Using Collection Shards to Study Retrieval Performance Effect Sizes

NICOLA FERRO, University of Padua, Italy
YUBIN KIM, UPMC Enterprises, USA
MARK SANDERSON, RMIT University, Australia

Despite the bulk of research studying how to more accurately compare the performance of IR systems, less attention is devoted to better understanding the different factors which play a role in such performance and how they interact. This is the case of shards, i.e. partitioning a document collection into sub-parts, which are used for many different purposes, ranging from efficiency to selective search or making test collection evaluation more accurate. In all these cases there is empirical knowledge supporting the importance of shards but we lack actual models that allow us to measure the impact of shards on system performance and how they interact with topics and systems. We use the general linear mixed model framework and present a model that encompasses the experimental factors of system, topic, shard, and their interaction effects. This detailed model allows us to more accurately estimate differences between the effect of various factors. We study shards created by a range of methods used in prior work and better explain observations noted in prior work in a principled setting and offer new insights. Notably, we discover that the topic*shard interaction effect, in particular, is a large effect almost globally across all datasets, an observation that, to our knowledge, has not been measured before.

CCS Concepts: • **Information systems** → **Test collections**; **Retrieval effectiveness**.

Additional Key Words and Phrases: shard effect; effectiveness model; GLMM; ANOVA

ACM Reference Format:
Nicola Ferro, Yubin Kim, and Mark Sanderson. 2019. Using Collection Shards to Study Retrieval Performance Effect Sizes. *ACM Transactions on Information Systems* 0, 0, Article 0 (January 2019), 38 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Information Retrieval (IR) researchers have accumulated empirical knowledge about which retrieval techniques work best by quantitatively measuring the performance of IR systems as *scores* [12, 27, 49]. One way of obtaining such scores is to use test collections, which offline, provide a means of measuring system performance. However, the measurement is beset with noise, of which, topics are a substantial source. In time, the noise was modeled so as to better understand it and provide a statistical means to cope.

Models showed that the variance in scores across topics was larger than that measured across systems. Tague-Sutcliffe and Blustein [56] and Banks et al. [5] studied this phenomenon using two-way *ANalysis Of VAriance (ANOVA)* and *General Linear Mixed Model (GLMM)*, which are complementary ways to model and analyse data. Such statistical processes model scores as a

Authors' addresses: Nicola Ferro, University of Padua, Italy, ferro@dei.unipd.it; Yubin Kim, UPMC Enterprises, USA, kimy10@upmc.edu; Mark Sanderson, RMIT University, Australia, mark.sanderson@rmit.edu.au.

1046-8188/2019/01-ART0 $15.00
https://doi.org/0000001.0000001

combination of factors and factor interactions. The models were extended to include a topic*system interaction [6, 45, 60].

Ferro et al. [17, 18] and Allan et al. [1] explained how comprehensive *performance models* are a key element to develop richer explanations of IR system performance. An improved model provides a stepping stone for developing predictive models, which are able to anticipate the performance of IR systems in new operational conditions.

A number of developments in IR system research have occurred that may require a re-examination of the models.

- There is increasing focus in IR on *shards*: partitions of a collection, often formed for efficiency reasons [34]. The efficiency arises from the fact that topics can be easier to serve from some shards than others. This aspect of IR systems is not currently modeled.
- There has also been a recent examination of a novel way to make test collection evaluation more accurate by sharding [60]. However, the results of the work are perhaps more noisy than was expected. A statistical model of this approach could perhaps explain the noise.
- Past work has shown that certain retrieval systems perform better on some shards than on others [51]. While some models have been built to understand that work [20], the models are simple and open for improvement.
- A track at TREC found that topics thought to be 'difficult' were actually difficult on one collection, but 'easy' on another [57]. While this observation is not directly related to shards, the documents of test collections are often samples of a wider population. That this topic variation exists suggests the collections may be under sampled. An understanding of this aspect of topic score variation might be better understood through shards.

All four developments have a common theme: the interaction of large groups of documents (collections or shards) with other elements of retrieval performance: topics and systems. While there is an understanding of the importance of shards to search, there is little work *measuring* that impact on system performance and no understanding of the impact relative to the other factors of evaluation schemes. Therefore, this paper presents an investigation of the effects of topics, systems, and *shards* on performance and an extension of a statistical model that includes shards.

We conducted an experimental analysis using *Text REtrieval Conference (TREC)*[1] collections from the Adhoc, Web, and Terabyte tracks using a range of document sharding methods. We considered different types of IR systems: those originally submitted to TREC tracks and our own set of systems, each optimized to a particular shard; this allowed for accounting for a wide range of IR scenarios: from adhoc retrieval to distributed and vertical search. The main contributions of the paper are:

- an extended performance model incorporating a shard factor and its interaction with other factors; coupled with an initial analysis of the extended model's ability to measure system performance more accurately than past models;
- a detailed study of the impact of different shard types on system performance, quantifying and better explaining effects seen in the past;
- a discussion of one component of the model: topic*shard interaction, which is measured to be one of the biggest effects in system performance. While shards have long been known to be important to performance, a measure of the effect was not observed before.

The paper is organized as follows: Section 2 conducts a review of the relevant literature on the topic; Section 3 provides background information on GLMM and ANOVA and on how they work; Section 4 introduces our methodology; Section 5 describes our experimental setup; Section 6 validates our methodology; Section 7 reports our experimental findings; Section 8 discusses and

[1] https://trec.nist.gov/

summarises the observed general trends; finally, Section 9 draws some conclusions and outlooks for future work.

2 LITERATURE REVIEW

We describe past work examining modeling system scores using ANOVA, the sharding of collections, and examinations of topic difficulty.

2.1 ANOVA

There has long been an understanding that performance measured on test collections is noisy. In 1968, Salton and Lesk [50] described the use of significance tests in IR evaluation, in part, because "*...it is difficult to draw hard conclusions because the variation in performance between individual [topics] is large...*".

While known about for many years, the magnitude of topic variation was not measured until Tague-Sutcliffe and Blustein [56] conducted an ANOVA of TREC-3 adhoc system results (i.e. runs). The authors compared the variation in performance across topics with that across systems, finding that the variation in the former was larger than the latter. The results suggested that "*...for some [topics] some [systems] are best and for other [topics] other [systems] are best.*" The interaction effects of systems and topics was not measured directly as only one observation of each topic*system combination existed in the TREC-3 data, while more replicates/observations would have been needed for each (topic, system) pair to estimate their interaction. Later, still using only one observation of each topic*system combination, Banks et al. [5] employed methods to approximate the interaction, suggesting it would be strong and significant.

Bodoff and Li [6] presented results where multiple topic*system measures were generated due to having > 1 relevance assessment. A topic*system interaction was reported, the magnitude of which was found to be less than the topic effect, but greater than the system effect. Robertson and Kanoulas [45] used a simulation to generate topic*system pair replicates and relied on GLMM and ANOVA to study the interaction, further showing its importance.

The factors used in an ANOVA analysis do not have to be the components of a test collection. Ferro and Silvello [21, 22] systematically varied the components of an IR system: stop list, stemmer, ranking model, etc. by using the grid-of-points approach proposed by Ferro and Harman [19]. The analysis allowed the researchers to understand the relative impact of each system component on performance.

More recently, Voorhees et al. [60] conducted experiments where the researchers randomly split TREC collections into shards, thus creating more replicates for each (topic, system) pair and allowing them to examine topic*system interactions. By modeling the interactions, Voorhees et al. were able to measure more significant differences between retrieval systems. However, the benefits of this approach were limited to collections split into two or three shards. Split any further, noise appeared to dominate measurement.[2]

2.2 Shards

Shards are a common feature of search engines, many of the popular commercial systems operate aggregated search across heterogeneous collections. In such systems, topics are directed to particular shards that are specialized for different domains of data (e.g. maps, news, video, images, etc.) [38, 55]. It is clear that in such search, some shards are better sources of relevant items than others for

[2]See also work by Yang et al. [61] who used a bootstrapping approach that in effect split document collections into multiple per-topic shards. The authors used the analysis to consider how accurately differences between retrieval systems were being measured.

particular topics. What is less clear is how shards within homogeneous collections interplay with performance scores. We first examine past work on shard interactions with topics and systems.

Test collections can be split into shards in different ways. For example, sharding is a common technique in distributed search where shards have been created based on geography, quality, and random assignment [11]. In federated search, multiple collections or the outputs of multiple search engines are aggregated. Selective search, which combines the ideas from distributed and federated search has been examined. Kulkarni and Callan [34] experimented with random, source-based, and topic-based shards to reduce the computational cost of large-scale search. While an increasingly common feature, the shards, such as these, have not been modeled in an ANOVA analysis.

Considering shards and retrieval performance, Zhai and Lafferty [63] experimented with different parameter settings for a retrieval system searching on different shards from a TREC collection. The shards were defined as source based (i.e. FT, LA Times, FR, etc.), the authors found that there were notably different parameter settings for each shard. Zhai and Lafferty [64] later incorporated a collection model into their probabilistic framework. The aim of the model was to capture qualities of the set of documents being searched. The researchers observed that the greatest influence on parameter setting for the model seemed to come from topic variation [65]. There were differences between shards, but not substantial. The incorporation of the collection model did not seem to bring notable benefits to retrieval performance.

Unaware of the past work of Zhai and Lafferty [63], Sanderson et al. [51] similarly split TREC adhoc collections into shards based on the source of the documents. The authors examined if rankings of systems generated from different shards were significantly different from each other. In contrast to the previous work, the authors found that they were, though the size of the measured effect was not clear. A followup article showed similar effects even when potentially confounding factors were removed from experiments [32].

More recently, Goswami et al. [24] tried finding an ideal ranking function that worked across the widest range of test collections. The authors approximated a form of learning-to-rank by searching in a space of possible ranking functions up to a maximum length of function complexity. In order to prevent an explosion of experiments, the authors used a staged process of testing the largest number of functions on a single test collection then selecting the 1,000 best functions and testing those functions on a wider range of test collections. The best resulting functions were tested on two final collections that were held out. In order to maximize the number of test collections available, the researchers sharded a number of TREC collections using a partitioning similar to Sanderson et al.'s approach. The results indicated that this process led to a more stable ranking function than some of the standard functions.

What the work here demonstrates is that across different shards, systems appears to be affected by the shards themselves. What is not clear, however, is how substantial, practical, or predictable the effect is in comparison to other effects. As an initial examination of shard effects, Ferro and Sanderson [20] incorporated shards into an ANOVA model. Studying a limited number of collections, the authors showed that such effects were present, but that shard*system interactions were minimal.

2.3 Topics Difficulty

The Robust Retrieval track of TREC [57] was formed to determine retrieval approaches that could improve the consistency of search by focusing on poorly performing topics. Topics measured to be difficult on one collection were run against a new collection. A result of the track was that *'...topics are not inherently easy or difficult in isolation–the difficulty depends on the interaction between the information need and information source'*. Groups that tried to predict topic difficulty by training on one collection found prediction on the other collection was *'largely unsuccessful,'*. One could view

the two collections as two shards, drawn from a larger population of documents. The nature of this topic*shard interaction was not examined.

3 GLMM AND ANOVA

A *General Linear Mixed Model (GLMM)* [39, 47] explains the variation of a dependent variable ("Data") in terms of 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 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, i.e. 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 (predictor) times the regression coefficient β_1, i.e. the slope of the regression line, plus a residual error ε_i, not explained by the model, which follows a Gaussian distribution with mean 0.

In GLMM terms, *ANalysis Of VAriance (ANOVA)* 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 do dependent variable score means differ and what proportion of variation in the dependent variable can be attributed to differences between specific experimental groups or conditions, as defined by the independent variable(s). An ANOVA can be regarded as a particular type of regression analysis that employs only categorical predictors.

The previous regression model is expressed in ANOVA terms as $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, where Y_{ij} is the i-th subject's dependent variable score in the j-th experimental condition, the parameter μ is the grand mean of the experimental condition population means that underlies all subjects' dependent variable scores, the parameter α_j is the effect of the j-th experimental condition and the random variable ε_{ij} is the error term, which reflects variation due to any uncontrolled 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, e.g., by using dummy coding.

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.

When it comes to independent variables they can be either *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 randomly and independently drawn set of levels that represent variation in 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 sampling of effect levels.

The experimental design determines how you compute the model and how you estimate its parameters. In particular, it is possible to have an *independent measures* design where different subjects participate to different experimental conditions (factors) or a *repeated measures* design, where each subject participates to all 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 Estimating the Model

Figure 1 shows the experimental layout for the ANOVA model reported in equation (1). This is the typical IR setting where you have a set of topics and a set of systems which are run against those topics; in ANOVA terms this is a crossed/factorial repeated measures design. Note that this is the same model used by Banks et al. [5] and Tague-Sutcliffe and Blustein [56] to analyse TREC data.

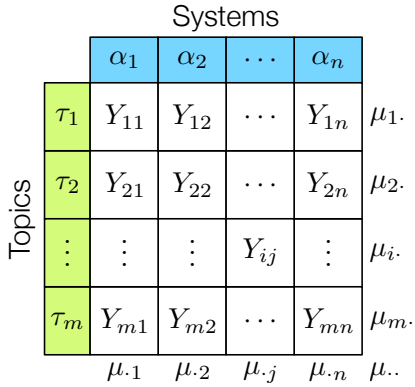


Fig. 1. Two-way ANOVA model for the topic and system effects.

$$Y_{ij} = \mu_{\cdot\cdot} + \tau_i + \alpha_j + \varepsilon_{ij} \tag{1}$$

where: $\mu_{\cdot\cdot}$ is the grand mean; τ_i with $i = 1, \ldots, m$ represents the effect of topics; α_j with $j = 1, \ldots, n$ represents the effect of systems; and, ε_{ij} is the residual error.

The model of equation (1) has the following estimators:

- grand mean

$$\hat{\mu}_{\cdot\cdot} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} Y_{ij}$$

- topic marginal mean and topic effect

$$\hat{\mu}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$$
$$\hat{\tau}_i = \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}$$

- system marginal mean and system effect

$$\hat{\mu}_{\cdot j} = \frac{1}{m} \sum_{i=1}^{m} Y_{ij}$$
$$\hat{\alpha}_j = \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$$

Therefore, the score predicted by the model is

$$\hat{Y}_{ij} = \hat{\mu}_{\cdot\cdot} + \hat{\tau}_i + \hat{\alpha}_j = \hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$$

and the prediction error is:

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})$$

3.2 Assessment of the Model

We can compute the *Sum of Squares (SS)*, *Degrees of Freedom (DF)*, *Mean Squares (MS)* and F statistics as follows:

- total effects

$$\begin{aligned}
\mathrm{SS}_{\mathrm{Total}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \hat{\mu}_{..})^2 \\
\mathrm{df}_{\mathrm{Total}} &= mn - 1 \\
\mathrm{MS}_{\mathrm{Total}} &= \frac{\mathrm{SS}_{\mathrm{Total}}}{\mathrm{df}_{\mathrm{Total}}}
\end{aligned}$$

- topic effects

$$\begin{aligned}
\mathrm{SS}_{\mathrm{Topic}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\tau}_i^2 = n\sum_{i=1}^{m}\hat{\tau}_i^2 = n\sum_{i=1}^{m}(\hat{\mu}_{i\cdot} - \hat{\mu}_{..})^2 \\
\mathrm{df}_{\mathrm{Topic}} &= m - 1 \\
\mathrm{MS}_{\mathrm{Topic}} &= \frac{\mathrm{SS}_{\mathrm{Topic}}}{\mathrm{df}_{\mathrm{Topic}}} \\
\mathrm{F}_{\mathrm{Topic}} &= \frac{\mathrm{MS}_{\mathrm{Topic}}}{\mathrm{MS}_{\mathrm{Error}}}
\end{aligned}$$

- system effects

$$\begin{aligned}
\mathrm{SS}_{\mathrm{System}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\alpha}_j^2 = m\sum_{j=1}^{n}\hat{\alpha}_j^2 = m\sum_{j=1}^{n}(\hat{\mu}_{\cdot j} - \hat{\mu}_{..})^2 \\
\mathrm{df}_{\mathrm{System}} &= n - 1 \\
\mathrm{MS}_{\mathrm{System}} &= \frac{\mathrm{SS}_{\mathrm{System}}}{\mathrm{df}_{\mathrm{System}}} \\
\mathrm{F}_{\mathrm{System}} &= \frac{\mathrm{MS}_{\mathrm{System}}}{\mathrm{MS}_{\mathrm{Error}}}
\end{aligned}$$

- error effects

$$\begin{aligned}
\mathrm{SS}_{\mathrm{Error}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\varepsilon}_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{..})\right)^2 \\
\mathrm{df}_{\mathrm{Error}} &= (m-1)(n-1) \\
\mathrm{MS}_{\mathrm{Error}} &= \frac{\mathrm{SS}_{\mathrm{Error}}}{\mathrm{df}_{\mathrm{Error}}}
\end{aligned}$$

Note that:

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Topic}} + \mathrm{SS}_{\mathrm{System}} + \mathrm{SS}_{\mathrm{Error}}$$

We can then compute the critical value for the F statistics of a factor, i.e. $\mathrm{F}_{crit} = \mathrm{F}_{\left(\mathrm{df}_{fact}, \mathrm{df}_{err}\right)}$, and determine its significance if $\mathrm{F}_{fact} > \mathrm{F}_{crit}$; this allows us also to obtain the p-value for that factor.

