Towards an Anatomy of IR System Component Performances

Nicola Ferro[a], Gianmaria Silvello[a]

[a]*Department of Information Engineering, University of Padua*
Via Gradenigo 6/b, Padua, Italy
{*ferro, silvello*}*@dei.unipd.it*
tel. +39 049 827 7500

Abstract

Information Retrieval (IR) systems are the prominent means for searching and accessing huge amounts of unstructured information on the Web and elsewhere. They are complex systems, constituted by many different components interacting together, and evaluation is crucial to both tune and improve them. Nevertheless, in the current evaluation methodology, there is still no way to determine how much each component contributes to the overall performances and how the components interact together. This hampers the possibility of a deep understanding of IR system behaviour and, in turn, prevents us from designing ahead which components are best suited to work together for a specific search task.

In this paper, we move the evaluation methodology one step forward by overcoming these barriers and beginning to devise an "anatomy" of IR systems and their internals. In particular, we propose a methodology based on the *General Linear Mixed Model (GLMM)* and *ANalysis Of VAriance (ANOVA)* to develop statistical models able to isolate system variance and component effects as well as their interaction, by relying on a *Grid of Points (GoP)* containing all the combinations of the analysed components.

We apply the proposed methodology to the analysis of two relevant search tasks – news search and Web search – by using standard TREC collections. We analyse the basic set of components typically part of an IR system, namely stop lists, stemmers and n-grams, and IR models. In this way, we derive insights about English text retrieval.

1. Introduction

A limitation of the current experimental methodology in *Information Retrieval (IR)* is that it allows us to evaluate IR systems and search engines only as a kind of "black-box", without an understanding of how their different components interact with each other and contribute to the overall performances. In other terms, the current experimental methodology considers system performances as indivisible and it cannot break them down into the contributions of the different components constituting an IR system. This severe impediment was pointed out a long time ago by (Robertson, 1981): "if we want to decide between alternative indexing strategies for example, we must use these strategies as part of a complete information retrieval system, and examine its overall performance (with each of the alternatives) directly".

This limitation has several drawbacks: it prevents us from gaining a thorough understanding of IR system performances; it precludes the possibility of knowing beforehand which combination of components is best suited for a specific search task or collection of documents (Fuhr, 2010,

2012); and, it hampers the possibility of determining in which components it is more convenient to invest effort and resources because they or their combination have the greatest impact in terms of performance gains.

Therefore, we would need to develop some sort of *anatomy* of IR system performances, meant as a set of methodologies which allow us to study and gain knowledge about the internals of IR system performances and how to combine them.

In this paper, we start to overcome the limitations of the current experimental methodology and we provide the means to estimate the effects of the different components of an IR system. In particular, we develop a methodology, based on the *General Linear Mixed Model (GLMM)* and *ANalysis Of VAriance (ANOVA)* (Rutherford, 2011), which makes use of a *Grid of Points (GoP)* containing all the possible combinations of inspected components (Ferro and Silvello, 2016a).

We create extensive GoPs covering 6 different stop lists, 6 types of stemmers, 8 flavours of n-grams, and 17 distinct IR models, representing the basic set of common components typically present in any IR system for English retrieval. Then, the proposed methodology allows us to break down the system performances into the contributions of these stops lists, stemmers or n-grams and IR models, as well as to study their interaction.

The main contributions of the paper are:

- the methodology for breaking down component effects and analysing the GoPs across multiple evaluation measures;
- the GoPs themselves, a valuable resource which can also be exploited for other kinds of analyses and which is available to the research community[1]. They also provide a very extensive sample of the most commonly used components for English retrieval;
- the application of the proposed methodology to two different search tasks, namely news search and Web search, and their thorough analysis in order to derive insights on English retrieval.

Note that the goal of the paper is not to analyse the effects of any possible kind of components, such as word compounding, entity extraction or query expansion, just to name a few, which you may find in an operational IR system. The goal is instead to introduce a methodology that allows us to break-down the effects of components and to show an application of this methodology to the basic and common set of components described above, highlighting what it allows us to understand in the context of two relevant search tasks, as news search and Web search are. Further research could then exploit the proposed methodology and apply it to other kinds of components and search tasks.

The paper is organized as follows: Section 2 presents related works; Section 3 introduces our methodology; Section 4 describes the experimental setup and the produced GoPs; Section 5 and 6 apply, respectively, the proposed methodology to the analysis of the news and Web search tasks; Section 7 analyses the component effects and their interaction with evaluation measures; finally, Section 8 draws some conclusions and provides an outlook for future work.

2. Related works

Experimental evaluation is a core activity in IR but it is very demanding in terms of both the time and effort needed to perform it. Therefore, it is usually carried out in publicly open and large-

[1] `http://gridofpoints.dei.unipd.it/`

scale evaluation campaigns at the international level, all based on the Cranfield paradigm (Cleverdon, 1997), which makes use of shared experimental collections. This allows for sharing the effort of producing large experimental collections and comparing state-of-the-art systems and algorithms on a common and reproducible ground.

The impossibility of testing a single component by setting it aside from the complete IR system is a long-standing and well-known problem in IR experimentation. Component-based evaluation methodologies have tried to tackle this issue by providing technical solutions for mixing different components without the need of building an entire IR system from scratch. However, even if these approaches allowed researchers to focus on the components of their own interest, they have not delivered estimates of the performance figures of each component yet.

When it comes to explaining system performances and differences between algorithms, it is commonly understood (Jayasinghe et al., 2015; Robertson and Kanoulas, 2012; Tague-Sutcliffe and Blustein, 1994) that system performances can be broken down to a reasonable approximation as

$$\text{system performances} = \text{topic effect} + \text{system effect} + \quad \text{topic/system interaction effect}$$

even though it is not always possible to estimate these effects separately, especially the interaction effect.

It is well-known that topic variability is greater than system variability (Tague-Sutcliffe and Blustein, 1994). Therefore, a lot of effort has been put into better understanding this source of variance (Robertson and Kanoulas, 2012) as well as making IR systems more robust to it, e.g. (Zhang et al., 2014), basically by trying to improve on the interaction effect. Nevertheless, with respect to an IR system, topic variance is a kind of "external source" of variation, which cannot be controlled by developers, but can only be taken into account to better deal with it.

On the other hand, system variance is a kind of "internal source" of variation, since it is originated by the choice of system components, it can be directly affected by developers by working on them, and it represents the intrinsic differences between algorithms.

