections: Hard (topic 321 in blue), Medium (topic 350 in red), and Easy (topic 397 in green). On the right, we expand all formulations*corpus interactions

for each of those highlighted topics, with the original TREC title query shown in black as a point of reference. We can see that regardless of whether a topic is classified as easy, medium, or hard, we can generally find at least one query reformulation for that topic in any of the three regions across the corpora.

5.4.2 Micro-Level ANOVA. Table 12 shows the results for model ($MD2_{mi}$) on the Robust 2004, CORE 2017 and CORE 2018 corpora. All the factors are again statistically significant.

Table 12. Model ($MD2_{mi}$) on the CORE 2018, CORE 2017 and Robust 2014 tracks using AP. The letter in the column "effect size" indicates whether the effect is large (L), medium (M) or small (S). For single factors, observe a similar overall behaviour to Model ($MD1_{mi}$). Overall, observe that the interactions including query formulations often have large effect sizes, indicating that using different formulations in combination with various components can induce dramatically different results. Furthermore, observe that several interactions of system components and the corpus have small or negligible effect sizes, indicating that the performance of these components are very similar in all of the corpora.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	1544.69	24	64.36	41853.88	<1e-6	0.8611	L
Formulations (Topic)	804.82	350	2.30	1495.32	<1e-6	0.7635	L
Stoplist	0.91	1	0.91	590.04	<1e-6	0.0036	(S)
Stemmer	0.02	1	0.02	15.46	<1e-6	0.0001	(S)
Model	19.64	8	2.45	1596.40	<1e-6	0.0730	M
Query Expansion	59.29	3	19.76	12851.05	<1e-6	0.1922	L
Corpus	33.04	2	16.52	10744.16	<1e-6	0.1171	M
Topic*Stoplist	1.14	24	0.05	30.83	<1e-6	0.0044	(S)
Topic*Stemmer	3.97	24	0.17	107.63	<1e-6	0.0156	S
Topic*Model	116.03	192	0.60	392.99	<1e-6	0.3172	L
Topic*Query Exp.	68.95	72	0.96	622.70	<1e-6	0.2165	L
Topic*Corpus	960.75	48	20.02	13015.90	<1e-6	0.7941	L
Form.*Stoplist	4.56	350	0.01	8.48	<1e-6	0.0159	S
Form.*Stemmer	18.23	350	0.05	33.86	<1e-6	0.0663	M
Form.*Model	84.64	2800	0.03	19.66	<1e-6	0.2438	L
Form.*Query Exp.	66.51	1050	0.06	41.19	<1e-6	0.2067	L
Form.*Corpus	527.09	700	0.75	489.66	<1e-6	0.6786	L
Corpus*Stoplist	0.05	2	0.03	17.41	<1e-6	0.0002	(S)
Corpus*Stemmer	0.16	2	0.08	52.31	<1e-6	0.0006	(S)
Corpus*Model	14.82	16	0.93	602.27	<1e-6	0.0561	S
Corpus*Query Exp.	11.55	6	1.93	1251.98	<1e-6	0.0443	S
Topic*Corpus*Stoplist	2.02	48	0.04	27.42	<1e-6	0.0078	(S)
Topic*Corpus*Stemmer	4.93	48	0.10	66.80	<1e-6	0.0191	S
Topic*Corpus*Model	76.70	384	0.20	129.88	<1e-6	0.2340	L
Topic*Corpus*Query Exp.	105.31	144	0.73	475.58	<1e-6	0.2967	L
Form.*Corpus*Stoplist	4.08	700	0.01	3.79	<1e-6	0.0119	S
Form.*Corpus*Stemmer	16.78	700	0.02	15.58	<1e-6	0.0593	S
Form.*Corpus*Model	104.42	5600	0.09	12.13	<1e-6	0.2778	L
Form.*Corpus*Query Exp.	65.57	2100	0.03	20.30	<1e-6	0.2001	L
Error	224.90	146250	0.00				
Total	4945.58	161999					

Table 12 shows the break-down by system factor for Table 11 by component contribution. As observed for the single-corpus case (see Table 7), the most important components are the IR model

Table 13. Ranking of topics analysis over 10,000 forward and backward permutations, 20,000 overall permutations. Observe that overall there is agreement between permutations of topics and rankings of formulations based on the relative performance. The corpus that is the easiest to target and find a given rank was CORE 2018, followed by Robust. The latter is an interesting outcome as several of the queries included in the Robust collection were known to be "hard" topics in previous TREC tracks. This finding provides additional evidence to the importance of the magnitude of the effect size of query formulations across corpora. Conversely, the ratio of exact rankings is small, indicating that the formulation task is rarely effortless.