3.3 Effect Size

Since the F statistics tend to increase and the p-value tends to decrease as the sample size increases, we also consider the *effect size* of a factor, which accounts for the amount of variance explained by the model, by means of an unbiased estimator [42, 48]:

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

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 also how much it matters.

The common rule of thumb [47] 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 is considered as zero.

4 METHODOLOGY

A typical IR experimental design measures systems against a set of topics applied to a whole document collection. Inspired by past work [20, 60], our design measures systems against a set of topics and a set of shards, producing replicates for each topic*system pair.

We use GLMM [47] to provide a model that explains a performance score (dependent variable) in terms of the factors (independent variables) we are interested in: topic, system, and shard effects together with their interactions. We use ANOVA to determine when dependent variables differ and what proportion of variation can be attributed to differences between specific factors. The design is a *crossed factorial repeated measure*, since every level of one factor is measured in combination with every level of the other factors and each subject participates in all experimental factors.

Formally, the model is:

$$Y_{ijk} = \mu_{...} + \underbrace{\tau_i + \alpha_j + \beta_k}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{3}$$

where: Y_{ijk} is the effectiveness score of the i-th subject (topic) in the j-th (system) and k-th (shard) factors; $\mu_{...}$ is the grand mean; $\tau_i = \mu_{i..} - \mu_{...}$ is the effect of the i-th subject, i.e. a topic, where $\mu_{i..}$ is the marginal mean of the i-th subject; $\alpha_j = \mu_{.j.} - \mu_{...}$ is the effect of the j-th factor, i.e. a system, where $\mu_{.j.}$ is the marginal mean of the j-th factor; $\beta_k = \mu_{..k} - \mu_{...}$ is the effect of the k-th factor, i.e. a shard, where $\mu_{..k}$ is the marginal mean of the k-th factor; $(\tau\alpha)_{ij}$ is the interaction between topics and systems; $(\tau\beta)_{ik}$ is the interaction between topics and shards; $(\alpha\beta)_{jk}$ is the interaction between systems and shards; finally, ε_{ijk} is the error committed by the model in predicting Y_{ijk}.

4.1 Assumptions

ANOVA is based on the following assumptions [35]:

- 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 [28, p. 205] observes that "the F-test is found to be remarkably insensitive to general nonnormality. In the commonly occurring case where the group sample sizes are equal, it is not very sensitive to heterogeneity of variance from group to group". Similarly, Mendenhall and Sincich [40] 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 unequal variances in the case of balanced design. On the other hand, violation of the third assumption may severely impact the F-test and hamper the drawn conclusions as noted, for example, by Scariano and Davenport [54].

IR performance scores are known to violate the first two ANOVA assumptions [12, 56]. Tague-Sutcliffe and Blustein [56] 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 noted very few differences in the analysis conducted on the transformed data and they decided to stay with the

untransformed scores, which are more easily interpretable. Carterette [12] observed that both the first two assumptions are violated because of performance scores typically being bounded in [0 1]; however Carterette concluded that ANOVA is robust to the kind of violations of normality due to IR performance scores and that also the violations of homoskedasticity have a fairly limited impact, in agreement with previous findings of Tague-Sutcliffe and Blustein [56].

In our case, we deal with tens of topics, tens or hundreds of systems, and tens of shards which are a fairly large sample size when it comes to robustness to violations of the normality assumption. Moreover, we adopt a balanced design where group sample sizes are equal, as discussed in the next section, and this limits the impact of violations of homoskedasticity. Finally, when it comes to violations of the independence, topics, systems, and shards are independent from each other.

4.2 Shards with no relevant documents

When sharding a collection, the relevance judgements (*qrels*) are also sharded. Because only a subset of documents are judged, some shards may not contain relevant documents for certain topics. In such cases, the score for all systems on that (topic, shard) pair is undefined. In ANOVA, missing data leads to an unbalanced design, which are challenging to estimate, e.g. SS may become dependent on the order in which the factors are entered into the model[3]. Moreover, as explained in the previous section, balanced designs are preferable because they are more robust to violations of the homoskedasticity assumption.

Ferro and Sanderson [20] obtained a balanced design by using only topics for which there were relevant documents in all shards, which reduced the number of available topics to less than half. The number of shards Ferro and Sanderson split their collections into was limited to four. As we wish to create more fine-grained shards, almost no topics were likely to exist with relevant documents in all shards. Voorhees et al. [60] formed random shards, which were resampled until all shards had relevant documents for all the topics. However, this is not feasible because most of our shards are not random and cannot be resampled, until you get at least one relevant document per shard. For example, when we group Web pages by their *Top Level Domain (TLD)*, we cannot change the shape of a shard even if it has no relevant document. Moreover, also in the case of Voorhees et al., the number of shards was limited to a maximum of five, where it is still likely to find a sample where all shards have at least one relevant document. However, when we explore different types of random shards, we consider several tens of shards and it is almost impossible to find a sample where all shards have at least one relevant document. In addition, the idea itself of resampling might be questioned because it introduces some bias and departs from randomness, even more when the number of shards is high and you risk to deterministically look for the configuration of shards, if any, where there is at least one relevant document per shard.

For these reasons, we decided to simulate a balanced design by substituting the undefined values with zeros. The effect of this choice was to lower average system effectiveness and to introduce a bias that topics with substituted zeros are treated the same as 'real' zeros, which are assigned to poor systems not retrieving relevant documents.

We explored five alternatives to substituting zeros for a (topic, shard) pair: substituting the median value of a system's overall effectiveness score; substituting the mean score; substituting the lower quartile of its score; substituting the upper quartile of its score; and, substituting with ones. We compared the impact of these different substitutions on our measurement scheme. We found no notable deviations with respect to either previous work or the interpolation alternatives. We further tested our balanced design comparing it with a reproduction of the balanced design Ferro and Sanderson used. We compared ANOVA tables from the two designs and found that the

[3]The literature describes methods to constrain SS, but with no agreement on the best.

statistical significance of the different effects was unchanged across the two tables. Similarly the general trends for the effect sizes were the same. Details about the evaluation measures used in the experiments are provided in Section 5 while details about the validation of our methodology are provided in Section 6.

Therefore, we use the "zeros" method for obtaining a balanced design.

5 EXPERIMENTAL SETUP

We describe the collections, shards, and measures used. Note, to ease the reproducibility of the experiments, code can be found at: https://bitbucket.org/frrncl/tois2018-fks/.

We used the following collections from TREC:

- **Adhoc track T08 [59]**: 528,155 documents of the TIPSTER disks 4-5 corpus minus congressional record; T08 provides 50 topics, each with binary relevance judgments and a pool depth of 100 (i.e. the top 100 results from each submitted system run were collected and judged by relevance assessors); 129 system runs were submitted to T08.
- **Web tracks T09 [25] and T10 [26]**: 1,692,096 documents of the WT10g Web corpus; T09 and T10 provide 50 topics each with graded relevance judgments – i.e. not relevant, relevant, highly relevant – and a pool depth of 100; 104 and 97 system runs were each submitted to T09 and T10, respectively.
- **Terabyte tracks T13 [14], T14 [15], and T15 [9]**: 25,205,179 documents of the GOV2 Web corpus; T13, T14, and T15 provide 50 topics each (note 49 are used in T13) with graded relevance judgments (three levels) and pool depths of 85, 100 and 50 documents, respectively; 70, 58, and 80 system runs were submitted to T13, T14, and T15, respectively.

T08 is used for validating our methodology, as described in Section 6, while T09, T10, T13, T14, and T15 are used for running the experiments, as described in Section 7.

In order to split run systems and pools across shards, we take the original systems and qrels over the whole collection and, for each topic, we keep only the documents present in the given shard, discarding all the others, which is the same procedure used in [20, 60]. For all the collections, we discard the runs in the lower quartile of performance in terms of *Average Precision (AP)* over the whole collection, to avoid excessive variance due to poorly performing runs.

We used measures for both binary and graded relevance:

- **Binary:** *Average Precision (AP)* [8], *Precision at k* (P@k) [10], *Rprec* [2], and *Rank-Biased Precision (RBP)* [41]. We calculated RBP by setting $p = 0.8$ representing a trade-off between a persistent and a remitting user. For T09, T10, T13, T14, and T15, we perform a lenient mapping of the relevance judgments by considering as relevant both highly relevant and relevant documents.
- **Graded:** *normalized Discounted Cumulated Gain (nDCG)* [30], *Expected Reciprocal Rank (ERR)* [13], and *Twist* [23]. For T07 we calculate nDCG using binary relevance by setting gain to 0 for non-relevant documents and to 5 for relevant. For T09, T10, T13, T14, and T15 we assign a weight of 0 to non-relevant documents, 5 to the relevant, and 10 to highly relevant. We use a log_{10} discounting function, which accounts for a reasonably persistent user, i.e. a user willing to scan at least the first 10 documents in the result list.

For the cut-offs of Precision and nDCG we considered the following values $k = 5, 10, 20, 50, 100$. We consider a confidence level $\alpha = 0.05$ to determine if a factor is statistically significant.

6 VALIDATION OF THE MODEL

First, in Section 6.1, we incrementally build the sophistication of a GLMM model from that used in past work [5, 56] to our proposed model as detailed in eq. (3), as it is typically done with step-wise

procedures for GLMM [47] where linear terms are introduced first then followed by interaction terms. In this way, we create a set of nested models, i.e. models where it is progressively added one more terms, where the simpler model is called *reduced (or restricted) model* and the more complex one is called *full (or complete) model*.

As suggested by Rutherford [47], we compare this set of nested models using the following F-statistic

$$F_{model} = \frac{{SS_{\text{Error,R}} - SS_{\text{Error,F}}}/{df_{\text{Error,R}} - df_{\text{Error,F}}}}{{SS_{\text{Error,F}}}/{df_{\text{Error,F}}}} \tag{4}$$

where $SS_{\text{Error,R}}$ and $SS_{\text{Error,F}}$ are the SS of the error in the reduced and full model, respectively; similarly, $df_{\text{Error,R}}$ and $df_{\text{Error,F}}$ are the DF of the error in the reduced and full model. This F-statistic is compared against the critical value $F_{crit} = F_{\left(df_{\text{Error,R}} - df_{\text{Error,F}}, df_{\text{Error,F}}\right)}$, and determine its significance if $F_{model} > F_{crit}$; this allows us also to obtain the p-value for that model.

Moreover, we validate the normality and homoskedasticity ANOVA assumptions. For the former, we use the Jarque-Bera normality test [29], which is known to be robust also to lack of homoskedasticity of the data [62]. For the latter, we use the Levene test [37].

Next, in Section 6.2, we validate the way a balanced design is achieved with respect to previous results by Ferro and Sanderson [20] on the same datasets. In particular, we compare the approach of Ferro and Sanderson who keep only the topics for which there is at least one relevant document in all the shards with respect to our approach to achieve a balanced design by substituting zeros for the missing data.

Then, in Section 6.3, we validate the way a balanced design is achieved with respect to possible alternative interpolation values instead of zeros for the missing data.

Finally, in Section 6.4, we validate the choice of discarding the systems in the lower quartile of performance.

To ease comparability with previous works, e.g. [20], we use the `T08` collection and the `TIPSTER_DS` shards, made according to the four `TIPSTER` sub-corpora by document source, namely Foreign Broadcast Information Service (`TIPFBIS`, 130,471 documents); Federal Register (`TIPFR`, 55,630 documents); Financial Times (`TIPFT`, 210,158 documents); and Los Angeles Times (`TIPLA`, 131,896 documents).

For brevity, we report detailed discussion only with AP but trends for the other measures are similar. For the same reasons, results on the other experimental collections mentioned in Section 5 are not reported even if they exhibit similar trends.

6.1 Building the model

A classic IR GLMM model considers only the topic (τ_i) and system (α_j) main effects:

$$Y_{ij} = \mu_{..} + \tau_i + \alpha_j + \varepsilon_{ij} \quad . \tag{5}$$

Note that this model is the same as the one by Banks et al. [5] and Tague-Sutcliffe and Blustein [56] but we apply it on the sharded data while both Banks et al. and Tague-Sutcliffe and Blustein apply it on the whole corpus instead.

Table 1 shows the ANOVA table for eq. (5). We observe that both the topic and the system effects are statistically significant: the former is a large effect while the latter is a small effect. We also note the model of eq. (5) leaves a large share of variance unexplained since the SS of the error is 946.99 with respect to a total SS of 1524.53, i.e. most variance is unexplained. As expected and previously discussed in Section 4.1, the Jarque-Bera test show that the normality assumption is not met; similarly, the Levene test shows that the homoskedasticity assumption is not met for both the topic and system effect.

Table 1. Reduced model of eq. (5). Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 2.449e-27.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	556.8975	49	11.3653	228.6854	0	0.3675
System	20.6407	95	0.2173	4.3718	4.883e-41	0.0164
Error	946.9991	19055	0.0497			
Total	1524.5373	19199				

Table 2. Reduced model of eq. (6). Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 8.572e-12. p-value of the Levene test for homoskedasticity of Shard effect: 3.469e-12.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	556.8975	49	11.3653	272.9818	0	0.4097
System	20.6407	95	0.2173	5.2186	1.26e-54	0.0204
Shard	153.7930	3	51.2643	1231.3168	0	0.1612
Error	793.2062	19052	0.0416			
Total	1524.5373	19199				

We add to the model the shard (β_k) main effect:

$$Y_{ijk} = \mu... + \tau_i + \alpha_j + \beta_k + \varepsilon_{ijk} \quad . \tag{6}$$

Table 2 shows the ANOVA table for eq. (6). The added shard effect is statistically significant, large size, and bigger than the system effect. As expected, the SS of the topic and system factors are the same while the SS of the error is reduced by the amount of the SS of the shard factor; similarly, the DF of the topic and system factors are the same while the DF of the error is slightly decreased by the amount of the DF of the shard factor. As a consequence, the MS of the error is decreased since it is given by the ratio between the SS (decreased) and the DF (almost the same). Therefore, the F-statistics of the topic and system factors are slightly increased, since the F-statistic is given by the ratio of the MS of the factor (the same) and MS of the error (reduced); this also increases the effect size of the factor, since $\hat{\omega}^2_{\langle fact \rangle}$ is increasing in the F-statistic.

Therefore, introducing the shard main effect has the positive outcome of increasing the amount of explained variance and decreasing the error which, in turn, also improves the estimation of the other factors. This is also supported by the validation of the model of eq. (6) with respect to the one of eq. (5). The F_{model} is 1231.31 with respect to a $F_{crit} = 2.60$, thus significantly different from the previous model with a p-value of 0. However, the SS of the error is still quite large (793.20) with respect to the total SS (1524.53), indicating that most of the variance is still unexplained. Finally, the Levene test shows that the homoskedasticity assumption is not met also for the shard effect.