The decomposition of performances into system and topic effects has been exploited by Banks et al. (1999) and Tague-Sutcliffe and Blustein (1994) to analyze *Text REtrieval Conference (TREC)* data; Carterette (2012) proposed model-based inference, using linear models and ANOVA, as an approach to multiple comparisons; Jayasinghe et al. (2015) used multivariate linear models to compare non-deterministic IR systems among them and with deterministic ones. In all these cases, the goal is a more accurate comparison among systems rather than an analysis and breakdown of system variance per se. Robertson and Kanoulas (2012) applied GLMM to the study of per-topic variance by using simulated data to generate more replicates for each (topic, system) pair in order to also estimate the topic/system interaction effect.Sanderson et al. (2012) and Jones et al. (2014) started a preliminary investigation of the sub-corpus effect exploiting correlation analysis while, recently, Ferro and Sanderson (2017) made it part of a whole model for system performance. Overall, to the best of our knowledge, this paper represents the first attempt to apply GLMM and ANOVA to the decomposition of system components effects.

The idea of creating all the possible combinations of components has been proposed by Ferro and Harman (2010), who noted that a systematic series of experiments on standard collections would have created a GoP, where (ideally) all the combinations of retrieval methods and components are represented, allowing us to gain more insights into the effectiveness of the different components and their interaction; this would have also called for the identification of *suitable baselines* with respect to which all the comparisons have to be made. Even though Ferro and Harman (2010) introduced

the idea of a GoP and how it could have been central to the decomposition of system component performances, they did not come up with a fully-fledged methodology for analysing such data and breaking down component performances, which is instead the contribution of the present work.

More recently, the proliferation of open source IR systems (Trotman et al., 2012) has greatly ameliorated the situation, allowing researchers to run systematic experiments more easily. Trotman et al. (2014) conducted a vertical exploration of variations of two IR models while the "Open-Source Information Retrieval Reproducibility Challenge" (Lin et al., 2016) provided several reproducible baselines over TREC and *Conference and Labs of the Evaluation Forum (CLEF)* collections. Overall, both these efforts added a few points to the ideal GoP mentioned above, but they do not propose any methodology for estimating the component effects. We move a step forward with respect to these works since we propose an actual methodology for exploiting such GoPs to decompose system performances and we rely on a much more fine-grained grid, in terms of the number of components and IR models experimented.

Finally, on a different angle, (Ferro, 2017) exploited GLMM and ANOVA together with GoPs as a means to study the behaviour of correlation among evaluation measures.

3. Methodology

We aim to decompose the effects of different components on the overall system performances. In particular, we are interested in investigating the effects of the components commonly present in almost any IR system: stop lists; *Lexical Unit Generator (LUG)*, namely stemmers or n-grams; and IR models, such as the vector space, the probabilistic model or the language models. For a detailed survey and description of the main IR system components refer to (Croft et al., 2009).

The methodology we adopt relies on the Cranfield paradigm (Cleverdon, 1997) based on experimental collections $\mathcal{C} = (D, T, GT)$ consisting of a set of documents D, a set of topics T, and the ground-truth GT which, for each topic $t \in T$, determines the documents $d \in D$ relevant for that topic.

We create a *Grid of Points (GoP)* on an experimental collection by running all the IR systems resulting from all the possible combinations of the considered components (stop list, LUG, IR model); stemmers and n-grams are considered as mutually exclusive LUG components, thus we do not consider IR systems using both stemmer and n-grams.

We employ a GLMM to explain 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.

In this context, we are interested both in single independent variables, i.e., the *main effects* of the different components alone, and their combinations, i.e., the *interaction effects* between components. Note that some independent variables are considered *fixed effects* – i.e., they have precisely defined levels, and inferences about its effect apply only to those levels – which in our case are different kinds of systems and components; and, some others are considered *random effects* – i.e., they describe a randomly and independently drawn set of levels that represent variation in a clearly defined wider population – which in our case are the topics.

In our experimental design, systems and their components are the experimental conditions (factors) while topics are the subjects. Given that each topic is processed by each system, we have a repeated *measure design* which has the advantages of reducing the error variance due to the greater similarity of the scores provided by the same subjects and increasing the statistical power for a fixed number of subjects.

The adopted design is the one typically used when ANOVA is applied to the analysis of the system performances in a track of an evaluation campaign, as in (Banks et al., 1999; Tague-Sutcliffe and Blustein, 1994), where the subjects are the topics and the factors are the system runs. Basically, in this context ANOVA is used to determine which experimental condition dependent variable score means differ, i.e. which systems are significantly different from others. On top of this and unlike the state-of-the-art, we are also interested in determining which proportion of variation is due to the topics and which one to the systems. Furthermore, we also aim to determine the proportion of variance explained by each component of a system.

3.1. Analysis of component effects

We define a three factors design where we manipulate factors A, B and C corresponding to the stop lists, the LUG and the IR models respectively; with this design we can also study the interaction between component pairs as well as the third order interaction between them.

In this design the systems are decomposed into three main constituents: (i) factor A (stop lists) with p levels where, for instance, A_1 corresponds to the absence of a stop list, A_2 to the indri stop list, A_3 to the terrier stop list and so on; (ii) factor B (LUG) with q levels where B_1 corresponds to the absence of a LUG, B_2 to the Porter stemmer, B_3 to the Krovetz stemmer and so on; (iii) factor C (IR models) with r levels where C_1 corresponds to BM25, C_2 to TF*IDF and so on.

The full GLMM for the described factorial ANOVA for repeated measures with three fixed factors (A, B, C) and a random factor ($T^{'}$) is:

$$Y_{ijkl} = \underbrace{\mu_{....} + \tau_i + \alpha_j + \beta_k + \gamma_l}_{\text{Main Effects}} + \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma_{kl} + \alpha\beta\gamma_{jkl}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}} \tag{3.1}$$

where: Y_{ijkl} is the score of the i-th subject in the j-th, k-th, and l-th factors; $\mu_{....}$ is the grand mean; τ_i is the effect of the i-th subject $\tau_i = \mu_{i...} - \mu_{....}$ where $\mu_{i...}$ is the mean of the i-th subject; $\alpha_j = \mu_{.j..} - \mu_{....}$ is the effect of the j-th factor, where $\mu_{.j..}$ is the mean of the j-th factor; $\beta_k = \mu_{..k.} - \mu_{....}$ is the effect of the k-th factor, where $\mu_{..k.}$ is the mean of the k-th factor; and, $\gamma_l = \mu_{...l} - \mu_{....}$ is the effect of the l-th factor where $\mu_{...l}$ is the mean of the l-th factor; ε_{ijkl} is the error committed by the model in predicting the score of the i-th subject in the three factors j, k, l. It consists of all the interaction terms between the random subjects and the fixed factors, such as $(\tau\alpha)_{ij}$, $(\tau\beta)_{ik}$ and so on, plus the error ϵ_{ijkl} which is any additional error due to uncontrolled sources of variance. As in the single factor design to calculate interaction effects with the subjects, you need to have replicates; when there is only one score per subject per factor the factor ϵ_{ijkl} cannot be separated from the interaction effects with the random subjects.