Overall Statistics			
	Fwd Perms	**Bck Perms**	**All Perms**
Ratio of exact rankings	2.06	2.18	2.12
Best Kendall's tau	0.8588±0.0017	0.8579±0.0017	0.8583±0.0012
Robust 2004			
	Fwd Perms	**Bck Perms**	**All Perms**
Ratio of exact rankings	0.65	0.91	0.78
Best Kendall's tau	0.8053±0.0021	0.8030±0.0022	0.8041±0.0015
CORE 2017			
	Fwd Perms	**Bck Perms**	**All Perms**
Ratio of exact rankings	0.55	0.43	0.49
Best Kendall's tau	0.7040±0.0032	0.7002±0.0032	0.7021±0.0023
CORE 2018			
	Fwd Perms	**Bck Perms**	**All Perms**
Ratio of exact rankings	0.93	0.88	0.91
Best Kendall's tau	0.7977±0.0023	0.7958±0.0023	0.7967±0.0016

and query expansion, as well as their interaction with topics, while stop list and stemmers and their interaction with topics have small-size effects. However, while the single-corpus case showed that the topic*query expansion interaction is almost twice the size of the topic*model interaction, in the multiple-corpora case the opposite is true, and now the topic*model interaction are roughly 1.5 times the size of the topic*query expansion interaction. This could possibly due to IR models being a sort of "filter" with respect to corpora, whose impact change from corpus to corpus.

We can also observe, for the first time, the interaction between components and query formulations: the interaction with IR model and query expansion components have a large-size effect, almost the same size in this case, while the interaction with stemmers is now a medium-size effect, suggesting that the "clustering" induced by a stemmer can have an important impact on query reformulations. These trends are also confirmed by the third order interactions, i.e. topic * <component> * corpus and formulations * <component> * corpus.

Finally, when interpreting the component and corpora interaction, the most important components are, again, IR models and query expansion which have strong, large-size effects, with query expansion being roughly 1.3 times larger than IR models, while the interaction with stop lists and stemmers are negligible in size, indicating a very consistent behavior across corpora. This finding differs from that of Zampieri et al. [82] who found that the corpus*query expansion interaction had a negligible effect size. This also suggests that components of an IR system may indirectly contribute to topic difficulty as IR models and query expansion are clearly also sensitive to variations in topics and query formulations.

5.5 Topic Difficulty

Given these findings, we are finally in a position to revisit the fundamental tenet in IR that has been explored from many different angles in the past – the notion of *topic difficulty*.

Table 13 summarizes the outcomes of the topic ranking analysis using 10,000 permutations. For each permutation we considered both the *forward* and *backward* (or reverse) permutation. Thus, 20,000 total permutations were evaluated. We observed that in 2.12% of all permutations (slightly less forward than backward ones), it was possible to exactly match the permuted ranking of topics targeted[5], while on the other permutations we have a mean Kendall's tau of 0.85, indicating that the queries selected to induce the desired topic rankings were consistently close to the arbitrary target ordering. Given that these results represent a lower bound, they provide strong evidence that ordering topics by relative effectiveness is not intrinsically difficult, and in fact can be "arbitrarily" easy or difficult across many different corpora and system combinations.

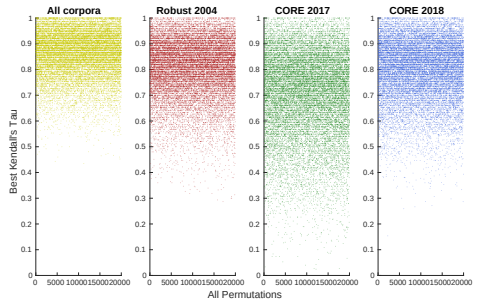
Table 13 also shows what happens when we restrict ourselves to using only a single collection to find the requested ranking of topics, which is a somewhat harder case. All of the collections exhibited similar behavior in terms of exact match ratio – 0.78% for Robust 2004, 0.49% for CORE 2017 and 0.91% for CORE 2018 – indicating that it is more difficult to find an exact solution for a random topic ordering on a single collection. On CORE 2018 this ratio was slightly higher, suggesting that in this case it is easier to find the requested ranking. When comparing the Kendall's tau results, Robust 2004 and CORE 2018 perform similarly while CORE 2017 was slightly worse, but in every case, there is a clear indication that it is possible to find topic orderings similar to the requested one, even on a single corpus.