We next add the interaction effects, starting from the system*shard interaction effect ($(\alpha\beta)_{jk}$)

$$Y_{ijk} = \mu... + \tau_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \quad . \tag{7}$$

Table 3 shows the ANOVA table for eq. (7). We can observe that, in this model, the system*shard interaction effect is not statistically significant and that it explains just a small portion of the variance, leaving the error still quite large. The fact that, in the model of eq. (7), the system*shard interaction effect is not significant may mean that this factor is actually not significant but it may also be a by-product of the large share of variance still not explained. The fact that this model is

Table 3. Reduced model of eq. (7). Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 7.143e-08. p-value of the Levene test for homoskedasticity of Shard effect: 1.325e-13.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	556.8975	49	11.3653	271.9507	0	0.4088
System	20.6407	95	0.2173	5.1989	2.73e-54	0.0204
Shard	153.7930	3	51.2643	1226.6660	0	0.1607
System*Shard	8.9032	285	0.0312	0.7475	0.9995	–
Error	784.3029	18767	0.0418			
Total	1524.5373	19199				

not able to more satisfactory explain the data is further supported by the validation of the model of eq. (7) with respect to the one of eq. (6). The F_{model} is 0.74 with respect to a $F_{crit} = 1.14$, thus not significantly different from the previous model with a p-value of 0.99. Note that this is the model proposed by Ferro and Sanderson [20].

We add the topic*system effect ($(\tau\alpha)_{ij}$), which is considered a quite prominent effect in the literature [5, 45, 60]

$$Y_{ijk} = \mu_{...} + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_{ij} + (\alpha\beta)_{jk} + \varepsilon_{ijk} \quad . \tag{8}$$

Table 4 shows the ANOVA table for eq. (8). The topic*system interaction effect turns out to be not statistically significant and this seems to contradict previous findings in the literature. However, this is mostly a by-product of this reduced model which does not yet explain the data in a fully satisfactory way, as we will further discuss later on when we will add also the topic*shard interaction effect. Indeed, the SS of the topic*system effect (115.74) is the third biggest after the SS of the topic effect (556.89) and the shard effect (153.79). However, the SS of the error is still quite big (668.55) with respect to the total SS (1524.53), even if the topic*system effect helps in explaining a good share of variance. At the same time, the topic*system effect is responsible for the highest DF (4655) and this substantially reduces the DF of the error (14112). As a consequence, despite the more variance explained, the MS of the error increases instead of descresing, because the decrease in the SS of the error is counter-balanced by the more marked decrease in its DF. Therefore, the F-statistics of all the other effects are slightly smaller than in the case of the model without the topic*system interaction (Table 3) and the same happens for their effect sizes, leading also to a not significant topic*system interaction. The fact that this model is not able to more satisfactory explain the data is further supported by the validation of the model of eq. (8) with respect to the one of eq. (7). The F_{model} is 0.52 with respect to a $F_{crit} = 1.03$, thus not significantly different from the previous model with a p-value of 1.00.

Finally, Table 5 shows the ANOVA table for the full model of eq. (3) where also the topic*shard interaction effect is taken into consideration. The topic*shard interaction effect is a large statistically significant effect. It explains a good share of the variance: SS (488.75) the second biggest after the topic effect (556.89). This substantially reduces the SS of the error (179.80), which now suggests that most of the variance is explained by the model, being the total SS 1524.53. The MS of the error is substantially reduced and this considerably increases the F-statistics of all the other factors and their effect sizes as well. As a consequence, the system*shard and topic*system interaction effects are now statistically significant, the former being a small size effect and the latter being a large size effect, as expected from previous findings in the literature. Overall, this suggests that the topic*shard interaction effect is not only a remarkably big effect but it may shadow the other interaction effects if not properly taken into account. These observations are further supported by

Table 4. Reduced model of eq. (8). Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 4.582e-05. p-value of the Levene test for homoskedasticity of Shard effect: 7.054e-09.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	556.8975	49	11.3653	239.9002	0	0.3788
System	20.6407	95	0.2173	4.5862	3.375e-44	0.0174
Shard	153.7930	3	51.2643	1082.0985	0	0.1445
System*Shard	8.9032	285	0.0312	0.6594	1	–
Topic*System	115.7480	4655	0.0249	0.5249	1	–
Error	668.5549	14112	0.0474			
Total	1524.5373	19199				

Table 5. Full model of eq. (3). Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 1.315e-143. p-value of the Levene test for homoskedasticity of Shard effect: 7.544e-06.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	556.8975	49	11.3653	882.7220	0	0.6923
System	20.6407	95	0.2173	16.8751	6.53e-256	0.0728
Shard	153.7930	3	51.2643	3981.6227	0	0.3835
System*Shard	8.9032	285	0.0312	2.4263	1.594e-34	0.0207
Topic*System	115.7480	4655	0.0249	1.9313	1.377e-183	0.1842
Topic*Shard	488.7522	147	3.3248	258.2357	0	0.6632
Error	179.8027	13965	0.0129			
Total	1524.5373	19199				

the validation of the full model of eq. (3) with respect to the reduced models. If we compare it to the immediately previous reduced model of eq. (8), the F_{model} is 258.23 with respect to a $F_{crit} = 1.20$, thus significantly different from the previous model with a p-value of 0. If we compare it to eq. (6), i.e. the model where only the main effects are considered but not the interaction ones, the F_{model} is 9.36 with respect to a $F_{crit} = 1.03$, thus significantly different from the previous model with a p-value of 0.

Therefore, the full model in eq. (3) allows us to explain most of the variance and to have a more accurate estimation of effect size and when they are statistically significant. To the best of our knowledge, this is the first time that a topic*shard interaction has been explicitly measured in IR system experiments.

6.2 Validation of the Balanced Design

To further test our design, we reproduce the design of Ferro and Sanderson [20] on the same dataset they used so as to compare it with our way of achieving a balanced design. Note that the reduced model of eq. (7) is the same as the model of eq. (2) of Ferro and Sanderson.

Table 6 shows the ANOVA table for the reduced model of eq. (7), computed as in [20], i.e. keeping only the topics with relevant documents across all the shards (15 out of 50)[4].

[4]Note that the p-values and the $\hat{\omega}^2_{\langle fact \rangle}$ of Table 6 are not exactly the same as those reported in the table of Figure 2(c) of Ferro and Sanderson [20]. The difference is due to the fact that Ferro and Sanderson use all the runs of `T08` while we discard the lowest quartile performing runs. If we use all the runs, we obtain exactly the same scores as Ferro and Sanderson but we

Table 6. Reduced model of eq. (7) computed as in Ferro and Sanderson [20]. This table should be compared to Table 3. Systems up to third quartile (top 75%) of performance. Balanced design obtained by keeping only the 15 topics which have at least one relevant document in each shard. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 0.9812. p-value of the Levene test for homoskedasticity of Shard effect: 8.69e-30.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	192.8895	14	13.7778	363.5912	0	0.4685
System	8.8407	95	0.0931	2.4558	3.017e-13	0.0234
Shard	25.3043	3	8.4348	222.5905	4.247e-136	0.1035
System*Shard	6.0021	285	0.0211	0.5558	1	–
Error	203.1861	5362	0.0379			
Total	436.2228	5759				

Comparing Table 6 (Ferro and Sanderson's way of obtaining a balanced design) to Table 3 (our way of obtaining a balanced design), we can observe that the statistical significance of the different effects is unchanged in the two tables as well as the general trends for the effect sizes.

However, differences can be seen in the size of the topic effect which is slightly smaller than in previous work, possibly due to smaller number of topics (15) used in Table 6 with respect to 50 topics used in Table 3. The system effect is just slightly smaller than in prior work. This is an expected by-product of using zeros where no relevant documents are found, because this lowers the average performance of all systems and reduces the performance variance across systems. That is, the different systems behave more similarly in our new design methodology compared to that of prior work and less of the variance is explained by system differences, thus producing a smaller effect size.

The zeros affect the shard effect sizes as well. Our new methodology reports stronger shard effects than prior work. The newly added zeros created a larger variance in individual shard performances. In the prior method, shards without relevant documents were excluded from analysis, meaning that there were fewer shards that mostly produced non-zero accuracy results. In our new methodology, shards without relevant documents always produce zero effectiveness, increasing both the number and the range of effectiveness scores produced, thereby increasing the performance variance across shards. That is, shards behave more differently under our new methodology and thus larger effect sizes are reported.

Table 7 shows the ANOVA table for the full model of eq. (3), computed as in [20], i.e. keeping only the topics with relevant documents across all the shards (15 out of 50). Note that Ferro and Sanderson did not compute this model but we do it to see what happens if we compute our full model using only the topics with at least one relevant document in each shard instead of interpolating missing values with zero.

We can see a similar behaviour as the one observed in the previous case: the trends for the significance and size of the effects are similar between Table 7 and Table 5. We can again observe a small decrease in the topic effect size, possibly due to using less topics in the Ferro and Sanderson's settings; the size of the systems effect size is almost the same, almost negligibly smaller in our approach, possibly due to the interpolation with zeros; the shard effect size is bigger in our approach, due to the interpolation with zeros; the interaction effects are bigger in our approach due to the more topics used with respect to Ferro and Sanderson and to the interpolation with zeros.

prefer to report in Table 6 the scores without considering the lowest quartile runs because this is consistent with all the other figures reported in this paper.

Table 7. Full model of eq. (3) computed as in Ferro and Sanderson [20]. This table should be compared to Table 5. Systems up to third quartile (top 75%) of performance. Balanced design obtained by keeping only the 15 topics which have at least one relevant document in each shard. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 1.46e-144. p-value of the Levene test for homoskedasticity of System effect: 2.871e-59. p-value of the Levene test for homoskedasticity of Shard effect: 4.523e-20.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	192.8895	14	13.7778	865.9267	0	0.6777
System	8.8407	95	0.0931	5.8488	6.083e-61	0.0740
Shard	25.3043	3	8.4348	530.1203	6.024e-290	0.2160
System*Shard	6.0021	285	0.0211	1.3236	0.0003466	0.0158
Topic*System	33.6924	1330	0.0253	1.5921	2.359e-27	0.1203
Topic*Shard	106.0086	42	2.5240	158.6325	0	0.5348
Error	63.4852	3990	0.0159			
Total	436.2228	5759				

If we compare the full model of eq. (3) to the reduced model of eq. (7), both computed as in Ferro and Sanderson, the F_{model} is 6.39 with respect to a $F_{crit} = 1.07$ with a p-value of 0. Therefore, also in this case, the full model turns out to be able to explain variance significantly better than the reduced model.

Overall, this analysis confirms that both the shard and the topic*shard effect size and significance are not a by-product of the choice of interpolating missing values but they are prominent effects instead. Indeed, they are present also when you restrict yourself only to topics which have relevant documents across all the shards. At the same time, it also emerges that the size and significance of the other effects – topics, system, topic*system, and system*shard – are not hampered or severely distorted by the choice of interpolating missing values.

Overall, the balanced design methodology used by Ferro and Sanderson and that of this paper both have advantages and drawbacks. The methodology of prior work eliminates undefined data and thus only work with topics and shards where there is explicit evidence. However, due to eliminating topics, there is less data overall and the remaining topics may be biased. Our methodology's use the full range of available topics and is able to study a wider variety of shard configurations, which in the prior work were artificially removed somehow biasing the estimation of the shard effect. However, our methodology cannot distinguish between a true zero effectiveness score and a lack of relevant documents. The different choices we made to achieve a balanced design produced differences in the effect sizes reported from prior work, still keeping a reasonable agreement in terms of general trends and what is statistically significant.

6.3 Validation of Balanced Design wrt Alternative Interpolation Values

As previously explained, we substitute missing data with zeros but it would be possible to resort to data interpolation techniques to determine alternative values which are smoother than zero.

We explore five alternatives for data interpolation: (i) `lq` makes use of the score's lower quartile value; (ii) `med` makes use of the median value of an effectiveness score; (iii) `mean` makes use of the score's mean; and, (iv) `uq` makes use of the score's upper quartile value; and, (v) one interpolates missing values with 1. Since some shards may generate missing relevant documents for all the topics, we determine these interpolation value across all the topics, all the systems, and all the shards, i.e. by putting all the available data together.

Tables 8 to 12 show the ANOVA table for the full model of eq. (3) for the `lq`, `med`, `mean`, `uq`, and `one` interpolation alternatives, respectively.

Table 8. Full model of eq. (3) with `1q` interpolation. This table should be compared to Table 5. Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to the lower quartile value, balancing value: 0.03243. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 1.315e-143. p-value of the Levene test for homoskedasticity of Shard effect: 7.544e-06.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	548.3131	49	11.1901	869.1152	0	0.6890
System	20.6407	95	0.2173	16.8751	6.53e-256	0.0728
Shard	130.1532	3	43.3844	3369.6008	0	0.3448
System*Shard	8.9032	285	0.0312	2.4263	1.594e-34	0.0207
Topic*System	115.7480	4655	0.0249	1.9313	1.377e-183	0.1842
Topic*Shard	456.2990	147	3.1041	241.0888	0	0.6477
Error	179.8027	13965	0.0129			
Total	1459.8599	19199				

Table 9. Full model of eq. (3) with `med` interpolation. This table should be compared to Table 5. Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to the median value, balancing value: 0.1801. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 1.315e-143. p-value of the Levene test for homoskedasticity of Shard effect: 7.544e-06.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	524.0837	49	10.6956	830.7098	0	0.6792
System	20.6407	95	0.2173	16.8751	6.53e-256	0.0728
Shard	50.7038	3	16.9013	1312.6962	0	0.1701
System*Shard	8.9032	285	0.0312	2.4263	1.594e-34	0.0207
Topic*System	115.7480	4655	0.0249	1.9313	1.377e-183	0.1842
Topic*Shard	353.1083	147	2.4021	186.5672	0	0.5869
Error	179.8027	13965	0.0129			
Total	1252.9904	19199				

Table 10. Full model of eq. (3) with `mean` interpolation. This table should be compared to Table 5. Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to the mean value, balancing value: 0.2676. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 1.315e-143. p-value of the Levene test for homoskedasticity of Shard effect: 7.544e-06.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	521.2183	49	10.6371	826.1680	0	0.6780
System	20.6407	95	0.2173	16.8751	6.53e-256	0.0728
Shard	25.4066	3	8.4689	657.7632	0	0.0931
System*Shard	8.9032	285	0.0312	2.4263	1.594e-34	0.0207
Topic*System	115.7480	4655	0.0249	1.9313	1.377e-183	0.1842
Topic*Shard	326.4212	147	2.2206	172.4669	0	0.5676
Error	179.8027	13965	0.0129			
Total	1198.1406	19199				

Comparing these tables to Table 5, we can observe that the statistical significance and the trends in the effect sizes are basically unchanged. As it can be expected, the shard effect is the most affected by the change in the interpolation strategy. Indeed, it decreases as the interpolation values increase passing from `1q`, to `med`, and to `mean` since its variation gets reduced and closer to the central part

Table 11. Full model of eq. (3) with up interpolation. This table should be compared to Table 5. Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to the upper quartile value, balancing value: 0.4186. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 1.315e-143. p-value of the Levene test for homoskedasticity of Shard effect: 7.544e-06.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	536.4064	49	10.9471	850.2422	0	0.6843
System	20.6407	95	0.2173	16.8751	6.53e-256	0.0728
Shard	19.9587	3	6.6529	516.7207	1.849e-318	0.0746
System*Shard	8.9032	285	0.0312	2.4263	1.594e-34	0.0207
Topic*System	115.7480	4655	0.0249	1.9313	1.377e-183	0.1842
Topic*Shard	340.7818	147	2.3182	180.0545	0	0.5782
Error	179.8027	13965	0.0129			
Total	1222.2415	19199				

Table 12. Full model of eq. (3) with one interpolation. This table should be compared to Table 5. Systems up to third quartile (top 75%) of performance. Balanced design where missing values are forced to one. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 1.315e-143. p-value of the Levene test for homoskedasticity of Shard effect: 7.544e-06.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	832.7732	49	16.9954	1320.0046	0	0.7710
System	20.6407	95	0.2173	16.8751	6.53e-256	0.0728
Shard	450.2567	3	150.0856	11656.9168	0	0.6455
System*Shard	8.9032	285	0.0312	2.4263	1.594e-34	0.0207
Topic*System	115.7480	4655	0.0249	1.9313	1.377e-183	0.1842
Topic*Shard	1109.7591	147	7.5494	586.3490	0	0.8176
Error	179.8027	13965	0.0129			
Total	2717.8835	19199				

of the score distribution; it then increases again with uq and one. The topic effect tends to very slightly decrease moving from lq, to med, and to mean while it increases back with uq and one, even if these changes are not appreciably different. The topic*shard interaction effect follows a similar pattern, slightly decreasing and then increasing, and it confirms to be one of the most prominent effects under all circumstances, i.e. whatever interpolation value is chosen or even when you perform no interpolation at all and keep only topics with at least one relevant document in each shard.