3.2. Analysis of component and measure effects

We define a four factor design which extends the three factors one defined above. The first three factors remain the same – i.e., A = stop list, B = lug, and C = model; whereas, the fourth factor D is represented by the different evaluation measures. Each available evaluation measure represents a different angle to weight and understand system performances. The goal of this new model is twofold: (i) to further dig into the components contribution without taking into account the influence of the adopted evaluation measure; (ii) to understand how measures interact with the different components.

Different evaluation measures are not directly comparable with each other, as you would expect from factorial design above, because they embed different user models. Therefore, even though all

the measure scores are in the range $[0, 1]$, AP = 0.20 is not the same as ERR = 0.20 because they exploit the $[0, 1]$ scale in different ways. In order to smooth these differences and make the scores more directly comparable, we normalized them by the maximum value achieved on the dataset and then reasoned in terms of ratios.

Basically, with this design the system variance is decomposed into factor A represented by the p levels of the stop lists, factor B with the q levels of the LUG, factor C with the r levels of the IR models and factor D with the s levels of the measures.

$$\begin{aligned} Y_{ijklm} = & \underbrace{\mu_{.....} + \tau_i + \alpha_j + \beta_k + \gamma_l + \delta_m}_{\text{Main Effects}} + \\ & \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \alpha\delta_{jm} + \beta\gamma_{kl} + \beta\delta_{km} + \gamma\delta_{lm}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijklm}}_{\text{Error}} \end{aligned} \tag{3.2}$$

where: Y_{ijklm} is the score of the i-th subject in the j-th, k-th, l-th, and m-th factors; $\mu_{.....}$ is the grand mean; τ_i is the effect of the i-th subject, i.e. topics; α_j is the effect of the j-th factor, i.e. stop lists; β_k is the effect of the k-th factor, i.e. stemmers or n-grams; γ_l is the effect of the k-th factor, i.e. IR models; δ_m is the effect of the m-th factor, i.e. measures.

For this analysis, we also consider the interaction effects among these factors, e.g. $\gamma\delta_{lm}$ is the interaction between an IR model and a measure. We do not consider third and fourth order interactions such as $\alpha\beta\gamma_{jkl}$ reporting the interaction between stop lists, LUGs and IR models, because they are rarely significant.

Finally, ε_{ijklm} is the error committed by the model in predicting the score of the i-th subject in the four factors j, k, l, m.

3.3. Effect size, multiple comparisons, and power analysis

We are not only interested in determining whether the factor effect is significant, but also which proportion of the performances variance is due to it, that is we need to estimate its *effect-size measure* (*Strength of Association (SOA)*). The SOA is a "standardized index and estimates a parameter that is independent of sample size and quantifies the magnitude of the difference between populations or the relationship between explanatory and response variables" (Sakai, 2014). We use the $\hat{\omega}^2_{\langle fact \rangle}$ SOA:

$$\hat{\omega}^2_{\langle fact \rangle} = \frac{df_{fact}(F_{fact} - 1)}{df_{fact}(F_{fact} - 1) + N} \tag{3.3}$$

which is an unbiased estimator of the variance components associated with the sources of variation in the design. F_{fact} is the F-statistics and df_{fact} are the degrees of freedom for the factor while N is the total number of samples.

The common rule of thumb (Rutherford, 2011) when classifying $\hat{\omega}^2_{\langle fact \rangle}$ effect size is: 0.14 and above is a large effect, 0.06–0.14 is a medium effect, and 0.01–0.06 is a small effect. $\hat{\omega}^2_{\langle fact \rangle}$ values could happen to be negative and in such cases they are considered as zero.

When conducting experiments, two types of error may happen. A *Type I* error occurs when a true null hypothesis is rejected and the significance level α is the probability of committing a Type I error. A *Type I* error occurs when a true null hypothesis is rejected and the significance level α is the probability of committing a Type I error. We keep Type I errors controlled by applying the Tukey *Honestly Significant Difference (HSD)* test with a significance level $\alpha = 0.05$. Tukey's

method is used in ANOVA to create confidence intervals for all pairwise differences between a given factor level means while controlling the family error rate.

A *Type II* error occurs when a false null hypothesis is accepted and it is concerned with the capability of the conducted experiment to actually detect the effect under examination. Type II errors are often overlooked because if they occur, although a real effect is missed, no misdirection occurs and further experimentation is very likely to reveal the effect.

The *power* is the probability of correctly rejecting a false null hypothesis when an experimental hypothesis is true

$$\text{Power} = 1 - \beta$$

where β (typically $\beta = 0.2$) is the Type II error rate.

To determine the power of an experiment, we compute the effect size parameter:

$$\phi = \sqrt{N \cdot \frac{\hat{\omega}^2_{\langle fact \rangle}}{1 - \hat{\omega}^2_{\langle fact \rangle}}} \tag{3.4}$$

and we compare it with its tabulated values for a given Type I error rate α to determine β.

4. Grid of points, measures and setup

4.1. Grid of points (GoP)

We considered three main components of an IR system: stop list, LUG and IR model. We selected a set of alternative implementations of each component and, by using the Terrier[2] open source system (Macdonald et al., 2012), we created a run for each system defined by combining the available components in all possible ways. The components we selected are:

- *Stop list:* nostop, indri, lucene, snowball, smart, terrier;
- *LUG:* nolug, weak Porter, Porter, snowball Porter, Krovetz, Lovins, 4grams, 5grams, 6grams, 7grams, 8grams, 9grams, 10grams;
- *Model:* BB2, BM25, DFIZ, DFRee, DirichletLM, DLH, DPH, HiemstraLM, IFB2, InB2, InL2, InexpB2, Js_KLs, LemurTFIDF, LGD, PL2, TFIDF.

The stoplists differ from each other by the number of terms composing them; specifically, indri has 418 terms, lucene has 33 terms, snowball has 174 terms, smart has 571 terms and terrier 733 terms.

As for the LUG component, we consider two distinct classes: stemmers and n-grams. Stemmers can be classified into aggressive and weak stemmers. One of the first stemmers developed for IR systems is Lovins; this is an iterative affix removal stemmer which removes the longest possible string of characters from a word, according to a set of rules. Lovins is the most aggressive stemmer amongst those we consider. The Porter algorithm and its variants (snowball and a weaker version) is inspired by the Lovins algorithm, but it adds machine-readable dictionaries and well-defined rules for morphology. The Krovetz algorithm adds a word-disambiguation algorithm to the Porter stemmer. n-grams are LUG components alternative to the stemmer which can be defined as a

[2] `http://www.terrier.org/`

contiguous sequence of n items from a given sequence of text. Character n-grams are a language-independent alternative to complex language-specific tokenization. We consider seven different n-grams lengths ranging from $n = 4$ to $n = 10$; we did not employ 3-grams or smaller lengths because they are very ineffective for the English language and the runs employing this component would be treated as outliers in the experiments.