Figure 10a is a visualization of the raw data which is summarized in Table 13. The Figure shows what happens when results are aggregated across all corpora as well as the outcome when each corpus is treated independently; the figure uses all available system configurations in our test set, i.e. we did not restrict our greedy algorithm only to the best configurations. Each permutation on the X-axis is plotted against the best Kendall's tau. We can clearly see that, regardless of corpora used, the Kendall's tau values tend to cluster above 0.6, suggesting that, for every permutation probed, it is quite possible to find at least one system which produces a similar ranking of topics being inspected. For the single corpora case, there is a higher likelihood of not finding a close mapping to the permutation being probed, but remains possible for many cases.

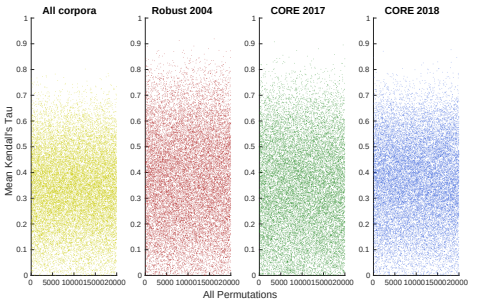
Figure 10b shows the mean Kendall's tau across the systems for each permutation. A very similar behavior can be observed across the different corpora, with values tending to below 0.5, again with a larger spread when only a single corpus is targeted. This suggests that the average behavior of systems is very noisy and that it is much more difficult to obtain a requested ranking of topics from a whole set of systems.

In Figure 10c each line represents a system for which the best Kendall's tau across all the permutations was computed. We can see strong evidence once again that nearly *every* system can find a solution for at least one permutation on each corpus, albeit with high variance in a few cases. There appears to often be more than one system which is able to find at least one exact match for a permutation being probed across all of the corpora. We note that CORE 2017 appears to be a little more difficult than the other two collections in this respect, since several of our system configurations tended to have worse overall effectiveness in this case. In Figure 10d each line is a system where the mean Kendall's tau across all the permutations is compared. When viewed from this perspective, it appears to be more difficult for a system to consistently find a close match for every permutation and this behavior is quite consistent in this respect across corpora, when enforcing a fixed system configuration. Again some corpora are more difficult than others, and small clusters of our system configurations are substantially more effective than the others.

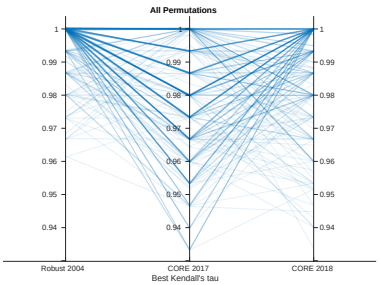
[5]The low rate of permutations for which it is possible to find an exact match provides further evidence that the corner case where a very poor query formulation for an "easy" topic still performs worse than a really good query formulation for a "difficult" topic is not a dominating factor, otherwise this value would be much higher.



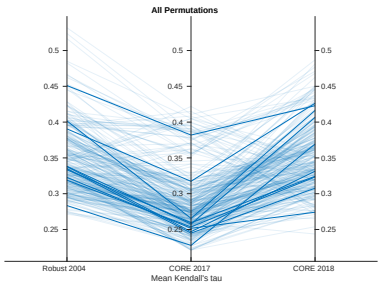
(a) Best Kendall's tau for each permutation. Each dot is the best system for that permutation.



(b) Mean Kendall's tau for each permutation. Each dot is the mean across systems for that permutation



(c) Best Kendall's tau for each system. Each line is a system, selecting its best Kendall's tau across permutations.



(d) Mean Kendall's tau for each system. Each line is a system, averaging its Kendall's tau across permutations.

Fig. 10. Ranking of topics analysis across the different corpora.

5.6 Topic Difficulty – Lessons Learned

We have presented evidence showing that topic difficulty is not an intrinsic property of an information need – meaning that query formulation based on a corpus and retrieval system, can be combined to sort topics arbitrarily based on a performance goal. While IR researchers have long been aware of the importance of query terms, the *magnitude* of the impact relative to other change to a system, such as the ranker, or even the introduction of query expansion has never been shown experimentally. ANOVA provides a powerful methodology to do it. While not discussed in detail in this work, the recent work of Liu et al. [45] show how similar query formulations are within a topic or when compared across different topics. It is remarkable how much performance can differ between two formulations of the same information need, with other factors being fixed.