The system, system*shard, and topic*system are exactly the same across all these tables. Indeed, when a shard causes a topic to not have any relevant document, this affects all the systems in the same way since all of them will miss that effectiveness score. Whatever interpolation values you choose – zero, med, mean, or lq – this will cause a shift in the system marginal means which is the same for all the systems; then, the SS of these effects are basically computed as the square of the data points minus these marginal means and so they are the same because the SS re-absorb these fixed shifts equal for all the systems. On the other hand, these missing data are not spread in the same way for topics and shards and so the interpolation values do not produce a fixed equal shift for all the topics and shards; this is why the SS where topics and shards are involved change across the tables.

Table 13. Full model of eq. (3) where all systems used. This table should be compared to Table 5. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 7.379e-140. p-value of the Levene test for homoskedasticity of Shard effect: 0.001717.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	543.1558	49	11.0848	731.1896	0	0.5810
System	158.3804	128	1.2373	81.6193	0	0.2857
Shard	140.9738	3	46.9913	3099.6945	0	0.2649
System*Shard	35.8976	384	0.0935	6.1665	3.235e-261	0.0714
Topic*System	215.9815	6272	0.0344	2.2715	0	0.2361
Topic*Shard	477.8748	147	3.2508	214.4364	0	0.5488
Error	285.2500	18816	0.0152			
Total	1857.5140	25799				

Finally, the one interpolation seems to be the least appropriate one since it over-emphasises the topic, shard, and topic*shard effects, still keeping the same trends in the of significance and effect size. Indeed, it turns a possibly frequent situation where a system has no chance to retrieve any relevant document into the best performance possible, systematically increasing the spread between existing low/medium performing systems – remember that the median performance is 0.1801 (Table 9) and the mean performance is 0.2676 (Table 10) – and an artificial best possible performance 1.0. Similar considerations, even if far less marked, can be done for the uq interpolation.

As a general consideration, interpolation might not be the most appropriate technique to apply in our context anyway. Indeed, interpolation inserts smoothed values where the real information is, instead, that there are no relevant documents in a shard for a given topic. In a sense, interpolation could mask reality and smooth the shard effect too much, as we have seen happening in Tables 9 to 8 as the interpolation values increase. Indeed, our context is very different from other cases where interpolation is somehow intrinsic. Consider, for example, a time series where you miss some data points but there is some sort of continuity - e.g. if the temperature is 30 degrees and after 2 hours is 20 degrees, it is reasonable to assume that after 1 hour it should be something like 25 degrees. Instead, in our context, there is no a-priori "sensible" value for a topic without relevant documents and thus, among the possible interpolation values, a zero might be the least worst approximation of what is truly happening, especially if you consider that for some measures, like AP, no value is possible in absence of relevant documents but for other measures, like P@k, you may consider 0 as the real value to be used.

6.4 Validation of Discarding the Lower Quartile Systems

In this section, we validate the choice of using systems up to the third quartile of performance by reporting the application of the full model of eq. (3) when using all the systems (Table 13), system up to the second quartile (Table 14), and systems in the top quartile (Table 15). These tables should be compare to Table 5 where systems up to the third quartile are used.

In general, you can note how the trends in terms of significance and size of the effects are are similar to those observed in the previous cases. However, as it can be expected, reducing the number of systems considered, moving from the fourth to the first quartile, reduces the variation in this factor and, consequently, its effect size; in turn, this gives room to the other effects to become relatively bigger.

Therefore, we will use systems up to the third quartile, since this does not hamper to observed trends and, when conducting analyses, it is usually preferred to remove low performing systems

Table 14. Full model of eq. (3) where systems up to second quartile (top 50%)/median of performance are used. This table should be compared to Table 5. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 0. p-value of the Levene test for homoskedasticity of System effect: 1.601e-139. p-value of the Levene test for homoskedasticity of Shard effect: 0.000197.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	387.0483	49	7.8989	609.8103	0	0.6998
System	7.9881	63	0.1268	9.7887	2.196e-88	0.0415
Shard	121.5241	3	40.5080	3127.2797	0	0.4229
System*Shard	4.7255	189	0.0250	1.9302	6.454e-13	0.0135
Topic*System	73.7020	3087	0.0239	1.8432	1.78e-105	0.1690
Topic*Shard	348.6851	147	2.3720	183.1225	0	0.6765
Error	119.9588	9261	0.0130			
Total	1063.6318	12799				

Table 15. Full model of eq. (3) where systems up to first quartile (top 25%) of performance are used. This table should be compared to Table 5. Balanced design where missing values are forced to zero. p-value of the Jarque-Bera test for normality: 0.001. p-value of the Levene test for homoskedasticity of Topic effect: 3.204e-205. p-value of the Levene test for homoskedasticity of System effect: 6.589e-99. p-value of the Levene test for homoskedasticity of Shard effect: 0.009965.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	190.1945	49	3.8815	264.6120	0	0.6687
System	3.0925	31	0.0998	6.8008	7.158e-28	0.0273
Shard	72.6746	3	24.2249	1651.4653	0	0.4362
System*Shard	2.4459	93	0.0263	1.7930	5.717e-06	0.0114
Topic*System	41.2468	1519	0.0272	1.8511	9.11e-54	0.1681
Topic*Shard	179.0731	147	1.2182	83.0464	0	0.6533
Error	66.8454	4557	0.0147			
Total	555.5728	6399				

to avoid outliers and an excess of variance, as done for example in [3, 46, 58], which may over-emphasise some effects instead.

Overall, according to the outcomes of this thorough validation section, since we have not observed remarkable deviations with respect to either previous work or interpolation alternatives, all further experiments in the paper will use the new "zeros" method for achieving a balanced design and systems up to the third quartile of performance.

7 EXAMINING SHARD EFFECTS

With a richer GLMM model and balanced experimental design, we now focus on a wide range of shards in order to better understand the shard effects noted in the past [20, 32, 52]. We apply our model to shards created by various ways of partitioning a collection that have been employed in past work. We have already examined shards split on publication source in Section 6. We now examine shards created by partitioning the collection by the *Top Level Domain (TLD)* of the URL of a web page (Section 7.1), shards created by clustering the document collection by topic (Section 7.2), and shards created by splitting on document type (Section 7.3).

Table 16. WT10g datasets split based on their TLD on T09 and T10 for various measures. In the left column, there are uneven size shards; in the right column, there are even size shards. For each effect, its effect size $\hat{\omega}^2$ is reported. Insignificant effects are in gray; small effects, light blue; medium, blue; and large, dark blue.

Measure	T09 – WT10g_TLD						T09 – WT10g_TLD_EVEN					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.08	0.01	0.37	0.03	0.03	0.68	0.30	0.03	0.07	<1e-2	0.10	0.67
R-prec	0.06	0.01	0.29	0.02	0.02	0.61	0.19	0.01	0.04	–	0.06	0.56
RBP	0.14	0.01	0.55	0.06	0.04	0.69	0.45	0.03	0.11	<1e-2	0.17	0.73
nDCG	0.12	0.01	0.66	0.05	0.04	0.74	0.42	0.03	0.13	<1e-2	0.13	0.75
ERR	0.09	0.01	0.48	0.03	0.02	0.68	0.37	0.02	0.09	<1e-2	0.10	0.73
Twist	0.07	0.01	0.37	0.03	0.02	0.63	0.20	0.02	0.05	<1e-2	0.06	0.58
P@5	0.11	0.01	0.49	0.05	0.03	0.64	0.40	0.03	0.09	<1e-2	0.14	0.69
P@10	0.13	0.01	0.55	0.05	0.03	0.68	0.43	0.03	0.11	<1e-2	0.16	0.73
P@20	0.14	0.01	0.58	0.06	0.04	0.70	0.44	0.03	0.12	<1e-2	0.16	0.75
P@50	0.13	0.01	0.61	0.06	0.04	0.71	0.40	0.02	0.11	<1e-2	0.15	0.77
P@100	0.13	0.01	0.63	0.06	0.04	0.73	0.38	0.02	0.11	<1e-2	0.13	0.77
nDCG@5	0.07	0.01	0.36	0.03	0.02	0.62	0.28	0.03	0.07	<1e-2	0.09	0.65
nDCG@10	0.08	0.01	0.42	0.03	0.02	0.65	0.33	0.04	0.09	<1e-2	0.10	0.69
nDCG@20	0.09	0.01	0.49	0.03	0.03	0.69	0.38	0.04	0.11	<1e-2	0.12	0.73
nDCG@50	0.11	0.01	0.58	0.04	0.03	0.73	0.41	0.04	0.13	<1e-2	0.13	0.75
nDCG@100	0.12	0.01	0.62	0.05	0.04	0.73	0.41	0.04	0.13	<1e-2	0.13	0.75

Measure	T10 – WT10g_TLD						T10 – WT10g_TLD_EVEN					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.09	<1e-2	0.45	0.02	0.02	0.68	0.34	0.01	0.09	<1e-2	0.09	0.69
R-prec	0.07	<1e-2	0.35	0.01	0.01	0.59	0.23	0.01	0.06	<1e-2	0.06	0.57
RBP	0.20	0.01	0.65	0.04	0.03	0.76	0.56	0.02	0.15	<1e-2	0.13	0.77
nDCG	0.13	0.01	0.73	0.02	0.03	0.75	0.44	0.01	0.16	0.01	0.11	0.77
ERR	0.12	<1e-2	0.56	0.02	0.02	0.70	0.42	0.01	0.12	<1e-2	0.07	0.73
Twist	0.08	<1e-2	0.42	0.01	0.01	0.60	0.23	0.01	0.06	<1e-2	0.05	0.59
P@5	0.16	0.01	0.58	0.03	0.02	0.71	0.48	0.02	0.11	<1e-2	0.10	0.71
P@10	0.20	0.01	0.65	0.04	0.03	0.76	0.56	0.02	0.16	0.01	0.12	0.77
P@20	0.24	0.01	0.69	0.04	0.03	0.80	0.57	0.01	0.19	0.01	0.11	0.80
P@50	0.25	<1e-2	0.70	0.04	0.03	0.81	0.51	0.01	0.18	0.01	0.09	0.81
P@100	0.27	<1e-2	0.72	0.03	0.02	0.84	0.48	0.01	0.18	<1e-2	0.07	0.82
nDCG@5	0.08	<1e-2	0.46	0.02	0.02	0.65	0.28	0.01	0.08	<1e-2	0.07	0.66
nDCG@10	0.08	0.01	0.52	0.02	0.02	0.67	0.34	0.02	0.10	<1e-2	0.09	0.71
nDCG@20	0.10	0.01	0.60	0.02	0.02	0.70	0.39	0.02	0.13	0.01	0.10	0.74
nDCG@50	0.11	0.01	0.66	0.03	0.03	0.73	0.41	0.02	0.15	0.01	0.11	0.75
nDCG@100	0.12	0.01	0.69	0.03	0.03	0.73	0.41	0.02	0.15	0.01	0.11	0.75

7.1 Sharding by Top Level Domain

The WT10g dataset is composed of web documents, which can be sharded by their by the TLD of their URL [34, 52]: all documents with a URL ending in .org are placed in one shard, all with .gov are placed in another, etc. for a total of 59 shards. Table 16 on the left shows the results of this shard organization.

The strongest overall effect in WT10g_TLD is the topic*shard interaction, indicating that certain topic*shard combinations produce higher variance in effectiveness than others. This is expected, since the organization of the documents by TLD produce loosely topic-related shards. For example, queries related to politics might be best served by the .gov shard. The strength of the topic*shard interaction is a global trend which we discuss further in Section 8.

Table 17. WT10g datasets split randomly with different sizes on T09 and T10 for various measures. In the left column, there are randomly split and unevenly sized shards; in the right column, there are randomly split and unevenly sized shards. For each effect, its effect size $\hat{\omega}^2$ is reported. Insignificant effects are in gray; small effects, light blue; medium, blue; and large, dark blue.

Measure	T09 – WT10g_RANDOM_UNEVEN						T09 – WT10g_RANDOM_EVEN					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.13	0.01	0.35	0.02	0.05	0.68	0.34	0.04	0.03	–	0.15	0.65
R-prec	0.10	0.01	0.25	0.02	0.03	0.60	0.20	0.02	0.02	–	0.09	0.52
RBP	0.21	0.02	0.57	0.07	0.07	0.65	0.59	0.07	0.03	–	0.30	0.69
nDCG	0.20	0.02	0.67	0.04	0.06	0.74	0.52	0.05	0.05	–	0.19	0.73
ERR	0.16	0.01	0.51	0.04	0.04	0.68	0.45	0.04	0.04	–	0.15	0.72
Twist	0.11	0.01	0.31	0.02	0.03	0.62	0.22	0.02	0.02	–	0.09	0.55
P@5	0.17	0.01	0.51	0.05	0.05	0.59	0.52	0.05	0.03	–	0.25	0.62
P@10	0.20	0.01	0.59	0.07	0.06	0.63	0.62	0.07	0.03	–	0.32	0.68
P@20	0.21	0.01	0.64	0.08	0.07	0.67	0.68	0.07	0.04	–	0.37	0.72
P@50	0.21	0.01	0.68	0.09	0.06	0.71	0.69	0.07	0.04	–	0.38	0.73
P@100	0.20	0.01	0.70	0.08	0.06	0.73	0.69	0.07	0.04	–	0.38	0.73
nDCG@5	0.12	0.01	0.36	0.03	0.04	0.65	0.33	0.04	0.03	–	0.14	0.61
nDCG@10	0.13	0.01	0.45	0.03	0.04	0.67	0.43	0.06	0.04	–	0.16	0.67
nDCG@20	0.15	0.02	0.54	0.04	0.05	0.70	0.50	0.06	0.05	–	0.19	0.71
nDCG@50	0.18	0.02	0.63	0.04	0.06	0.73	0.52	0.05	0.05	–	0.19	0.72
nDCG@100	0.19	0.02	0.65	0.04	0.06	0.73	0.52	0.05	0.05	–	0.19	0.72

Measure	T10 – WT10g_RANDOM_UNEVEN						T10 – WT10g_RANDOM_EVEN					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.18	0.01	0.39	0.01	0.05	0.70	0.34	0.02	0.03	–	0.15	0.63
R-prec	0.13	<1e-2	0.27	0.01	0.03	0.60	0.22	0.01	0.01	–	0.08	0.49
RBP	0.29	0.01	0.66	0.05	0.06	0.71	0.69	0.04	0.04	–	0.26	0.69
nDCG	0.22	0.01	0.70	0.02	0.07	0.75	0.50	0.03	0.05	<1e-2	0.21	0.73
ERR	0.19	0.01	0.52	0.02	0.04	0.67	0.48	0.02	0.03	–	0.14	0.66
Twist	0.13	0.01	0.33	0.01	0.03	0.61	0.23	0.01	0.02	–	0.08	0.52
P@5	0.23	0.01	0.59	0.04	0.05	0.65	0.61	0.03	0.03	–	0.21	0.61
P@10	0.27	0.01	0.66	0.05	0.06	0.69	0.74	0.04	0.04	<1e-2	0.29	0.68
P@20	0.33	0.01	0.72	0.05	0.06	0.75	0.82	0.05	0.06	<1e-2	0.35	0.75
P@50	0.36	0.01	0.77	0.05	0.06	0.80	0.82	0.06	0.06	<1e-2	0.35	0.76
P@100	0.36	0.01	0.79	0.04	0.05	0.83	0.82	0.05	0.06	<1e-2	0.34	0.76
nDCG@5	0.15	0.01	0.41	0.02	0.04	0.66	0.27	0.02	0.03	–	0.13	0.60
nDCG@10	0.16	0.01	0.48	0.02	0.04	0.68	0.37	0.03	0.03	–	0.16	0.65
nDCG@20	0.17	0.01	0.56	0.02	0.05	0.70	0.46	0.04	0.04	<1e-2	0.20	0.70
nDCG@50	0.19	0.01	0.65	0.03	0.06	0.72	0.47	0.04	0.04	<1e-2	0.20	0.71
nDCG@100	0.20	0.01	0.66	0.03	0.06	0.73	0.47	0.04	0.04	<1e-2	0.20	0.71

The WT10g_TLD shards produce small system*shard and topic*system interaction effects and a large shard effect. The system*shard interaction indicates that certain systems are somewhat better suited for searching certain shards than other systems. The large shard effect indicates that regardless of topic or system, there are a few shards that produce the highest effectiveness and that most relevant documents across all topics appear in a few shards. This effect is stronger than the topic effect.