The models we employ are classified into the three main approaches currently adopted by search engines: (i) the vector space model: TFIDF and LemurTFIDF; (ii) the probabilistic model – including the BM25 models and the *Divergence From Randomness (DFR)* models, and in particular: BB2 (Bose-Einstein model), DFIZ, DFRee (hyper-geometric model), DLH (parameter free DFR model), DPH (DFR model with Popper's normalization), IFB2 (ITF with Bernoulli's processes), InB2 (a variant of the PL2), InL2 (IDF model for randomness with Laplace succession), InexpB2 and PL2 (Poisson estimation using Laplace succession)); and, (iii) the language models, and in particular: DirichletLM (LM with Bayesian smoothing and a Dirichlet Prior), HiemstraLM (Hiemstra's LM), Js_KLs (Jefrreys' divergence with Kullback Leibler's divergence) and LGD.

We consider the stemmers and n-grams and mutually exclusive LUG components because they represent very different approaches to lexical unit generation and, typically, without prejudice to the other components, they lead to alternative pipelines. Thus, we end up with two distinct groups of runs, one using the stemmers and one using the n-grams; the `nolug` component is common to both these groups instantiated as `nostemmer` for the stemmer group and as `nograms` for the n-grams one. The stemmer group defines a $6 \times 6 \times 17$ factorial design with a GoP consisting of 612 runs; the n-grams group defines a $6 \times 8 \times 17$ factorial design with a grid of points consisting of 816 runs.

Note that the above GoPs consider whole components and do not further break them down into sub-components. For example, IR models are not broken down into their constituents, as studied by Zobel and Moffat (1998). In addition, for IR models, we use the default setup of their parameters, as provided by Terrier, and we do not explore here variations of this setup, which would lead to much bigger GoPs: this is left for future work.[3].

4.2. Experimental collections

We used the following standard and shared collections: TREC 07 Adhoc track, TREC 08 Adhoc track, TREC 09 Web track and TREC 10 Web track. TREC 07 and TREC 08 focus on a news search task and adopt a corpus of about 528K news documents (i.e., disk 4 and 5 of the TIPSTER collection minus the Congressional Record); both TREC 07 and 08 provide 50 different topics and a set of binary relevance judgments – i.e., given a topic $t \in T$ a document $d \in D$ is judged either not relevant or relevant. TREC 09 and TREC 10 focus on a Web search task and adopt a corpus of 1.7M Web pages (i.e., the WT10g collection); both TREC 09 and 10 are composed of 50 different topics and a set of graded relevance judgments – i.e., not relevant, relevant and highly relevant.

We conducted the experiments on two collections composed of 100 topics each, one created by combining the topics of TREC 07 with those of TREC 08 and the topics of TREC 09 with those of TREC 10.

4.3. Evaluation measures

We evaluated the GoPs by employing nine different evaluation measures: AP, P@10, Rprec, RBP, nDCG, nDCG@20, ERR, ERR@20 and Twist.

[3]To ease reproducibility, the code is available at: `http://gridofpoints.dei.unipd.it/`.

Average Precision (AP) represents the "gold standard" measure in IR, known to be stable and informative, with a natural top-heavy bias and an underlying theoretical basis as approximation of the area under the precision/recall curve.

Precision at Ten (P@10) is the classic precision measure with cut-off at the first 10 retrieved documents. This measure is defined for binary relevance judgments and is particularly well-suited for evaluating Web tasks.

RPrec is precision calculated with cut-off at the recall base – i.e., the total number of relevant documents for a given topic determined by the ground truth.

Rank-Biased Precision (RBP) (Moffat and Zobel, 2008) is built around a user model based on the utility a user can achieve by using a system: the higher, the better. The model it implements is that a user always starts from the first document in the list and then s/he progresses from one document to the next with a probability p. We calculated RBP by setting $p = 0.8$ which represent a good trade-off between a very persistent and a remitting user.

These measures are based on binary relevant judgments and thus can be naturally applied to TREC 07 and TREC 08. For TREC 09 and TREC 10, we performed a lenient mapping of the relevance judgments by considering as relevant both highly relevant and relevant documents.

Normalized Discounted Cumulated Gain (nDCG) (Järvelin and Kekäläinen, 2002) is the normalized version of the widely-known DCG which discounts the gain provided by each relevant retrieved document proportionally to the rank at which it is retrieved. nDCG is defined for graded relevance judgments and it is one of the most common measures used for evaluating Web search tasks. For TREC 07 and 08, we calculated nDCG in a binary relevance setting by giving gain 0 to non-relevant documents and gain 5 to the relevant ones; whereas, for TREC 09 and TREC 10 we assign a weight 0 to non-relevant documents, 5 to the relevant ones and 10 to the highly relevant ones. Furthermore, we used a log_{10} discounting function. nDCG is calculated up to the last relevant retrieved document, whereas nDCG@20 is calculated up to rank position 20.

Expected Reciprocal Rank (ERR) (Chapelle et al., 2009) is a measure defined for graded relevance judgments and for evaluating navigational intent. It is particularly top-heavy since it highly penalizes systems placing not-relevant documents in high positions. For TREC 07 and TREC 08, we calculated ERR in a binary relevance setting as we have done for nDCG. ERR is calculated up to the last relevant retrieved document, whereas ERR@20 is calculated with cut-off at 20, thus focusing on the top retrieved documents.

Twist (Ferro et al., 2016) is a measure for informational intents, which handles both binary and graded relevance. Twist adopts a user model where the user scans the ranked list from top to bottom until s/he stops, and returns an estimate of the effort required by the user to traverse the ranked list. Twist evaluates systems from the viewpoint of the avoidable effort for their users by accounting for their fatigue while visiting a non-ideal ranking of documents; thus, it evaluates IR systems from a different angle i.e., user effort than other measures such as nDCG and ERR which are more focused on user's gain.

5. News search analysis

We use the three-way GLMM of Section 3.1 to carry out the analysis of IR system components. In the following, we focus on AP as reference measure; the analysis across measures is reported in Section 7.

Table 1 reports the estimated ω^2 SOA for all the main and interaction effects and, within parentheses, the p-values for all the ANOVA three-way tests we conducted; this table presents the

results divided by search task (news search in the upper part or Web search in the lower part) and, within a search task, by LUG group (stemmers or n-grams).