This information can be used to further improve retrieval systems in IR as well as other related areas such as product, movie, or music recommendation. How can such a finding help researchers develop more effective retrieval systems? Firstly, it is worth noting that current evaluation paradigms usually consider a single formulation for each topic. All of the main evaluation campaigns, such as TREC or CLEF, allow participants to produce a single run for a given set of topics and queries. Such an arrangement prevents us to observe system behavior with small changes to each query. Automatic reformulation and query expansion (as shown in Tables 7 and 12) tend to have a large impact on the performance, one that is consistently greater than the retrieval model. If we consider

carefully the conclusions reached in this work, we can see the potential value of investigating it further, in many different scenarios and applications, none of which are happening today.

We believe that multiple formulations are a key omission in our current evaluation campaigns, and we are hopeful future campaigns will incorporate them into their methodology. The cost of collecting new data is certainly a limiting factor in every decision, but note that there is a high "bang-for-buck" here as query formulations for a single topic are all trying to find the *same* relevant documents, and there is often higher overlap in the documents retrieved. Query formulations can easily be collected through click-logs, produced automatically using ontologies or written directly by the assessors or crowdworkers. This additional information can be used to improve current evaluation methodology, with a small impact on the cost to develop a collection while providing substantial benefits. Using the ANOVA framework proposed here, practitioners can study formulations, systems, collections, and their associated interactions in exhaustive detail. More comprehensive failure analysis tools allow us to build more reliable systems, and to identify and eliminate tail cases which lead to poor performance under certain conditions.

Other applications can also benefit from the outcomes in this work. One IR research field which can benefit from a more theoretically grounded concept definition of "topic difficulty" is QPP. In QPP evaluation, queries are commonly ranked according to the effectiveness score produced by the QPP engine when mapped to some targeted evaluation metric such as AP or nDCG. Then, a correlation measure such as Kendall's τ or Spearman's ρ is computed between the two rankings and the quality of the QPP approach is derived from the rank correlation. When topic difficulty is so heavily bounded to the specific query being used, we should ask ourselves whether results observed for a specific corpus might be entirely different if formulations for the topics can be changed. Before we can fully understand this, multiple formulations should be considered in the evaluation. The work done using reference lists in QPP [62, 70] could easily be adapted to explore this notion further. Building reference lists for multiple formulations of topics would enable us to develop systems that better generalize with user behavior, as we are not able to know a priori which formulation a user might choose.

Another area that might benefit is conversational IR. Traditional evaluation in conversational IR requires sequences of utterances of a conversation between a system and a user. Again, it is common to have only one conversation per topic, and each utterance represents, individually, a query in a sequence. Such utterances lack variability that might occur in similar conversations in different users. As the number of possible formulations of human utterances grows exponentially, we are often limited only a few different dialogues in a campaign – a problem which has been previously discussed in the research community [4, 26, 55]. Nevertheless, the same "bang-for-buck" opportunities exist when expanding the date for each topic. A great deal of information can be created which is invaluable to practitioners but that have a much lower cost to collection creation and curation and if additional "new" topics are added. Our work demonstrates the importance of multiple formulations in the traditional IR, and can easily be extended in conversational IR.

To conclude, the lessons learned from our work have a number of potential applications and extensions. First, we have empirically observed the effect size of query formulations in IR evaluation, which is, in itself, an important issue warranting further attention. We show how query formulations interact with other components commonly found in a typical retrieval pipeline. These interactions were found often to be significant, and are rarely negligible in size. Their omission from current practices in IR evaluation should be reconsidered in the research community. If our goal is to model real performance of systems, collections creators should explore how to best include multiple formulations of each topic. Their use should also be standard practice for system builders as it is a valuable tool to perform a detailed performance analysis in complex retrieval software that is composed of multiple components – all of which can have unexpected interaction which can

degrade (or improve) the overall retrieval performance. Our isolation of multiple formulations of topics has allowed us to study in detail the concept of "topic difficulty", which we now understand to be a construct of a specific retrieval configuration – the collection, the system and the query which represents the topic under consideration – and not a property intrinsic to the topic alone. The malleability of relative performance which can be induced when of multiple query variations of a topic are available enable "topic difficulty" to be fully controlled in a collection, and raises important questions in current several of the communities current evaluation practices and potentially in related fields, such as QPP and conversational IR. That is, the distribution of which topics perform well or poorly can be arbitrarily reordered using query formulations such that relative system performances change, as their performance may be better or worse depending on the specific query choice. The implications of this observation should not be underestimated. The research community is heavily reliant on test collections and state-of-the-art performance comparisons, which are routinely used to verify that new approaches outperform current ones. However, if winning and losing can be manipulated based only on the query being used and everything else is the same – including the human relevance labels for the topics, are our current practices as reliable as we need them to be? Many open questions remain, Query formulation and its role in evaluation warrants further study in the IR community.