It is tempting to declare that organizing documents by TLD is an ideal way to sort web documents for applications such as selective search, where the goal is to minimize the number of shards accessed per query for efficiency reasons. However, a confounding factor exists in the WT10g_TLD shards;

the shard sizes are uneven. 48% of documents are in the .com shard and the top three TLDs (.com, .edu, and .org) contain 71% of all documents.

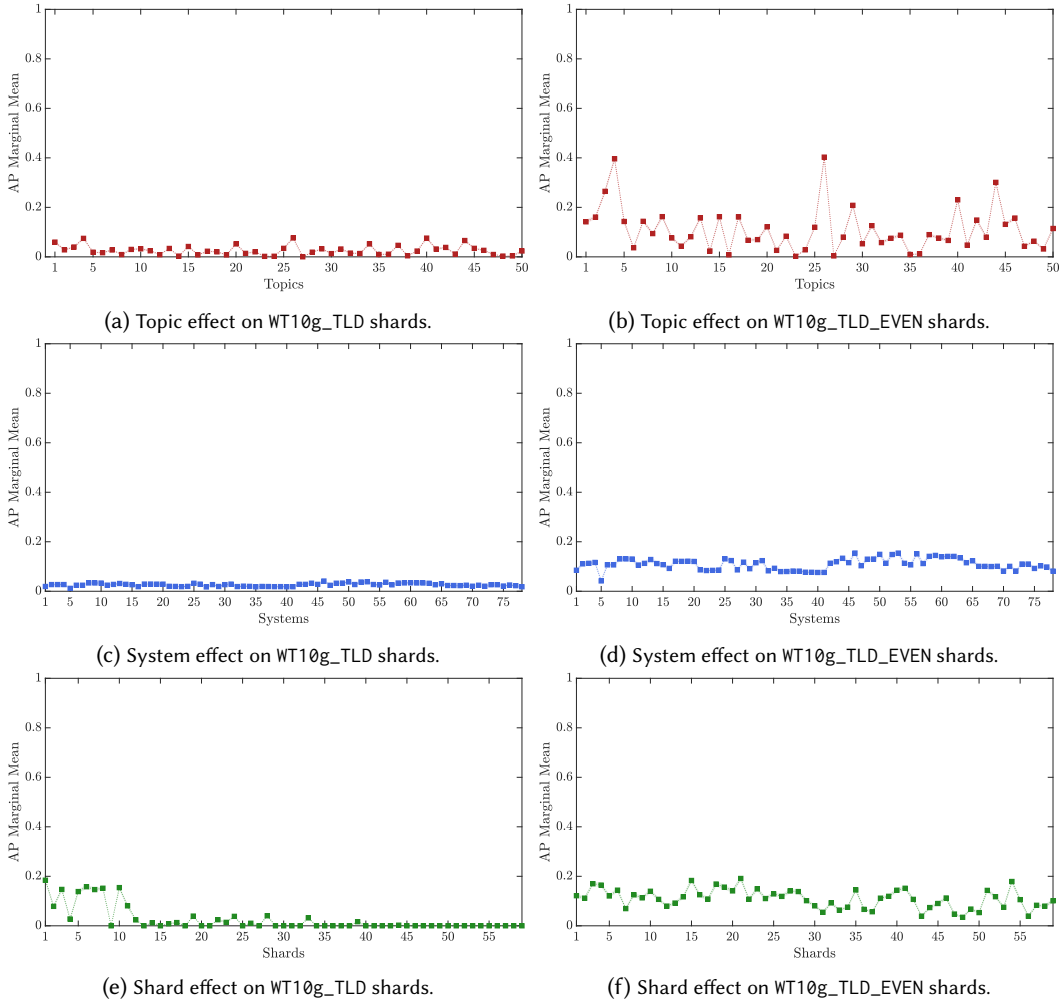
To better understand the effects of shard sizes, we created and examined three other shard sets, each with 59 shards: WT10g_TLD_EVEN (Table 16 on the right), WT10g_RANDOM_UNEVEN (Table 17 on the left), and WT10g_RANDOM_EVEN (Table 17 on the right). In WT10g_TLD_EVEN, the documents were ordered by TLD, but the shard boundaries were drawn such that each shard would contain the same number of documents; i.e. the .com shard was split into multiple shards, whereas smaller TLD shards were merged. In WT10g_RANDOM_UNEVEN, documents are randomly assigned to shards, where shard sizes are the same as WT10g_TLD; i.e. the shard sizes are skewed. In WT10g_RANDOM_EVEN documents are randomly assigned to shards of equal size.

Comparing WT10g_TLD_EVEN to WT10g_RANDOM_EVEN, we see that moving from TLD organised shards to the randomly organised shards results in the increase in the topic, system, and topic*system effects. In contrast, the shard effect in randomly organized shards drops from a combination of medium and large effects to small effects. The system*shard effects drop from small to insignificant. Therefore, organizing shards by TLD produces some shard effect, and our work confirms the observations of Sanderson et al. [52] and Kulkarni and Callan [34] with more concrete measurements.

Comparing WT10g_TLD with WT10g_RANDOM_UNEVEN, there is little difference in the measured effect size: shard effects are large; the system*shard effect is mostly small. The lack of change between WT10g_TLD and WT10g_RANDOM_UNEVEN indicates that shards heavily skewed in size create a shard effect independent of the internal organization of their constituent documents. The undesirability of size skewed shards have been noted previously for efficiency reasons [33]. However, our results suggest that they may also impede drawing accurate conclusions about the effectiveness of sharding methodologies. Therefore, when using size-skewed shards to evaluate the retrieval accuracy of a novel method as was the case in Puppin et al. [44] and Pobelte and Baeza-Yates [43], we recommend that a careful analysis should done to separate the shard effects caused by the size of the shard as opposed to the contents of the shard.

Figure 2 shows the main effect plots for the topic effects (Figure 2a and Figure 2b), system (effects Figure 2c and Figure 2d), and shard effects (Figure 2e and Figure 2f) using AP on T09 for the WT10g_TLD and WT10g_TLD_EVEN sharding schemes; the figure reports the marginal mean of AP for each of these effects. Figure 3 instead shows the interaction effect plots for the topic*system effects (Figure 3a and Figure 3b), topic*shard effects (Figure 3c and Figure 3d), and system*shard effects (Figure 3e and Figure 3f). An interaction effects plot displays the levels of one factor on the X axis and has a separate line for the marginal means of each level of the other factor on the Y axis. For example, in Figure 3b each step of the X axis corresponds to one of the 50 topics used in T09; each line corresponds to one of the 104 system runs submitted to T09; and, the Y axis is the marginal mean of AP for each system with respect to each topic. An interaction effects plot allows us to understand whether the effect of one factor depends on the level of the other factor. Indeed, two parallel lines indicate that no interaction occurred, whereas nonparallel lines indicate an interaction between factors; the more nonparallel the lines are, the greater the strength of the interaction. For example, in Figure 3b, we highlighted one of the system runs with a thick black line: you can note how, as expected, the performance of the system varies from topic to topic, being more effective for some topics and less for others; moreover, the line corresponding to the highlighted system crosses several times the lines corresponding to other systems and this is an indication of interaction.

Figures 2 and 3 allow us to appreciate some of the phenomena described above by visually inspecting the underlying analysed data. Note that T10, random sharding schemes, and other measure are omitted from Figures 2 and 3 for brevity but trends are similar.



(a) Topic effect on WT10g_TLD shards.

(b) Topic effect on WT10g_TLD_EVEN shards.

(c) System effect on WT10g_TLD shards.

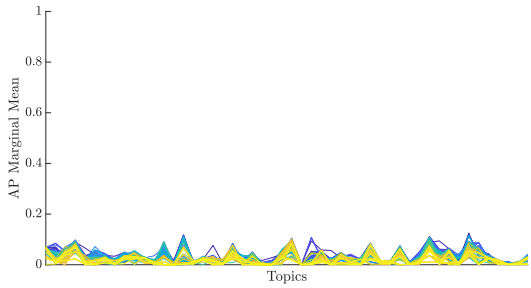
(d) System effect on WT10g_TLD_EVEN shards.

(e) Shard effect on WT10g_TLD shards.

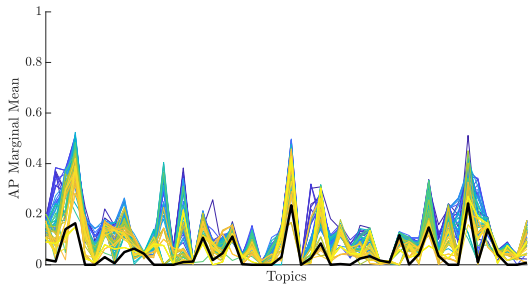
(f) Shard effect on WT10g_TLD_EVEN shards.

Fig. 2. Main effects plots using AP on T09 for different TLD sharding schemes. T10, other random shard schemes, and other measure are omitted for brevity but trends are similar.

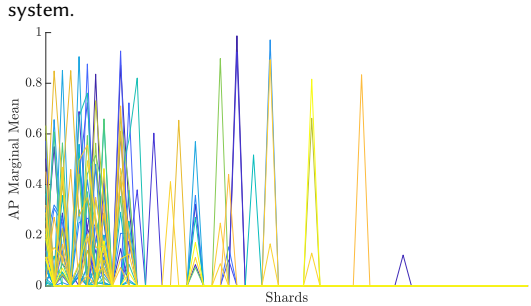
From Figure 2 we can observe that uneven size shards (on the left) tend to depress the topic and system effect while even size shards (on the right) have the opposite outcome. This behaviour comes from the fact that, when using uneven shards, there are several shards without any relevant documents and this lowers the overall performance across both topics and systems; on the other hand, even size shards are typically characterized by a more even distribution of relevant documents across shards, allowing topics and systems to have more chances for performing better. Moreover, uneven size shards happen to concentrate the shard effect only in few shards, the biggest ones, while even size shards exhibit a shard effect more distributed across all the shards. Also in this case, this behaviour reflects the concentration of relevant documents only in few shards which achieve higher performance with respect to evenly spreading relevant documents across shards and achieving a more uniform distribution of performance.



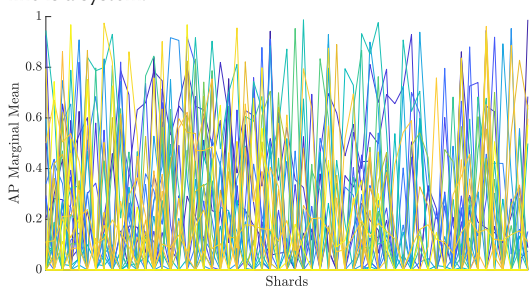
(a) Topic*System effect on WT10g_TLD shards. Each line is a system.



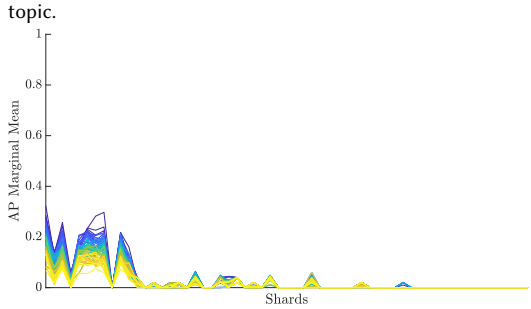
(b) Topic*System effect on WT10g_TLD_EVEN shards. Each line is a system.



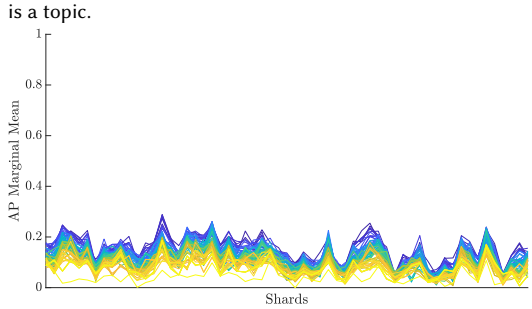
(c) Topic*Shard effect on WT10g_TLD shards. Each line is a topic.



(d) Topic*Shard effect on WT10g_TLD_EVEN shards. Each line is a topic.



(e) System*Shard effect on WT10g_TLD shards. Each line is a system.



(f) System*Shard effect on WT10g_TLD_EVEN shards. Each line is a system.

Fig. 3. Interaction effects plots using AP on T09 for different TLD and sharding schemes. T10, other random shard schemes, and other measures are omitted for brevity but trends are similar.

From Figure 3 we can observe how much bigger the topic*shard interaction effect is with respect to the others, as shown by the "very messy" set of lines which keep crossing each other (Figure 3c and Figure 3d). On the other hand, the system*shard interaction effect is the weakest one with lines which follow a more regular pattern and tend to be more parallel, being characterized by less crossing among the lines (Figure 3e and Figure 3f). Finally, the topic*system interaction effect falls between the two (Figure 3a and Figure 3b), being consistent with previous findings in the literature about the size of this effect. As in the previous case of Figure 2, we note how uneven shard sizes happen to concentrate the interaction within fewer shards, the biggest ones containing most of the relevant documents, while even shard sizes happen to distribute it across all the shards. Finally, we note how even shard sizes substantially increase the topic*system interaction with respect to the case of uneven shard sizes. Indeed, since even size shards typically exhibit a more

even distribution of relevant documents across shards, there are more cases, i.e. shards, where a system can be good (or not) at retrieving relevant documents for a given topic and this boosts the topic*system interaction effect.

7.2 Sharding by topic cluster

Selective search is a distributed search architecture which uses topically focused shards to improve the efficiency of large-scale search. Dai et al. [16] present a query-log biased clustering method for creating shards for selective search. The clusters are initially seeded with topics from the AOL query log and during clustering, terms common in the query log are favoured by the similarity metric, creating shards with a strong topic skew. We use the 199 shards of the Gov2 dataset created by [16] and apply our analysis (`GOV2_199`). The results are presented in Table 18 on the left. Across the three years of TREC submissions, one can see a large shard and topic*shard effect. These results corroborate and encapsulate several observations about selective search shards made in prior work [33, 34].

The strength of the topic*shard interaction confirms the purpose of the selective search shards, which were designed specifically to represent different topics. Indeed, the topic*shard interaction effects are the most prominent and make up the greatest majority of the total variance in `GOV2_199` than any other experiment. The single-factor shard effect also quantifies an observation made in prior work: topic imbalances exist in queries and certain selective search shards are globally more popular than others [33].

Given the specialized nature of the shards in selective search, we wondered if we could design systems specialized to each shard that would produce better rankings, and thus producing larger system*shard effects.

Therefore, we created fifty "systems" where each system performs RM3 query expansion [36] on given test collection topics with a particular shard. In RM3 query expansion is a pseudo relevance feedback algorithm. The initial query is submitted to a search index and the top retrieved documents are used to generate expansion terms that are likely to co-occur with the initial query. The expansion terms and initial query are then interpolated for the final retrieval.