For the stemmers group, we can see that stop lists are a medium size effect followed by IR models and stemmers, which are both small size effects. All second- and third-order interactions between components are not significant with the notable exception of the Stop lists * IR models interaction, which is a medium size effect and the biggest among all the effects. Overall, this suggest that stops lists are a key component for the stemmers case and that they even influence the IR models.

From the main effects plot of the stop lists in the upper part of Figure 1, we can see that there is a substantial difference between systems employing or not a stop list while the impact of different stop lists is less marked, even though the corresponding Tukey HSD plot shows that the `lucene` stop list, i.e. the shortest one, is not in the top group, thus marking a significant loss in performances with respect to the others.

When it comes to stemmers, the main effects and the Tukey HSD plots show that the best performances are obtained by the `krovetz` stemmer, followed by the Porter-based stemmers; systems employing the `lovins` stemmer or not employing a stemmer at all report significant lower performances, even if `lovins` is still significantly better than no stemming.

Finally, as far as IR models are concerned, the top group is composed, in descending order, by `inexpb2`, `inb2`, and `inl2`, which are three variations of the DFR probabilistic model; it is also possible to identify a second group including the `tfidf` (a vector space model), `jskls` (a second generation DFR model) and `lgd` (a bridge between language and DFR models). It is interesting to note that, for news search, the effect of `bm25` alone is not in the first two top performing groups, despite the fact that it is one of the widest employed models in IR and often the default choice when employing off-the-shelf search engines.

The interaction plot in the upper right of Figure 1 shows the joint effects of stop lists and IR models: we can see that there is some interaction, in general, due to the fact that lines tend to cross but, mostly, we can see that there are IR models which significantly suffer from the lack of a stop list, namely `ifb2`, `bb2`, `pl2`, `bm25`, and `dfiz`. This means that a generally well performing model, as `bm25` is, changes performances a lot with or without a stop list – e.g. *Mean Average Precision (MAP)* 0.2381 with `lucene` stop list and `krovetz` stemmer versus MAP 0.1575 without stop list and without `krovetz` stemmer; it also implies that the same model may perform worse than a generally lower performing model, but less influenced by stop lists, as in the case of `dirichletlm` – MAP 0.1967 with `lucene` stop list and `krovetz` stemmer and MAP 0.1910 without stop list and without `krovetz` stemmer.

In particular, the small interaction between stop lists and language models has been argued by Zhai and Lafferty (2004): "the effects of stop word removal should be better achieved by exploiting language modeling techniques". The interaction plot in Figure 1 supports this claim since the effect of the stop list on `hiemstralm` is almost null and on `dirichletlm` is very small. Moreover, there is a greater interaction between the stop list and the `lgd` model, probably because this is a hybrid model that bridges between language models and the DFR models, which typically exhibit a sizable interaction with stop lists. Overall, we observe a kind of increasing interaction between stop lists and the family of probabilistic models: almost no interaction for the language models; some interaction for the DFR models; and, a lot of interaction for the classic `bm25` model. Finally, the vector space model, namely `tfidf` and `lemurtfidf`, shows almost no interaction with the stop lists.

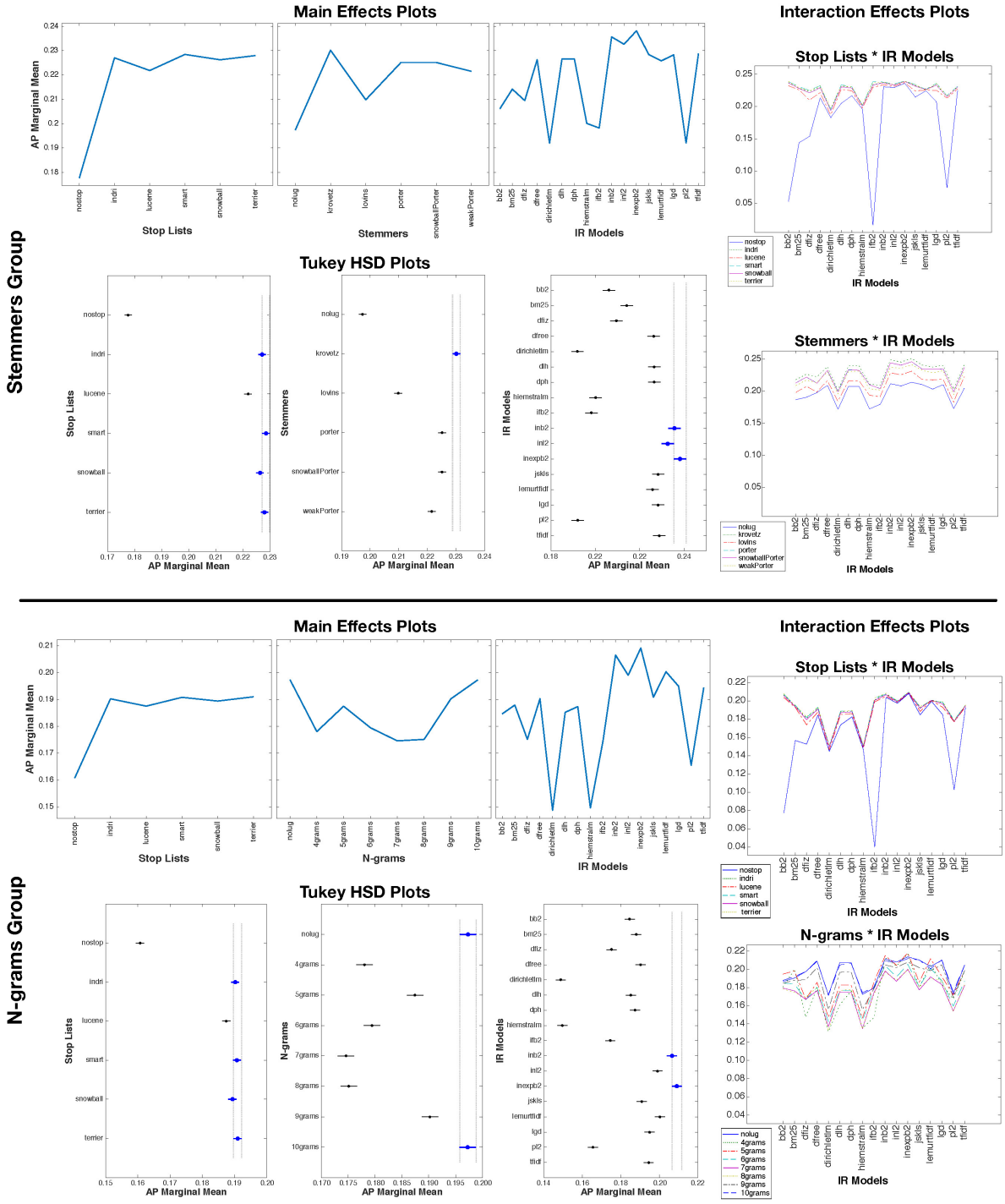

Figure 1: TREC 07-08 news search: main effects, interaction effects, and Tukey HSD plots for average precision on stemmers and n-grams groups.