5.7 Efficiency and Scalability

While we have discussed several new avenues of future work in Section 5.6, we have not discussed one of the important challenges we encountered, which is the efficiency and scalability of current ANOVA modeling techniques. Query formulations and the wide-spread availability of publicly available retrieval systems such at Terrier allowed us to produce far more data than we could incorporate into our models. Every factor introduced into a model results in an order of magnitude increase in the number possible combinations, all of which must be ran and included. For example, 9 rankers, 4 query expansion models, 4 stemmers, stopping (2) on 3 collections requires 864 retrieval runs, each of which are composed of hundreds (or even thousands) of queries to run (50 topics of 25 query formulations is 1,250 total queries). So, a rather modest number of individual component choices require retrieval results for 1,080,000 queries total! Unfortunately, that is only the cost to create the initial dataset. Running the ANOVA model on the resulting data has a significant computational cost as well, and our current experiments we limited primarily by the RAM available, with our largest server having 1.5TB of RAM. We are aware of very few other studies using ANOVA in the IR community using data at this scale. The bottom line is that it is costly to do large-scale ANOVA analysis, and the current software available to perform ANOVA analysis (Matlab or R) were not designed to use data of this magnitude. But for IR researchers who are interested in designing efficient and scalable algorithms, a scalable and efficient ANOVA framework for CUDA and other GPU related hardware would be a valuable contribution to the community. Remarkable achievements are possible using GPU hardware in the Deep Learning community. However, most of these efforts are dedicated to building new NLP/ML models, and not in leveraging it to evaluate models we create. We can and should be using this new hardware to improve current evaluation techniques too.

6 CONCLUSION

In this work, we have presented a comprehensive ANOVA analysis that compares the effect sizes across multiple corpora and retrieval system configurations. We have also generalized previous model configurations in order to incorporate a new nesting factor which maps an information need (topic) to multiple query formulations. The removal of the constraint of a 1:1 mapping between a query and a topic has led to several interesting observations which have important

implications on the notion of topic difficulty. The models were designed to enable us to systematically analyze multiple component interactions in a single study, several of which have not been possible before, such as query formulation*system interactions. The configuration shows that component interactions are strongly influenced by query formulation, which consistently increases the effect size across multiple component interactions and configurations.

We also propose an analysis methodology, based on a permutation algorithm, to further explore topic difficulty. Based on this new knowledge, we were able to show conclusive evidence that topic difficulty is not invariant, and therefore care should be taken when relying on topic ordering to evaluate the quality of various prediction models, as is common practice for QPP.

Leveraging the lessons learned in our current ANOVA design, our future work will directly explore QPP methods and how factors impact prediction quality. We would also like to allow latent factors to emerge more easily, such as the role of unique relevant documents that are retrieved in each system configuration. Finally, prior work has demonstrated that unique relevant documents have an important role in evaluation pooling [17, 75] but, no systematic studies on how they affect system performance have been conducted, or the relationship between topics and query formulations on unique relevant documents, or how this might change across corpora.

ACKNOWLEDGEMENTS

We thank the reviewers for their helpful comments. The work is partially funded by the "DAta BenchmarK for Keyword-based Access and Retrieval" (DAKKAR) Starting Grants project sponsored by University of Padua and Fondazione Cassa di Risparmio di Padova e di Rovigo, University of Padova Strategic Research Infrastructure Grant 2017: "CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione", Australian Research Council Grant DP190101113, and the Israel Science Foundation under grant no. 1136/17.

REFERENCES

[1] J. Allan, D. K. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. M. Voorhees. 2018. TREC 2017 Common Core Track Overview, See [76].

[2] J. Allan, D. K. Harman, E. Kanoulas, and E. M. Voorhees. 2019. TREC 2018 Common Core Track Overview. In *The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-331, Washington, USA.

[3] Giambattista Amati, Claudio Carpineto, and Giovanni Romano. 2004. Query Difficulty, Robustness, and Selective Application of Query Expansion. In *Advances in Information Retrieval*, Sharon McDonald and John Tait (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 127–137.