In our experiments, each "system" created expanded queries from the given topic, where expansion terms were generated from the top retrieved documents from a particular shard. 50 expansion terms were generated from the top 10 documents, and the new and expanded queries were weighed equally (0.5). In essence, this creates systems that were tuned to one particular shard. These shard-specific systems are then used to retrieve documents from all shards and these result lists are used in analyses, in lieu of the TREC submitted systems[5].

Table 18 on the right presents the results for the fifty shard-specific systems. These simple, shard-tuned systems also produced small system*shard interactions. Our findings support that of Jones et al. [31], which attempted to create customised relevance scoring functions for shards but did not find accuracy gains. While this does not mean creating systems that are tuned for individual shards are impossible, it clearly requires more careful methods.

[5]Note that this choice, which represents what actually happens in practice when you rely on shards for selective search, somehow departs from the independence assumption behind ANOVA, since some systems are specifically target to some shards. As explained by Scariano and Davenport, this violation of the ANOVA assumption may impact on the outcomes of the F-test. While from the results reported in Table 18 on the right this seems to not actually change the general trends observed also in the other experiments, it is left for future work to investigate the extent this violations may have impacted on the results.

Table 18. Selective search shards and system alternatives on GOV2 for T13, T14, and T15. In the left column, there are selective search shards with original TREC systems. In the right column, there are selective search shards with fifty different shard-specific systems created through RM3 query expansion with shard language models. For each effect, its effect size $\hat{\omega}^2$ is reported. Insignificant effects are in gray; small effects, light blue; medium, blue; and large, dark blue.

Measure	T13 – GOV2_199						T13_199s – GOV2_199					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.05	0.01	0.20	0.02	0.01	0.63	0.32	<1e-2	0.65	<1e-2	0.01	0.94
R-prec	0.03	<1e-2	0.16	0.01	0.01	0.57	0.19	<1e-2	0.48	<1e-2	<1e-2	0.90
RBP	0.07	0.01	0.37	0.02	0.01	0.79	0.36	<1e-2	0.77	<1e-2	0.01	0.96
nDCG	0.07	0.01	0.40	0.03	0.02	0.75	0.42	<1e-2	0.82	0.01	0.01	0.96
ERR	0.06	<1e-2	0.30	0.01	0.01	0.72	0.25	<1e-2	0.67	<1e-2	<1e-2	0.93
Twist	0.04	<1e-2	0.20	0.01	0.01	0.61	0.19	<1e-2	0.53	<1e-2	<1e-2	0.90
P@5	0.06	0.01	0.30	0.02	0.01	0.73	0.26	<1e-2	0.67	<1e-2	<1e-2	0.94
P@10	0.07	0.01	0.36	0.02	0.01	0.80	0.29	<1e-2	0.74	<1e-2	<1e-2	0.96
P@20	0.07	0.01	0.39	0.03	0.01	0.84	0.33	<1e-2	0.79	<1e-2	0.01	0.97
P@50	0.06	0.01	0.40	0.02	0.01	0.87	0.32	<1e-2	0.81	<1e-2	0.01	0.98
P@100	0.06	<1e-2	0.39	0.02	0.01	0.89	0.35	<1e-2	0.84	<1e-2	0.01	0.99
nDCG@5	0.04	0.01	0.19	0.01	0.01	0.60	0.23	<1e-2	0.55	<1e-2	<1e-2	0.91
nDCG@10	0.05	0.01	0.22	0.02	0.01	0.64	0.26	<1e-2	0.62	<1e-2	<1e-2	0.93
nDCG@20	0.05	0.01	0.26	0.02	0.01	0.67	0.29	<1e-2	0.67	<1e-2	<1e-2	0.93
nDCG@50	0.07	0.01	0.32	0.02	0.01	0.71	0.35	<1e-2	0.75	0.01	0.01	0.95
nDCG@100	0.07	0.01	0.36	0.02	0.01	0.73	0.39	<1e-2	0.79	0.01	0.01	0.96

Measure	T14 – GOV2_199						T14_199s – GOV2_199					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.08	<1e-2	0.34	0.01	0.01	0.77	0.28	<1e-2	0.67	<1e-2	<1e-2	0.94
R-prec	0.05	<1e-2	0.25	<1e-2	0.01	0.69	0.15	<1e-2	0.49	<1e-2	<1e-2	0.89
RBP	0.11	<1e-2	0.54	0.01	0.01	0.87	0.31	<1e-2	0.80	<1e-2	0.01	0.96
nDCG	0.16	0.01	0.60	0.02	0.02	0.87	0.46	<1e-2	0.86	0.01	0.01	0.97
ERR	0.08	<1e-2	0.43	0.01	0.01	0.81	0.23	<1e-2	0.70	–	<1e-2	0.94
Twist	0.05	<1e-2	0.29	<1e-2	0.01	0.72	0.16	<1e-2	0.53	–	<1e-2	0.90
P@5	0.08	<1e-2	0.45	0.01	0.01	0.83	0.20	<1e-2	0.69	<1e-2	<1e-2	0.93
P@10	0.09	<1e-2	0.53	0.01	0.01	0.87	0.25	<1e-2	0.77	<1e-2	<1e-2	0.96
P@20	0.10	<1e-2	0.59	0.01	0.01	0.91	0.28	<1e-2	0.82	<1e-2	<1e-2	0.97
P@50	0.11	<1e-2	0.65	0.01	0.01	0.94	0.30	<1e-2	0.87	<1e-2	<1e-2	0.98
P@100	0.10	<1e-2	0.65	0.01	0.01	0.95	0.32	<1e-2	0.89	0.01	<1e-2	0.99
nDCG@5	0.06	<1e-2	0.33	0.01	0.01	0.74	0.20	<1e-2	0.60	<1e-2	<1e-2	0.92
nDCG@10	0.08	0.01	0.37	0.01	0.01	0.77	0.25	<1e-2	0.65	<1e-2	<1e-2	0.93
nDCG@20	0.10	0.01	0.43	0.02	0.01	0.80	0.31	<1e-2	0.73	<1e-2	0.01	0.94
nDCG@50	0.13	0.01	0.52	0.02	0.02	0.84	0.39	<1e-2	0.81	<1e-2	0.01	0.96
nDCG@100	0.14	0.01	0.54	0.02	0.02	0.85	0.44	<1e-2	0.84	0.01	0.01	0.97

Measure	T15 – GOV2_199						T15_199s – GOV2_199					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.08	<1e-2	0.30	<1e-2	0.01	0.80	0.25	<1e-2	0.60	<1e-2	0.01	0.95
R-prec	0.05	<1e-2	0.20	<1e-2	0.01	0.73	0.14	<1e-2	0.43	<1e-2	<1e-2	0.91
RBP	0.11	<1e-2	0.45	<1e-2	0.01	0.89	0.30	<1e-2	0.73	–	<1e-2	0.97
nDCG	0.17	<1e-2	0.58	0.01	0.02	0.90	0.37	<1e-2	0.80	<1e-2	0.01	0.97
ERR	0.09	<1e-2	0.39	<1e-2	0.01	0.85	0.24	<1e-2	0.65	–	<1e-2	0.95
Twist	0.06	<1e-2	0.24	<1e-2	0.01	0.75	0.15	<1e-2	0.46	–	<1e-2	0.91
P@5	0.08	<1e-2	0.36	<1e-2	0.01	0.84	0.19	<1e-2	0.59	–	<1e-2	0.94
P@10	0.10	<1e-2	0.44	<1e-2	0.01	0.89	0.27	<1e-2	0.71	<1e-2	<1e-2	0.97
P@20	0.10	<1e-2	0.48	<1e-2	0.01	0.91	0.31	<1e-2	0.76	<1e-2	<1e-2	0.98
P@50	0.10	<1e-2	0.54	<1e-2	0.01	0.94	0.37	<1e-2	0.84	<1e-2	<1e-2	0.99
P@100	0.11	<1e-2	0.61	0.01	0.01	0.96	0.41	<1e-2	0.89	<1e-2	<1e-2	0.99
nDCG@5	0.07	<1e-2	0.28	<1e-2	0.01	0.77	0.17	<1e-2	0.52	–	<1e-2	0.92
nDCG@10	0.08	<1e-2	0.32	0.01	0.01	0.80	0.21	<1e-2	0.58	–	<1e-2	0.93
nDCG@20	0.10	<1e-2	0.38	0.01	0.01	0.83	0.25	<1e-2	0.65	<1e-2	<1e-2	0.95
nDCG@50	0.12	0.01	0.45	0.01	0.01	0.86	0.33	<1e-2	0.73	<1e-2	0.01	0.96
nDCG@100	0.14	0.01	0.51	0.01	0.02	0.88	0.36	<1e-2	0.77	<1e-2	0.01	0.96

Table 19. PDF and not PDF split shards on GOV2 for T13, T14, and T15. For each effect, its effect size $\hat{\omega}^2$ is reported. Insignificant effects are in gray; small effects, light blue; medium, blue; and large, dark blue.

Measure	**T13 – GOV2_PDF_NOTPDF**					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.67	0.42	0.01	0.18	0.37	0.24
R-prec	0.63	0.35	0.03	0.19	0.29	0.20
RBP	0.58	0.24	0.11	0.11	0.29	0.18
nDCG	0.64	0.60	<1e-2	0.43	0.45	0.17
ERR	0.42	0.12	0.14	0.07	0.10	0.13
Twist	0.60	0.41	0.03	0.25	0.30	0.20
P@5	0.48	0.18	0.08	0.08	0.24	0.14
P@10	0.58	0.24	0.10	0.11	0.30	0.17
P@20	0.65	0.30	0.14	0.12	0.31	0.21
P@50	0.73	0.32	0.28	0.15	0.32	0.30
P@100	0.75	0.29	0.38	0.16	0.25	0.37
nDCG@5	0.44	0.17	0.03	0.08	0.18	0.20
nDCG@10	0.53	0.23	0.04	0.11	0.25	0.21
nDCG@20	0.57	0.31	0.03	0.14	0.30	0.21
nDCG@50	0.62	0.40	0.01	0.22	0.38	0.26
nDCG@100	0.61	0.43	–	0.27	0.38	0.28

Measure	**T14 – GOV2_PDF_NOTPDF**					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.76	0.25	<1e-2	0.04	0.39	0.51
R-prec	0.70	0.16	<1e-2	0.03	0.33	0.49
RBP	0.71	0.10	0.05	0.02	0.29	0.49
nDCG	0.75	0.35	0.03	0.05	0.51	0.34
ERR	0.58	0.03	0.05	0.01	0.17	0.40
Twist	0.69	0.20	–	0.02	0.33	0.46
P@5	0.60	0.06	0.02	0.01	0.19	0.40
P@10	0.69	0.08	0.04	0.01	0.26	0.45
P@20	0.75	0.11	0.07	0.01	0.25	0.51
P@50	0.82	0.14	0.22	0.02	0.27	0.55
P@100	0.82	0.14	0.34	0.03	0.24	0.56
nDCG@5	0.60	0.06	–	0.02	0.19	0.43
nDCG@10	0.68	0.09	–	0.02	0.26	0.45
nDCG@20	0.71	0.12	–	0.02	0.25	0.44
nDCG@50	0.73	0.15	0.03	0.04	0.30	0.44
nDCG@100	0.69	0.19	0.10	0.05	0.34	0.36

Measure	**T15 – GOV2_PDF_NOTPDF**					
	$\hat{\omega}^2_{\langle \text{Tpc} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Sys*Shr} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Sys} \rangle}$	$\hat{\omega}^2_{\langle \text{Tpc*Shr} \rangle}$
AP	0.75	0.13	<1e-2	0.01	0.49	0.46
R-prec	0.70	0.09	0.01	0.01	0.40	0.42
RBP	0.69	0.06	0.09	0.02	0.29	0.49
nDCG	0.78	0.17	0.03	0.04	0.58	0.44
ERR	0.54	0.03	0.14	0.01	0.15	0.43
Twist	0.68	0.10	0.04	0.02	0.42	0.40
P@5	0.56	0.03	0.05	0.02	0.19	0.38
P@10	0.67	0.05	0.10	0.02	0.23	0.48
P@20	0.76	0.06	0.20	0.03	0.33	0.56
P@50	0.82	0.09	0.44	0.03	0.31	0.57
P@100	0.86	0.10	0.61	0.02	0.32	0.64
nDCG@5	0.53	0.04	0.01	0.02	0.20	0.34
nDCG@10	0.59	0.05	<1e-2	0.02	0.24	0.34
nDCG@20	0.62	0.06	0.01	0.03	0.32	0.33
nDCG@50	0.66	0.11	0.06	0.04	0.42	0.37
nDCG@100	0.71	0.16	0.14	0.05	0.49	0.37

7.3 Sharding by document type

Sanderson et al. [52] created and analyzed shards based on document type: PDF and non-PDF documents. We applied our model to these shards (GOV2_PDF_NOTPDF) for a more in-depth analysis. The results are presented in Table 19. In the case of the GOV2_199 shards in Section 7.2, we saw no notable change in effects across the three years of running the TREC track. However, this is not necessarily unexpected as the development of systems across the three years occurred before the sharding took place. In contrast, the study of GOV2_PDF_NOTPDF does examine a sharding in place as the systems were developed over the three years.

In contrast to prior experiments, a medium/large system effect is present, especially in T13, where there is also a large system*shard interaction. In subsequent years of the track, the system*shard interaction becomes much smaller. The shard effect is small or not significant, especially in T14. Overall, this suggests that these two documents types similarly contributed to system effectiveness (small size/not significant shard effect) but the systems behaved quite differently (large system*shard interaction), some were more effective on PDF documents[6] and others on non-PDF documents. These results enrich the findings of Sanderson et al. [52], who found that system rankings on PDF documents were not predictive of rankings on non-PDF. In addition, the decreasing system*shard interaction effect over subsequent years confirms the data in Table 4 of Sanderson et al., which showed some evidence that the non-predictive effect was stronger in T13 than later years.

8 DISCUSSION

In addition to observations specific to particular shards, our experiments revealed several global trends spanning across the datasets.

8.1 Extended Performance Model

In section 6 the extended performance model was validated. Extending the model with a shard, system*shard, and topic*shard factor reduced the level of error when compared to past models [6, 45, 60]. It would appear that the extended model has the potential to ensure that the differences between retrieval systems can be measured more accurately. There was some suggestion of this potential when Ferro and Sanderson [20] examined the number of statistically significant differences between systems using a simpler version of the performance model. However, a detailed examination of this topic will be left for future work.

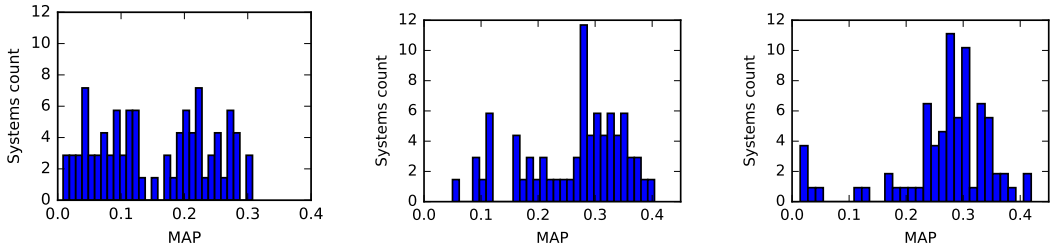
8.2 Impact of Skewed Shards

In many of the collections experimented on, much of the system*shard effect seems to be explained by the skewed nature of the shards that were formed. It is worth considering this finding when examining past results. Zhai and Lafferty [63] found that notably different parameter settings were needed to optimise a system on shards in their experiments. The skewed size of those shards was not discussed by those researchers. Similarly, Goswami et al. [24] did not consider the impact of skew in their results.

8.3 Impact of Convergence of Adopted Techniques in System Development

The system and system*shard interaction effects decreased sharply from T13 to subsequent years in GOV2_PDF_NOTPDF. We speculate this is because the first iterations of a TREC track have a wider variety of systems and in later years see a convergence of system effectiveness, as the best strategies from earlier years are propagated through the community and as later systems are trained on earlier track results. This effect is shown in Figure 4, which presents a histogram of MAP scores of

[6]Note, the 'PDF documents' were composed only of text, which was extracted from the original PDF documents [14].