Table 1: Summary of component effects for TREC 07-08 (news search) and TREC 09-10 (Web search). Each cell reports the estimated ω^2 SoA for the specified effects and, within parentheses, the p-value for those effects. Not significant effects are in light gray; small size effects are in light blue; medium size effects are in blue; and large size effects are in dark blue.

TREC 07-08 news search										
Lexical Units	**Effects**	**AP**	**P@10**	**R-prec**	**RBP**	**nDCG@20**	**nDCG**	**ERR@20**	**ERR**	**Twist**
Stemmers	$\hat{\omega}^2_{\langle \text{Stop Lists} \rangle}$	0.0677 (0.00)	0.0675 (0.00)	0.0819 (0.00)	0.0766 (0.00)	0.0770 (0.00)	0.1417 (0.00)	0.0581 (0.00)	0.0587 (0.00)	0.1041 (0.00)
	$\hat{\omega}^2_{\langle \text{Stemmers} \rangle}$	0.0266 (0.00)	0.0127 (0.00)	0.0228 (0.00)	0.0125 (0.00)	0.0167 (0.00)	0.0527 (0.00)	0.0031 (0.00)	0.0030 (0.00)	0.0327 (0.00)
	$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0465 (0.00)	0.0474 (0.00)	0.0598 (0.00)	0.0546 (0.00)	0.0594 (0.00)	0.0847 (0.00)	0.0431 (0.00)	0.0431 (0.00)	0.0712 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	-0.0000 (0.47)	-0.0000 (0.52)	-0.0001 (0.69)	-0.0001 (0.81)	-0.0001 (0.68)	0.0001 (0.31)	-0.0002 (0.99)	-0.0002 (0.99)	-0.0000 (0.59)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.1169 (0.00)	0.1432 (0.00)	0.1630 (0.00)	0.1615 (0.00)	0.1660 (0.00)	0.2751 (0.00)	0.1329 (0.00)	0.1347 (0.00)	0.2175 (0.00)
	$\hat{\omega}^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0005 (1.00)	0.0005 (0.01)	-0.0007 (1.00)	-0.0001 (0.62)	-0.0006 (1.00)	-0.0006 (1.00)	-0.0001 (0.59)	-0.0001 (0.58)	-0.0003 (0.92)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0060 (1.00)	-0.0056 (1.00)	-0.0059 (1.00)	-0.0061 (1.00)	-0.0060 (1.00)	-0.0057 (1.00)	-0.0059 (1.00)	-0.0059 (1.00)	-0.0058 (1.00)
N-grams	$\hat{\omega}^2_{\langle \text{Stop Lists} \rangle}$	0.0222 (0.00)	0.0282 (0.00)	0.0268 (0.00)	0.0305 (0.00)	0.0301 (0.00)	0.0392 (0.00)	0.0232 (0.00)	0.0235 (0.00)	0.0343 (0.00)
	$\hat{\omega}^2_{\langle \text{N-grams} \rangle}$	0.0146 (0.00)	0.0068 (0.00)	0.0182 (0.00)	0.0067 (0.00)	0.0141 (0.00)	0.0248 (0.00)	0.0080 (0.00)	0.0080 (0.00)	0.0216 (0.00)
	$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0522 (0.00)	0.0521 (0.00)	0.0553 (0.00)	0.0559 (0.00)	0.0593 (0.00)	0.0641 (0.00)	0.0368 (0.00)	0.0369 (0.00)	0.0575 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{N-grams} \rangle}$	0.0010 (0.00)	0.0007 (0.00)	0.0011 (0.00)	0.0008 (0.00)	0.0009 (0.00)	0.0024 (0.00)	0.0009 (0.00)	0.0010 (0.00)	0.0017 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.0536 (0.00)	0.0749 (0.00)	0.0711 (0.00)	0.0851 (0.00)	0.0826 (0.00)	0.1117 (0.00)	0.0734 (0.00)	0.0749 (0.00)	0.0986 (0.00)
	$\hat{\omega}^2_{\langle \text{N-grams} \times \text{IR Models} \rangle}$	0.0079 (0.00)	0.0057 (0.00)	0.0066 (0.00)	0.0062 (0.00)	0.0056 (0.00)	0.0123 (0.00)	0.0050 (0.00)	0.0051 (0.00)	0.0090 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{N-grams} \times \text{IR Models} \rangle}$	-0.0048 (1.00)	-0.0043 (1.00)	-0.0039 (1.00)	-0.0040 (1.00)	-0.0040 (1.00)	-0.0016 (1.00)	-0.0037 (1.00)	-0.0036 (1.00)	-0.0026 (1.00)
TREC 09-10 Web search										
Lexical Units	**Effects**	**AP**	**P@10**	**R-prec**	**RBP**	**nDCG@20**	**nDCG**	**ERR@20**	**ERR**	**Twist**
Stemmers	$\hat{\omega}^2_{\langle \text{Stop Lists} \rangle}$	0.0617 (0.00)	0.0465 (0.00)	0.0467 (0.00)	0.0543 (0.00)	0.0696 (0.00)	0.1628 (0.00)	0.0308 (0.00)	0.0313 (0.00)	0.0728 (0.00)
	$\hat{\omega}^2_{\langle \text{Stemmers} \rangle}$	0.0066 (0.00)	0.0015 (0.00)	0.0030 (0.00)	0.0013 (0.00)	0.0025 (0.00)	0.0184 (0.00)	0.0009 (0.00)	0.0009 (0.00)	0.0041 (0.00)
	$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0879 (0.00)	0.0727 (0.00)	0.0618 (0.00)	0.0955 (0.00)	0.0806 (0.00)	0.1543 (0.00)	0.0664 (0.00)	0.0659 (0.00)	0.0726 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	-0.0002 (0.99)	-0.0002 (0.92)	-0.0003 (1.00)	-0.0003 (1.00)	-0.0003 (1.00)	-0.0001 (0.67)	-0.0001 (0.79)	-0.0001 (0.77)	-0.0002 (0.95)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.0917 (0.00)	0.1059 (0.00)	0.0899 (0.00)	0.1224 (0.00)	0.1383 (0.00)	0.2967 (0.00)	0.0860 (0.00)	0.0878 (0.00)	0.1408 (0.00)
	$\hat{\omega}^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0004 (0.97)	-0.0002 (0.85)	-0.0002 (0.88)	-0.0005 (1.00)	-0.0003 (0.90)	0.0002 (0.21)	-0.0006 (1.00)	-0.0006 (1.00)	-0.0003 (0.95)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0061 (1.00)	-0.0059 (1.00)	-0.0058 (1.00)	-0.0062 (1.00)	-0.0059 (1.00)	-0.0061 (1.00)	-0.0057 (1.00)	-0.0057 (1.00)	-0.0056 (1.00)
N-grams	$\hat{\omega}^2_{\langle \text{Stop Lists} \rangle}$	0.0210 (0.00)	0.0223 (0.00)	0.0188 (0.00)	0.0244 (0.00)	0.0247 (0.00)	0.0493 (0.00)	0.0149 (0.00)	0.0152 (0.00)	0.0312 (0.00)
	$\hat{\omega}^2_{\langle \text{N-grams} \rangle}$	0.0399 (0.00)	0.0277 (0.00)	0.0369 (0.00)	0.0352 (0.00)	0.0317 (0.00)	0.0615 (0.00)	0.0240 (0.00)	0.0240 (0.00)	0.0546 (0.00)
	$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0646 (0.00)	0.0802 (0.00)	0.0498 (0.00)	0.0978 (0.00)	0.0729 (0.00)	0.0978 (0.00)	0.0676 (0.00)	0.0673 (0.00)	0.0713 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{N-grams} \rangle}$	0.0022 (0.00)	0.0013 (0.00)	0.0017 (0.00)	0.0019 (0.00)	0.0012 (0.00)	0.0028 (0.00)	0.0017 (0.00)	0.0017 (0.00)	0.0023 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.0433 (0.00)	0.0638 (0.00)	0.0424 (0.00)	0.0723 (0.00)	0.0650 (0.00)	0.1292 (0.00)	0.0487 (0.00)	0.0500 (0.00)	0.0790 (0.00)
	$\hat{\omega}^2_{\langle \text{N-grams} \times \text{IR Models} \rangle}$	0.0049 (0.00)	0.0076 (0.00)	0.0037 (0.00)	0.0084 (0.00)	0.0093 (0.00)	0.0133 (0.00)	0.0044 (0.00)	0.0044 (0.00)	0.0055 (0.00)
	$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{N-grams} \times \text{IR Models} \rangle}$	-0.0044 (1.00)	-0.0034 (1.00)	-0.0042 (1.00)	-0.0036 (1.00)	-0.0038 (1.00)	-0.0018 (1.00)	-0.0042 (1.00)	-0.0041 (1.00)	-0.0023 (1.00)