[4] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). In *Dagstuhl Reports*, Vol. 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[5] J. A. Aslam and V. Pavlu. 2007. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proc. ECIR*. 198–209.

[6] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2015. User Variability and IR System Evaluation. In *Proc. 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, R. Baeza-Yates, M. Lalmas, A. Moffat, and B. Ribeiro-Neto (Eds.). ACM Press, New York, USA, 625–634.

[7] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2017. Retrieval consistency in the presence of query variations. In *Proc. SIGIR*. 395–404.

[8] D. Banks, P. Over, and N.-F. Zhang. 1999. Blind men and elephants: Six approaches to TREC data. *Inf. Retr.* 1, 1 (1999), 7–34.

[9] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. 1993. The effect of multiple query variations on information retrieval system performance. In *Proc. SIGIR*. 339–346.

[10] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. 1995. Combining the evidence of multiple query representations for information retrieval. *Inf. Proc. & Man.* 31, 3 (1995), 431–448.

[11] R. Benham and J. S. Culpepper. 2017. Risk-reward trade-offs in rank fusion. In *Proc. ADCS*. 1:1–1:8.

[12] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. Mackenzie. 2018. Towards efficient and effective query variant generation. 62–67.

[13] R. Benham, L. Gallagher, J. Mackenzie, T. T. Damessie, R.-C. Chen, F. Scholer, A. Moffat, and J. S. Culpepper. 2017. RMIT at the 2017 TREC CORE track. In *Proc. TREC*.

[14] R. Benham, L. Gallagher, J. Mackenzie, B. Liu, X. Lu, F. Scholer, A. Moffat, and J. S. Culpepper. 2018. RMIT at the 2018 TREC CORE track. In *Proc. TREC*.

[15] D. Bodoff and P. Li. 2007. Test theory for assessing IR test collections. In *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando (Eds.). ACM Press, New York, USA, 367–374.

[16] R. L. Brennan. 2001. *Generalizability Theory*. Springer-Verlag, New York, USA.

[17] C. Buckley, D. Dimmick, I. Soboroff, and E. M. Voorhees. 2007. Bias and the Limits of Pooling for Large Collections. *Information Retrieval* 10, 6 (December 2007), 491–508.

[18] C. Buckley, G. Salton, J. Allan, and A. Singhal. 1995. Automatic query expansion using SMART: TREC 3. In *Proc. TREC*.

[19] C. Buckley and E. M. Voorhees. 2005. Retrieval System Evaluation. In *TREC. Experiment and Evaluation in Information Retrieval*, D. K. Harman and E. M. Voorhees (Eds.). MIT Press, Cambridge (MA), USA, 53–78.

[20] K. P. Burnham and D. R. Anderson. 2002. *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach* (2nd ed.). Springer-Verlag, Heidelberg, Germany.

[21] D. Carmel and E. Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool Publishers.

[22] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. 2009. Million Query Track 2009 Overview.. In *TREC*.

[23] B. A. Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 4:1–4:34.

[24] S. Cronen-Townsend and W. B. Croft. 2002. Quantifying query ambiguity. In *Proc. HLT*. 104–109.

[25] T. Dammesie, F. Scholer, and J. S. Culpepper. 2016. The Influence of Topic Difficulty, Relevance Level, and Document Ordering on Relevance Judging. In *Proc. ADCS*. 41–48.

[26] G. Faggioli, M. Ferrante, N. Ferro, R. Perego, and N. Tonellotto. 2021. Hierarchical Dependence-Aware Evaluation Measures for Conversational Search. In *Proc. of SIGIR (SIGIR 2021)*.

[27] G. Faggioli and N. Ferro. 2021. System Effect Estimation By Sharding: A Comparison Between ANOVA approaches to Detect Significant Differences. In *Proc. 43nd European Conference on IR Research (ECIR 2021)*.

[28] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, and Scholer F. 2021. An Enhanced Evaluation Framework For Query Performance Prediction. In *Proc. 43nd European Conference on IR Research (ECIR 2021)*.

[29] N. Ferro. 2018. IMS @ TREC 2017 Core Track, See [76].

[30] N. Ferro and D. Harman. 2010. CLEF 2009: Grid@CLEF Pilot Track Overview. In *Proc. CLEF*. 552–565.

[31] N. Ferro, Y. Kim, and M. Sanderson. 2019. Using Collection Shards to Study Retrieval Performance Effect Sizes. *ACM Trans. Inf. Sys.* 5, 44 (2019), 59.