(a) T13. (b) T14. (c) T15.

Fig. 4. Distribution of MAP scores for submitted systems in TREC Terabyte tracks using the GOV2 corpus.

runs submitted to the TREC Terabyte track. More subtle versions of the trend can be observed in the tables reported here for other shard sets. In later iterations of the track, smaller system effects and system*shard interaction effects are reported. To the best of our knowledge, this is the first time that results from ANOVA appear to track changes over the iterations of an evaluation track.

With the exception of T13 in GOV2_PDF_NOTPDF seen above, minimal system*shard interaction effects were reported. Jones et al. [31] attempted to build an aggregated system of relevance scoring methods customised to each shard, but found no benefit from taking such an approach. The reason for this difficulty is now explained by our results. The small system*shard interaction effect indicates that while some form of interaction is taking place, exploiting it for the benefit of overall search effectiveness is likely to be hard and would involve using very different approaches from standard techniques.

While the authors would have welcomed the opportunity to explore this form of shard further, we know of no other test collection on which we could experiment. The Gov2 collection appears to be the only collection we know of that is composed of documents that are notably different in type but that can be relevant to information needs of similar intent.

8.4 Impact of Random Shards

Large topic*shard interaction effects are present across every experiment and were the dominant factor of our measurments, especially in GOV2_199, which were designed to be topic-focused. To our surprise, the topic*shard interaction effect was also strong in randomly organized shards as well (Tables 16 and 17). To investigate further, we created four additional evenly-sized, randomly organized shard sets containing $n = \{25, 10, 5, 2\}$ shards, presented in Table 20.

As the number of shards decreases, the topic*shard interaction effects become overshadowed by the other larger effects such as the topic effect. However, even in two randomly created shards, the topic*shard interaction effect is a major factor in variance; this is due to the sparsity of relevance data and the fact that not all relevant documents are created equal. While the relevant documents are distributed fairly evenly in these random shards, some are easier to find for the submitted systems and are ranked higher, creating a skew in the actual scores in metrics such as MAP. Case in point, when the factors are compared across different metrics, rank-sensitive metrics such as AP, nDCG show topic*shard effects being comparatively stronger than other factors, where as rank insensitive metrics such as P@100 show a comparatively weaker topic*shard effect. E.g. in WT10g_RANDOM_5, the topic, system, and topic*shard effect sizes for AP are 0.52, 0.17 and 0.60 respectively, where as for P@100 they are 0.95, 0.41 and 0.84 respectively.

We speculate that the limits in effectiveness of the methodology described by Voorhees et al. [60] may be due to the topic*shard effects not being taken into account in that work.

Table 20. Effect of different number of shards, split randomly into 2, 5, 10 and 25 evenly sized shards, on WT10g and T09. T10 is omitted for brevity but trends are similar. For each effect, its effect size $\hat{\omega}^2$ is reported. Insignificant effects are in gray; small effects, light blue; medium, blue; and large, dark blue.

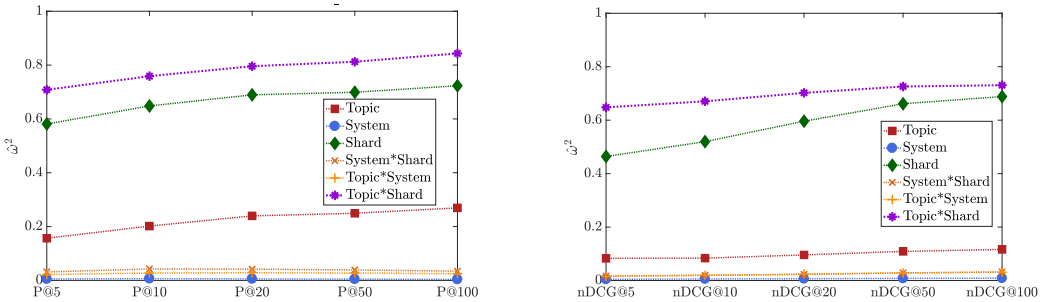
Measure	T09 – WT10g_RANDOM_2						T09 – WT10g_RANDOM_5					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.61	0.19	0.02	–	0.42	0.48	0.52	0.17	0.04	–	0.40	0.59
R-prec	0.50	0.13	0.01	–	0.30	0.32	0.41	0.10	0.01	–	0.29	0.45
RBP	0.76	0.27	<1e-2	–	0.61	0.38	0.73	0.20	0.02	–	0.53	0.46
nDCG	0.84	0.48	0.02	–	0.72	0.72	0.75	0.33	0.06	–	0.61	0.75
ERR	0.54	0.12	<1e-2	–	0.34	0.35	0.52	0.09	0.01	–	0.27	0.52
Twist	0.53	0.17	0.01	–	0.35	0.36	0.50	0.13	0.01	–	0.31	0.48
P@5	0.64	0.18	0.01	–	0.50	0.30	0.60	0.13	0.01	–	0.40	0.37
P@10	0.77	0.27	–	–	0.60	0.32	0.75	0.21	0.02	–	0.56	0.42
P@20	0.87	0.38	–	–	0.74	0.36	0.85	0.26	0.02	–	0.69	0.46
P@50	0.94	0.46	–	0.01	0.85	0.39	0.93	0.35	0.04	<1e-2	0.80	0.56
P@100	0.96	0.53	–	<1e-2	0.89	0.41	0.95	0.41	0.05	–	0.84	0.64
nDCG@5	0.55	0.12	<1e-2	–	0.37	0.38	0.41	0.11	0.02	–	0.29	0.50
nDCG@10	0.62	0.17	<1e-2	–	0.42	0.49	0.48	0.15	0.04	–	0.34	0.57
nDCG@20	0.68	0.21	0.01	–	0.47	0.55	0.55	0.18	0.04	–	0.43	0.62
nDCG@50	0.77	0.30	0.01	–	0.60	0.64	0.66	0.24	0.05	–	0.53	0.67
nDCG@100	0.79	0.36	0.02	–	0.65	0.66	0.72	0.29	0.05	–	0.58	0.72

Measure	T09 – WT10g_RANDOM_10						T09 – WT10g_RANDOM_25					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.42	0.12	0.03	<1e-2	0.30	0.61	0.38	0.07	0.04	–	0.22	0.62
R-prec	0.30	0.06	0.02	–	0.19	0.48	0.25	0.04	0.02	<1e-2	0.13	0.49
RBP	0.71	0.16	0.02	–	0.48	0.56	0.67	0.11	0.03	–	0.40	0.64
nDCG	0.67	0.20	0.02	–	0.45	0.71	0.60	0.10	0.05	–	0.29	0.71
ERR	0.50	0.07	0.02	<1e-2	0.23	0.62	0.50	0.06	0.03	–	0.20	0.68
Twist	0.37	0.08	0.02	–	0.21	0.52	0.29	0.04	0.03	<1e-2	0.13	0.52
P@5	0.59	0.11	0.01	–	0.37	0.47	0.57	0.07	0.02	–	0.32	0.54
P@10	0.74	0.16	0.02	–	0.52	0.51	0.70	0.11	0.03	–	0.43	0.62
P@20	0.84	0.21	0.03	–	0.64	0.59	0.78	0.13	0.03	–	0.51	0.68
P@50	0.90	0.26	0.04	–	0.73	0.68	0.84	0.15	0.04	–	0.59	0.74
P@100	0.93	0.29	0.05	–	0.78	0.72	0.84	0.15	0.04	–	0.59	0.74
nDCG@5	0.36	0.09	0.02	0.01	0.24	0.54	0.31	0.06	0.03	–	0.18	0.54
nDCG@10	0.40	0.12	0.02	<1e-2	0.29	0.56	0.40	0.08	0.04	–	0.22	0.59
nDCG@20	0.50	0.15	0.02	–	0.35	0.60	0.51	0.10	0.04	–	0.25	0.65
nDCG@50	0.61	0.18	0.02	–	0.42	0.67	0.59	0.11	0.05	–	0.29	0.71
nDCG@100	0.66	0.21	0.02	–	0.45	0.70	0.59	0.11	0.05	–	0.29	0.71

8.5 Impact of Measure Cut-off and Shard Size

When experimenting with different rank cut-offs for precision and nDCG, the effect sizes across all factors increase (or stay almost constant) as the deeper ranks are considered, as visualized in Figure 5. This is consistent with the knowledge that shallow metrics are noisy [53]; at shallower ranks, more of the variance is attributed to error. At deeper ranks, the error decreases as noise is reduced and the effect of the factors are more clearly visible. E.g. in WT10g_RANDOM_EVEN, MS_{err} of P@5 is 0.0044 vs. <0.0001 of P@100, more than two orders of magnitude in difference.

Figure 6 shows how the effect size for the different factors varies as the cut-off of nDCG increases and the type of random shards changes; on the left, there are different types of shards; on the right, there are different numbers of random even shards. The figure is based on the data of Tables 16



(a) Precision.

(b) Normalized Discounted Cumulated Gain.

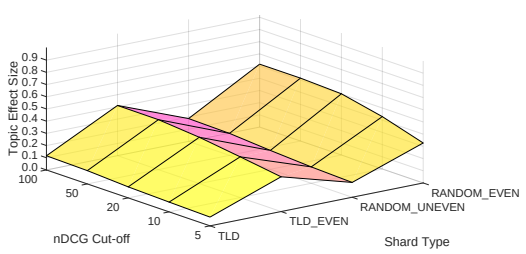
Fig. 5. Effect sizes for different document cut-off values on T10 with WT10g_TLD shards.

and 17 for the different types of shards on the left, and Table 20 for the different number of random even shards on the right. The figure shows the case T09 and WT10g for nDCG but similar trends can be observed also for precision at different cut-offs and in the case T10 of WT10g. Note that the figure does not show system*shard interaction, because it is often not significant.

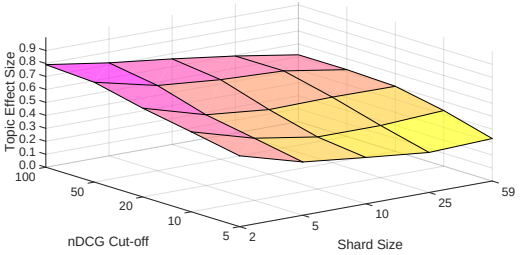
As general trends, we can confirm that increasing the document cut-off values increases the effect size. We can also observe that, in the case of random even shards (right part of Figure 6), increasing the size of the shards, i.e. decreasing their number, increases the size of the effects. Finally, even size shards (left part of Figure 6) tend to increase the effect size with respect to uneven size shards, apart from the Shard effect which is lowered. All these trends can be attributed to variations in the distribution of relevant document in the different cases. When you increase the cut-off of an evaluation measure, you give it the possibility of "seeing" more relevant documents, i.e. those that are retrieved in lower parts of the ranking, and this enhances the different effects. When you have even shards, decreasing their numbers allows for more relevant documents in the same shards; this reduces the shard effect and gives room for enhancing the other effects. This phenomenon is further highlighted by the difference between uneven and even shards: in the former case the shard effect is very pronounced due to the skewed distribution of relevant documents concentrated in few shards; in the latter case, the even distribution of relevant documents across shards reduces the shard effect.

There are a few exceptions to these general trends: the shard effect increases for uneven shard sizes and decreases for even ones; moreover, in the case of even size shards, it increases as the number of shards increases, i.e. when the size of the shards decreases. This suggests that, in both cases, smaller size shards impact more on system performance because, probably, there are fewer systems able to perform well on smaller shards.

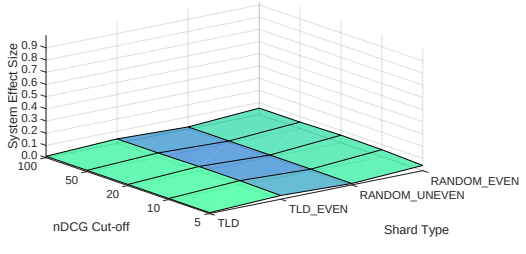
The other exception is the topic*shard interaction effect, which confirms to be one of the most prominent effects in all the cases. It increases more as the shard size decreases, in the case of random even shards, and it is more sensitive to the document cut-off than to the type of shards (even or uneven size). Overall, this further supports the importance of the topic*shards effect which seems not to be just a by product of the skewness/size of the shards but rather a more intrinsic form of interaction between topics and shards under different circumstances.



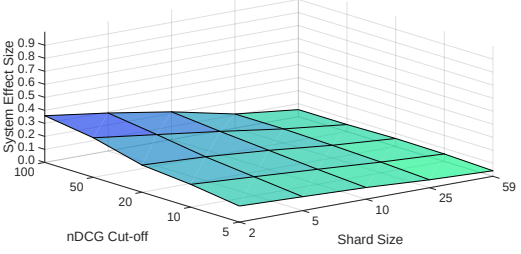
(a) Topic effect.



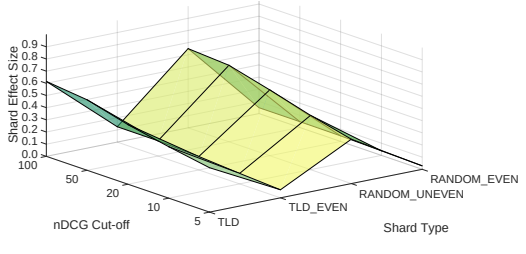
(b) Topic effect.



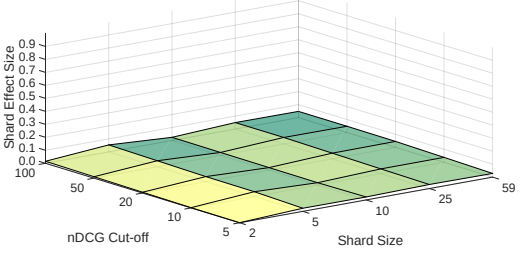
(c) System effect.



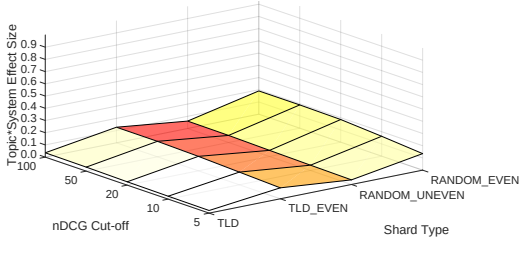
(d) System effect.



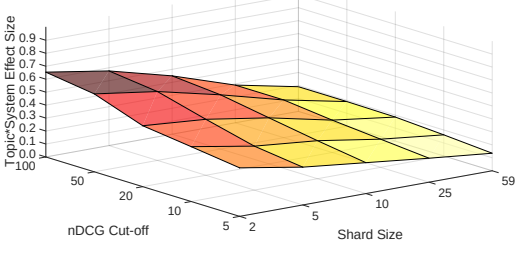
(e) Shard effect.



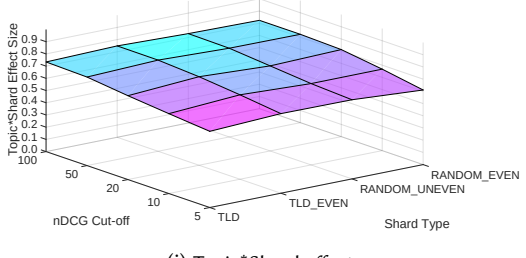
(f) Shard effect.



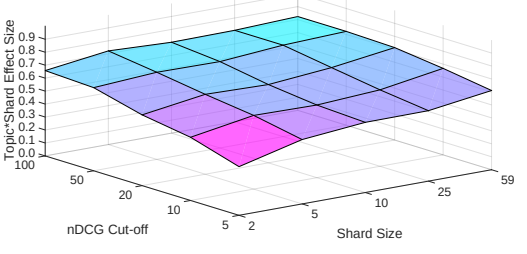
(g) Topic*System effect.



(h) Topic*System effect.



(i) Topic*Shard effect.



(j) Topic*Shard effect.