The interaction between stemmers and IR models reported in Table 1 is almost null and not significant but, as we will discuss in Section 7, we are lacking the statistical power to be able to detect it. Nevertheless, if we consider the interaction plot for stemmers * IR models shown in Figure 1, with the previous caveat about the size and significance of the effect, we notice that all the IR models are somehow affected by the use of a stemmer and some of them – `bm25`, `hiemstralm`, `inb2`, `inl2`, `inexpb2`, `lemurtfidf` – seems to be affected more by the lack of a stemmer.

Overall, this supports the importance of the stop lists noted above and it is quite a striking result since, nowadays, it is commonly thought that stop lists were more useful in the past to reduce the index size and skip useless calculations for efficiency reasons while their impact on effectiveness should have been limited and basically hidden by IR models, since they assign almost zero weight to stop words.

For the n-grams group, we can see from Table 1 that all the effects, with the exception of third order interactions, become significant even if they are all now small size effects, compared to the small and medium size they were in the case of the stemmers group. We can observe that the effect of IR models becomes more pronounced than the effect of the stop lists, while n-grams have less of an impact than stemmers, about half of the effect size. Moreover, the interaction Stop lists * IR models is as big as the IR models effect alone and this suggests that stop lists play a central role also in the case of n-grams. The interaction between stop lists and n-grams and IR models and n-grams, even though significant, is almost negligible in terms of size.

From the lower part of Figure 1, as far as the main effects of stop lists are concerned, we observe the same behaviour as in the case of stemmers: there is a marked difference between using or not a stop list and `lucene` stop list is the only one outside the top group. Regarding the main effects of n-grams, they are not effective for news search, since the top performing systems are those not employing n-grams or those employing the 10-grams, which is a very mild form of character sequencing. In particular, we can see that the 5-grams component is statistically as effective as the 9-grams component – i.e., they are the second best performing group; from 6- to 8-grams we can see a progressive degradation of performances. For the IR models effect, the top group consists of `inexpb2` and `inb2`, while the second top group consists of `lemurtfidf` and `inl2`; therefore, in comparison to the stemmers group, `inexpb2` and `inb2` confirm themselves as the top performing approaches, while in the second groups we always find a vector space model (either `tfidf` or `lemurtfidf`). Regarding the interaction between stop lists and IR models, we observe also in the n-grams case a behaviour similar to that of the stemmers group: `ifb2`, `bb2`, `pl2`, `bm25`, and `dfiz` models greatly suffer from the lack of a stop list.

The interaction plot of n-grams with IR models shows marked interaction. Moreover, even though systems employing `nolug` perform in general better than those employing no form of character sequencing, some models benefit from employing a 4- or 5-grams component, as with `bm25`, `inexpb2`, and `lemurtfidf` where the use of 5-grams increases performances with respect to not using them.

Overall, we can appreciate that n-grams interact with IR models more than stemmers, even though they are less effective than stemmers for English retrieval as also noticed by Büttcher et al. (2010). Moreover, stop lists play a central role both alone and in conjunction with IR models and their importance is basically independent of the LUG component adopted.