[32] N. Ferro and M. Sanderson. 2017. Sub-corpora impact on system effectiveness. In *Proc. SIGIR*. 901–904.

[33] N. Ferro and M. Sanderson. 2019. Improving the Accuracy of System Performance Estimation by Using Shards. In *Proc. 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, B. Piwowarski, M. Chevalier, E. Gaussier, Y. Maarek, J.-Y. Nie, and F. Scholer (Eds.). ACM Press, New York, USA, 805–814.

[34] N. Ferro and G. Silvello. 2016. A general linear mixed models approach to study system component effects. In *Proc. SIGIR*. 25–34.

[35] N. Ferro and G. Silvello. 2018. Toward an anatomy of IR system component performances. 69, 2 (2018), 187–200.

[36] C. Hauff, D. Kelly, and L. Azzopardi. 2010. A comparison of user and system query performance predictions. In *Proc. CIKM*. 979–988.

[37] Ben He and Iadh Ounis. 2004. Inferring Query Performance Using Pre-retrieval Predictors. In *String Processing and Information Retrieval*, Alberto Apostolico and Massimo Melucci (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 43–54.

[38] P. K. Ito. 1980. Robustness of ANOVA and MANOVA test procedures. In *Handbook of Statistics – Analysis of Variance*, P. R. Krishnaiah (Ed.), Vol. 1. Elsevier, The Netherlands, 199–236.

[39] T. Jones, A. Turpin, S. Mizzaro, F. Scholer, and M. Sanderson. 2014. Size and Source Matter: Understanding Inconsistencies in Test Collection-Based Evaluation. In *Proc. 23rd International Conference on Information and Knowledge Management (CIKM 2014)*, X. Li, X. S. Wang, M. Garofalakis, I. Soboroff, T. Suel, and M. Wang (Eds.). ACM Press, New York, USA, 1843–1846.

[40] M. G. Kendall. 1948. *Rank correlation methods*. Griffin, Oxford, England.

[41] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (March 1951), 79–86.

[42] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin, New York, USA.

[43] Kui-Lam Kwok, Laszlo Grunfeld, HL Sun, Peter Deng, and N Dinstl. 2004. TREC 2004 Robust Track Experiments Using PIRCS.. In *TREC*. National Institute of Standards and Technology. NIST.

[44] V. Lavrenko and W. B. Croft. 2001. Relevance-based language models. In *Proc. SIGIR*. 120–127.

[45] B. Liu, N. Craswell, O. Kurland, and J. S. Culpepper. 2019. A Comparative Analysis of Human and Automatic Query Variants. In *Proc. ICTIR*. 47–50.

[46] Xiaolu Lu, Oren Kurland, J. Shane Culpepper, Nick Craswell, and Ofri Rom. [n. d.]. Relevance Modeling with Multiple Query Variations. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019*. 27–34.

[47] S. Maxwell and H. D. Delaney. 2004. *Designing Experiments and Analyzing Data. A Model Comparison Perspective* (2nd ed.). Lawrence Erlbaum Associates, Mahwah (NJ), USA.

[48] R. McGill, J. W. Tukey, and W. A. Larsen. 1978. Variations of Box Plots. *The American Statistician* 32, 1 (February 1978), 12–16.

[49] W. Mendenhall and T: Sincich. 2012. *A Second Course in Statistics. Regression Analysis* (7th ed.). Prentice Hall, USA.

[50] Stefano Mizzaro. 2008. The Good, the Bad, the Difficult, and the Easy: Something Wrong with Information Retrieval Evaluation?. In *Advances in Information Retrieval*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 642–646.

[51] S. Mizzaro and S. Robertson. 2007. HITS hits TREC: exploring IR evaluation results with network analysis. In *Proc. SIGIR*. 479–486.

[52] S. Olejnik and J. Algina. 2003. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods* 8, 4 (December 2003), 434–447.

[53] Jovan Pehcevski, James A. Thom, Anne-Marie Vercoustre, and Vladimir Naumovski. 2009. Entity ranking in Wikipedia: utilising categories, links and topic difficulty prediction. *Information Retrieval* 13 (2009), 568–600. https://doi.org/10.1007/s10791-009-9125-9

[54] J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis. 2007. Incorporating Term Dependency in the DFR Framework. In *Proc. SIGIR*. 843–844.

[55] G. Penha and C. Hauff. 2020. Challenges in the evaluation of conversational search systems. *CEUR Workshop Proceedings* 2666.