Fig. 6. Effect sizes for nDCG at document cut-offs (5, 10, 20, 50, 100) on T09 and WT10g; left: different types of shards (TLD, TLD_EVEN, RANDOM_UNEVEN, RANDOM_EVEN); right: different number of random even shards (2, 5, 10, 25, 59).

Table 21. Overall summary of the effect sizes across all the tracks and evaluation measures. Each cell reports the number of times an event happens and, within parenthesis, its ratio with respect to the total cases.

Factor	Not Significant	Small Size	Medium Size	Large Size	Total
Topic	0 (0.00%)	13 (3.69%)	65 (18.47%)	274 (77.84%)	352 (100.00%)
System	0 (0.00%)	194 (55.11%)	76 (21.59%)	82 (23.30%)	352 (100.00%)
Shard	11 (3.12%)	176 (50.00%)	34 (9.66%)	131 (37.22%)	352 (100.00%)
System*Shard	127 (36.08%)	202 (57.39%)	14 (3.98%)	9 (2.56%)	352 (100.00%)
Topic*System	0 (0.00%)	104 (29.55%)	47 (13.35%)	201 (57.10%)	352 (100.00%)
Topic*Shard	0 (0.00%)	0 (0.00%)	2 (0.57%)	350 (99.43%)	352 (100.00%)

8.6 Overall Trends

Table 21 is based on the data reported in Tables 16, 17, 18 on the left[7], 19, and 20 plus the different types of random even and uneven shards on `T10`, which are the same setting as those of Table 20 where we did not report `T10` results just for space reasons. Table 21 summarizes, across all the tracks and evaluation measures, how many times each effect is not significant, small size, medium size, and large size.

We can observe as the topic*shard interaction effect is almost always a large-size effect followed by the topic effect, which so far was deemed the biggest effect [50, 56], followed by the shard effect itself. To the best of our knowledge, a topic*shard interaction has not been measured before, however, we speculate that the impact of its effect has been observed in past results. Voorhees [57] described experiments with the robust track where topics that were thought to be 'hard' were re-run on another collection. Some of those presumed hard topics were found to be not hard, just topics that had few relevant documents in the original collection. This appears to be an instantiation of the topic*shard effect. However, including that effect in a model of performance scores has not been proposed before.

It is important to note that the three of the major effects described here depend only on the experimental collection and how it is designed and they do not involve IR systems. The system effect is a small size effect two thirds of the times while the remaining third is evenly split between medium and large size. The scale of the former effects relative to the effects that measure the impact of systems shows how challenging it is to measure changes in retrieval effectiveness.

The topic*system interaction effect is, roughly, half of the time a large size and half of the time a small size. This is an important effect, as also suggested by previous work [5, 45, 60], and by far the most prominent effect involving IR systems, even if no more one of the biggest effects. The system*shard interaction effect is most of the times a small size effect and in a quarter of the cases it is even not statistically significant. As previously discussed, this supports the previous findings about the difficulty in developing systems customized for shards [31].

9 CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the development of richer and more explanatory models of IR system performance, a research theme whose urgency has been recently prompted [1, 17]. The contributions of this work are threefold.

First, we developed a GLMM model accounting for effects of topics, systems, and shards as well as all their interactions. We incrementally built the model and showed that it substantially reduces the size of the error, increasing confidence in the effects seen. That is, our model explains most of

[7]The right side of Table 18 contains the "fifty" systems optimised for the 199 shards which, as discussed in Section 7.2, may violate the independence assumption of ANOVA. Even if the behaviour of the effect sizes using those systems follows the overall trends, we have not counted them in Table 21 to avoid any possible bias.

the variance in the accuracy of systems and creates accurate estimations of the size and significance of effects. We validated our choices for a balanced design by comparing it to Ferro and Sanderson [20] and exploring alternative interpolation methods.

Second, we conducted an extensive and systematic investigation over several TREC collections and across many evaluation measures to get an understanding of how the effects vary and what influences them. By applying our model to various types of shard used in prior work, we quantified and better explained effects described in the past. A shard effect appears to be caused by a number of factors: heavily skewed shards have a large shard effect due to their sizes alone; topical clustering of shards cause large effects; domain ordering cause medium to large effects; finally, document types cause weak effects at best. While the number of topics needed in a test collection is a richly studied area, see e.g. Buckley and Voorhees [7] or Sakai [49], the potential sparsity of relevance data is less examined. Our results suggest that one must be careful when designing experiments that partition collections randomly, since the even or uneven distribution of documents across the shards as well as the shard size have an impact on the observed results. Notably, our results explain the observations of Voorhees et al. [60], where large differences were found between two different sets of evenly-sized random shards, which were expected to be just representative replicates of the whole collections.

Finally, we are the first to note the strength of topic*shard interactions compared to other factors of variance in a principled setting. This not only suggests that one has to take care when splitting test collections for experimental purposes but also that the way in which experimental collections are designed, in terms of topics and documents, can greatly affect the outcomes of measurements, as also discussed above. Indeed, even if we do not explicitly split them, our document collections are intrinsically made up of different parts – obvious examples are sport, economics, politics in newspapers or different subjects in a digital library catalog. Such was the size of the topic*shard effect, we suggest that it should be a primary concern in test collection design, as the topic factor alone is, since it can greatly affect the observed system performance, even if it is a factor completely external to the examined systems.

Future work will concern further investigation on how the distribution of relevant documents across topics and shards affects the different factors we have studied in this paper and how this should be taken into consideration while designing experimental collections.

Moreover, a future application of our approach can be the study of what happens in a multilingual setting. Indeed, evaluation campaigns such as CLEF, NTCIR, and FIRE always strived to develop topics able to retrieve documents across collections in multiple languages and this was a major challenge. Our methods could thus help in interpreting performance of multilingual IR systems.

10 ACKNOWLEDGMENTS

This research is supported in part by the Australian Research Council's Discovery Projects Scheme (DP170102726) and by the NSF grant (IIS-1302206).

REFERENCES

[1] J. Allan, J. Arguello, L. Azzopardi, P. Bailey, T. Baldwin, K. Balog, H. Bast, N. Belkin, K. Berberich, B. von Billerbeck, J. Callan, R. Capra, M. Carman, B. Carterette, C. L. A. Clarke, K. Collins-Thompson, N. Craswell, W. B. Croft, J. S. Culpepper, J. Dalton, G. Demartini, F. Diaz, L. Dietz, S. Dumais, C. Eickhoff, N. Ferro, N. Fuhr, S. Geva, C. Hauff, D. Hawking, H. Joho, G. J. F. Jones, J. Kamps, N. Kando, D. Kelly, J. Kim, J. Kiseleva, Y. Liu, X. Lu, S. Mizzaro, A. Moffat, J.-Y. Nie, A. Olteanu, I. Ounis, F. Radlinski, M. de Rijke, M. Sanderson, F. Scholer, L. Sitbon, M. D. Smucker, I. Soboroff, D. Spina, T. Suel, J. Thom, P. Thomas, A. Trotman, E. M. Voorhees, A. P. de Vries, E. Yilmaz, and G. Zuccon. 2018. Research Frontiers in Information Retrieval – Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (June 2018).

[2] J. A. Aslam, E. Yilmaz, and V. Pavlu. 2005. A Geometric Interpretation of R-precision and Its Correlation with Average Precision, See [4], 573–574.

[3] J. A. Aslam, E. Yilmaz, and V. Pavlu. 2005. The Maximum Entropy Method for Analyzing Retrieval Measures, See [4], 27–34.

[4] R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait (Eds.). 2005. *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005).* ACM Press, New York, USA.

[5] D. Banks, P. Over, and N.-F. Zhang. 1999. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval* 1, 1-2 (May 1999), 7–34.

[6] David Bodoff and Pu Li. 2007. Test theory for assessing IR test collections. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 367–374.

[7] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*. ACM, New York, NY, USA, 33–40. https://doi.org/10.1145/345508.345543

[8] 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.

[9] S. Büttcher, C. L. A. Clarke, and I. Soboroff. 2007. The TREC 2006 Terabyte Track. In *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-272, Washington, USA.

[10] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. 2007. Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements. 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, 63–70.

[11] Jamie Callan. 2002. Distributed information retrieval. *Advances in information retrieval* (2002), 127–150.

[12] 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.

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

[14] C. L. A. Clarke, N. Craswell, and I. Soboroff. 2004. Overview of the TREC 2004 Terabyte Track. In *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-261, Washington, USA.

[15] C. L. A. Clarke, F. Scholer, and I. Soboroff. 2005. Overview of the TREC 2005 Terabyte Track. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-266, Washington, USA.

[16] Zhuyun Dai, Chenyan Xiong, and Jamie Callan. 2016. Query-Biased Partitioning for Selective Search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. ACM, 1119–1128.

[17] N. Ferro, N. Fuhr, G. Grefenstette, J. A. Konstan, P. Castells, E. M. Daly, T. Declerck, M. D. Ekstrand, W. Geyer, J. Gonzalo, T. Kuflik, K. Lindén, B. Magnini, J.-Y. Nie, R. Perego, B. Shapira, I. Soboroff, N. Tintarev, K. Verspoor, M. C. Willemsen, and J. Zobel. 2018. Manifesto from Dagstuhl Perspectives Workshop 17442 – From Evaluating to Forecasting Performance: How to Turn Information Retrieval, Natural Language Processing and Recommender Systems into Predictive Sciences. *Dagstuhl Manifestos, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany* 7, 1 (2018), 96–139.

[18] N. Ferro, N. Fuhr, G. Grefenstette, J. A. Konstan, P. Castells, E. M. Daly, T. Declerck, M. D. Ekstrand, W. Geyer, J. Gonzalo, T. Kuflik, K. Lindén, B. Magnini, J.-Y. Nie, R. Perego, B. Shapira, I. Soboroff, N. Tintarev, K. Verspoor, M. C. Willemsen, and J. Zobel. 2018. The Dagstuhl Perspectives Workshop on Performance Modeling and Prediction. *SIGIR Forum* 52, 1 (June 2018), 91–101.

[19] N. Ferro and D. Harman. 2010. CLEF 2009: Grid@CLEF Pilot Track Overview. In *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers*, C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda (Eds.). Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany, 552–565.

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

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

[22] N. Ferro and G. Silvello. 2018. Toward an Anatomy of IR System Component Performances. *Journal of the American Society for Information Science and Technology (JASIST)* 69, 2 (February 2018), 187–200.
[23] N. Ferro, G. Silvello, H. Keskustalo, A. Pirkola, and K. Järvelin. 2016. The Twist Measure for IR Evaluation: Taking User's Effort Into Account. *Journal of the American Society for Information Science and Technology (JASIST)* 67, 3 (2016), 620–648.
[24] Parantapa Goswami, Eric Gaussier, and Massih-Reza Amini. 2017. Exploring the space of information retrieval term scoring functions. *Information Processing & Management* 53, 2 (2017), 454–472.
[25] D. Hawking. 2000. Overview of the TREC-9 Web Track. In *The Ninth Text REtrieval Conference (TREC-9)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-249, Washington, USA, 87–103.
[26] D. Hawking and N. Craswell. 2001. Overview of the TREC-2001 Web Track. In *The Tenth Text REtrieval Conference (TREC 2001)* , E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-250, Washington, USA, 61–67.
[27] D. A. Hull. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, R. Korfhage, E. Rasmussen, and P. Willett (Eds.). ACM Press, New York, USA, 329–338.
[28] 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.
[29] C. M. Jarque and A. K. Bera. 1987. A Test for Normality of Observations and Regression Residuals. *International Statistical Review* 55, 2 (1987), 163–172.
[30] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (October 2002), 422–446.
[31] Timothy Jones, Falk Scholer, Andrew Turpin, Stefano Mizzaro, and Mark Sanderson. 2015. Different rankers on different subcollections. In *European Conference on Information Retrieval*. Springer, 203–208.
[32] Timothy Jones, Andrew Turpin, Stefano Mizzaro, Falk Scholer, and Mark Sanderson. 2014. Size and Source Matter: Understanding Inconsistencies in Test Collection-Based Evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 1843–1846. https://doi.org/10.1145/2661829.2661945
[33] Yubin Kim, Jamie Callan, J. Shane Culpepper, and Alistair Moffat. 2017. Efficient distributed selective search. *Information Retrieval Journal* 20, 3 (2017), 221–252.
[34] Anagha Kulkarni and Jamie Callan. 2010. Document Allocation Policies for Selective Searching of Distributed Indexes. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. ACM, 449–458.
[35] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin, New York, USA.
[36] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 120–127.
[37] I. Levene. 1960. Robust Tests for Equality of Variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann (Eds.). Stanford University Press, USA, 278–292.
[38] Bo Long and Yi Chang. 2014. *Relevance Ranking for Vertical Search Engines* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
[39] S. Maxwell and H. D. Delaney. 2004. *Designing Experiments and Analyzing Data. A Model Comparison Perspective* (2nd ed.). Lawrence Erlbaum Associates, Mahwah (NJ), USA.
[40] W. Mendenhall and T: Sincich. 2012. *A Second Course in Statistics. Regression Analysis* (7th ed.). Prentice Hall, USA.
[41] A. Moffat and J. Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2:1–2:27.
[42] 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.
[43] B. Pobelte and R. Baeza-Yates. 2008. Query-sets: using implicit feedback and query patterns to organize web documents. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, 41–50.
[44] Diego Puppin, Fabrizio Silvestri, and Domenico Laforenza. 2006. Query-driven Document Partitioning and Collection Selection. In *Proceedings of the 1st International Conference on Scalable Information Systems*. 34.
[45] 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.

[46] S. E. Robertson, E. Kanoulas, and E. Yilmaz. 2010. Extending Average Precision to Graded Relevance Judgments. In *Proc. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, F. Crestani, S. Marchand-Maillet, E. N. Efthimiadis, and J. Savoy (Eds.). ACM Press, New York, USA, 603–610.
[47] A. Rutherford. 2011. *ANOVA and ANCOVA. A GLM Approach* (2nd ed.). John Wiley & Sons, New York, USA.
[48] 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.
[49] T. Sakai. 2016. Topic set size design. *Information Retrieval* 19, 3 (June 2016), 256–283.
[50] Gerard Salton and Michael E Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the ACM (JACM)* 15, 1 (1968), 8–36.
[51] Mark Sanderson, Andrew Turpin, Ying Zhang, and Falk Scholer. 2012. Differences in effectiveness across sub-collections. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1965–1969.
[52] Mark Sanderson, Andrew Turpin, Ying Zhang, and Falk Scholer. 2012. Differences in Effectiveness Across Sub-collections. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1965–1969. https://doi.org/10.1145/2396761.2398553
[53] M. Sanderson and J. Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, See [4], 162–169.
[54] 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.
[55] Milad Shokouhi, Luo Si, et al. 2011. Federated search. *Foundations and Trends® in Information Retrieval* 5, 1 (2011), 1–102.
[56] J. M. Tague-Sutcliffe and J. Blustein. 1994. A Statistical Analysis of the TREC-3 Data. In *The Third Text REtrieval Conference (TREC-3)*, D. K. Harman (Ed.). National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 385–398.
[57] Ellen M Voorhees. 2006. The TREC 2005 robust track. In *ACM SIGIR Forum*, Vol. 40. ACM, 41–48.
[58] E. M. Voorhees and C. Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. Hyon Myaeng (Eds.). ACM Press, New York, USA, 316–323.
[59] E. M. Voorhees and D. K. Harman. 1999. Overview of the Eigth Text REtrieval Conference (TREC-8). In *The Eighth Text REtrieval Conference (TREC-8)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-246, Washington, USA, 1–24.
[60] E. M. Voorhees, D. Samarov, and I. Soboroff. 2017. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)* 36, 2 (September 2017), 12:1–12:21.
[61] Meng Yang, Peng Zhang, and Dawei Song. 2018. A Study of Per-Topic Variance on System Comparison. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1181–1184.
[62] B. W. Yap and C. H. Sim. 2011. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation* 81, 12 (2011), 2141–2155.
[63] Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 334–342.
[64] ChengXiang Zhai and John Lafferty. 2002. Two-stage language models for information retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 49–56.
[65] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* 22, 2 (2004), 179–214.

Received 30 July 2018; revised 26 November 2018; accepted 28 January 2019