[56] J. Pérez-Iglesias and L. Araujo. 2010. Standard deviation as a query hardness estimator. In *Proc. SPIRE*. 207–212.

[57] J. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proc. SIGIR*. 275–281.

[58] Fiana Raiber and Oren Kurland. 2014. Query-performance prediction: setting the expectations straight. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin (Eds.). 13–22.

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

[60] J. J. Rocchio. 1971. *The SMART retrieval system: Experiments in automatic document processing*. Prentice Hall, Englewood Cliffs, NJ.

[61] Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. 2017. Do Easy Topics Predict Effectiveness Better Than Difficult Topics?. In *Advances in Information Retrieval*, Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (Eds.). Springer International Publishing, Cham, 605–611.

[62] Haggai Roitman. 2017. An Enhanced Approach to Query Performance Prediction Using Reference Lists. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 869–872. https://doi.org/10.1145/3077136.3080665

[63] A. Rutherford. 2011. *ANOVA and ANCOVA. A GLM Approach* (2nd ed.). John Wiley & Sons, New York, USA.

[64] T. Sakai. 2014. Metrics, Statistics, Tests. In *Bridging Between Information Retrieval and Databases - PROMISE Winter School 2013, Revised Tutorial Lectures*, N. Ferro (Ed.). Lecture Notes in Computer Science (LNCS) 8173, Springer, Heidelberg, Germany, 116–163.

[65] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. 2012. Differences in effectiveness across sub-collections. In *Proc. CIKM*. 1965–1969.

[66] S. M. Scariano and J. M. Davenport. 1987. The Effects of Violations of Independence Assumptions in the One-Way ANOVA. *The American Statistician* 41, 2 (1987), 123–129.

[67] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proc. SIGIR*. 623–632.

[68] R. J. Shavelson and N. M. Webb. 1991. *Generalizability Theory. A Primer*. SAGE Publishing, USA.

[69] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. 2011. LambdaMerge: Merging the results of query reformulations. In *Proc. WSDM*. 795–804.

[70] A. Shtok, O. Kurland, and D. Carmel. 2016. Query Performance Prediction Using Reference Lists. *ACM Trans. Inf. Sys.* 34, 4 (2016), 19:1–19:34.
[71] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Sys.* 30, 2 (2012), 11.
[72] J. M. Tague-Sutcliffe and J. Blustein. 1994. A Statistical Analysis of the TREC-3 Data. In *Proc. TREC*. 385–398.
[73] E. M. Voorhees. 2004. Overview of the TREC 2004 Robust Retrieval Track.. In *Proc. TREC*.
[74] E. M. Voorhees. 2005. Overview of the TREC 2005 Robust Retrieval Track.. In *Trec*.
[75] E. M. Voorhees. 2018. On Building Fair and Reusable Test Collections using Bandit Techniques. In *Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018)*, A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Broder, M. J. Zaki, S. Candan, A. Labrinidis, A. Schuster, and H. Wang (Eds.). ACM Press, New York, USA, 407–416.
[76] E. M. Voorhees and A. Ellis (Eds.). 2018. *The Twenty-Sixth Text REtrieval Conference Proceedings (TREC 2017)*. National Institute of Standards and Technology (NIST), Special Publication 500-324, Washington, USA.
[77] E. M. Voorhees, D. Samarov, and I. Soboroff. 2017. Using Replicates in Information Retrieval Evaluation. *ACM Trans. Inf. Sys.* 36, 2 (2017), 12:1–12:21.
[78] M. P. Wand and M. C. Jones. 1995. *Kernel Smoothing*. Chapman and Hall/CRC, USA.
[79] J. Xu and W. B. Croft. 1996. Query expansion using local and global document analysis. In *Proc. SIGIR*. 4–11.
[80] E. Yilmaz, J. A. Aslam, and S. E. Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *Proc. SIGIR*. 587–594.
[81] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. 2005. Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In *Proc. SIGIR*. 512–519.
[82] F. Zampieri, K. Roitero, J. S. Culpepper, O. Kurland, and S. Mizzaro. 2019. On Topic Difficulty in IR Evaluation: The Effect of Systems, Corpora, and System Components. In *Proc. SIGIR*. 909–912.
[83] O. Zendel, A. Shtok, F. Raiber, O. Kurland, and J. S. Culpepper. 2007. Information Needs, Queries, and Query Performance Prediction. In *Proc. SIGIR*. 395–404.

Received August 2020; revised February 2021; accepted June 2021