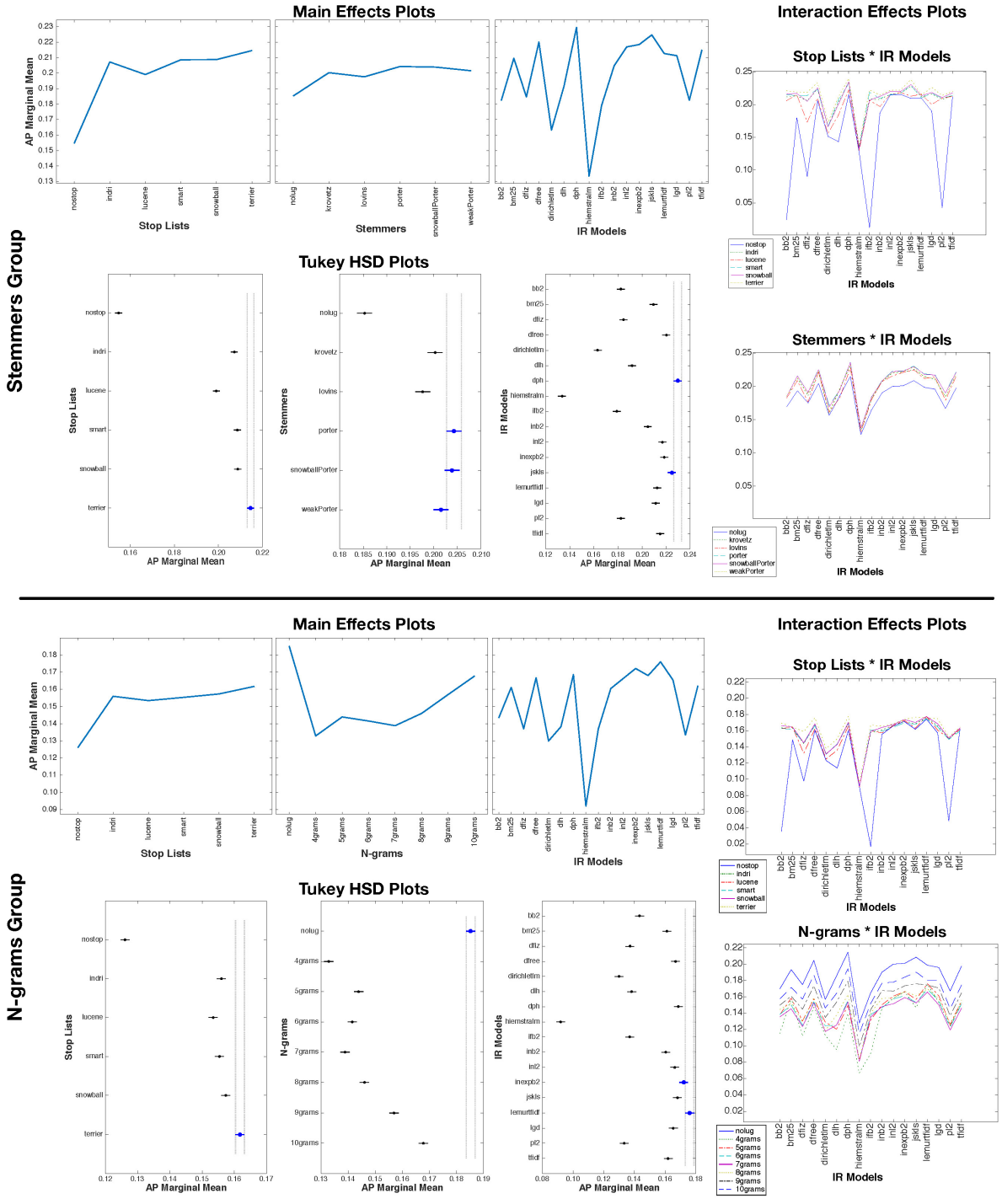

Figure 2: TREC 09-10 Web search: main effects, interaction effects, and Tukey HSD plots for average precision on stemmers and n-grams groups.

6. Web search analysis

Table 1 shows the Web search task, which displays a trend similar to the case of the news search tasks, even though it is more pronounced and with some notable differences we will discuss.

For the stemmer group, the stop lists and IR models yield medium size effects while the stemmers show a small size effect, almost negligible in this case; regarding interaction effects, the Stop lists * IR models yields the biggest effect, which is of medium size, while all the others are not significant.

From the main effects and Tukey HSD plots of the stop lists in Figure 2, we can see that, again, there is a substantial difference between systems employing or not a stop list, while the top performing stop list is `terrier`, i.e. the longest one, followed by a second group constituted by `snowball`, `smart`, and `indri`; `lucene` is still the lowest performing stop lists. The fact that there is a clearer distinction between stop lists than in the news search case, and that the longest stop list is the top performing one, lead us to hypothesize that the noisy Web context benefits more from the aggressive filtering of a longer stop list.

As far as stemmers are concerned, the Porter-based stemmers constitute the top group in the case of Web search, while `krovetz` and `lovins` stay together in the second group, well above the group employing no stemmer at all. With respect to the news search case, the less aggressive stemmers perform better for Web search and this may be motivated again by the hypothesis that the noisy Web context benefits more from avoiding further noise due to over-stemming.

Regarding IR models, the top group is constituted by `dph`, a DFR probabilistic model, and `jskls`, a second generation DFR model requiring no parameter tuning, while the second group includes `dfree`, `inexpb2`, `inl2`, and `tfidf`; therefore, we observe a kind of swap between the first and the second group with respect to the news search case.

The interaction plot in the upper right of Figure 2 shows the joint effects of stop lists and IR models: it shows even more interaction with respect to the news search case and the number of IR models which severely suffer from a lack of a stop list increases – now `ifb2`, `bb2`, `pl2`, `dfiz`, `dlh`, and `bm25` – even though in this case `bm25` is much less impacted from the lack of a stop list. Moreover, language models exhibit a higher interaction with stop lists for the Web search task than for news search, in particular for `dirichletlm`; by contrast, both the vector space models, i.e. `tfidf` and `lemurtfidf`, confirm a low interaction with the stop lists as observed for the news search task.

Regarding the interaction between stemmers and IR models, bearing in mind the caveats pointed out in the previous section, in the Web search scenario a number of models seem to be affected by the lack of a stemmer, namely `bm25`, `inb2`, `inl2`, `inexpb2`, `jskls`, `lemurtfidf`, `lgd`, `pl2`, and `tfidf`. Moreover, in this case, the difference in interaction between not using and using a stemmer is higher than the difference between the use of different stemmers.

For the n-grams group, all the effects, apart from third order interactions, are significant and the IR models yield the biggest effect, a medium size one, followed by n-grams, Stop lists * IR models interaction, and Stop lists, which are all small size effects. Regarding the news search task, we can observe the more pronounced role of the IR Models and n-grams effects, while the impact of stop lists is a bit reduced, with the interaction Stop lists * IR models still being the second biggest effect. From the lower part of Figure 2, we can note again the impact of using a stop list also in the case of n-grams with `terrier`, which is the top performing stop list, followed by all the others now in the same second group. This supports the hypothesis of the benefits of a longer stop list in the noisy Web context, made even noisier by the application of n-grams. The n-grams component

Table 2: Statistical power ($1-\beta$ error probability) for news search (TREC 07-08) and Web search (TREC 09-10) with $\alpha = 0.05$. High power (0.90-1.00) is highlighted in dark blue; good power (0.80-0.90) is in medium blue; low power (0.70-0.80) is in light blue; insufficient power (below 0.70) is in light gray. SL×M is an abbreviation for the Stop list * IR models and LUG×M for the LUG * IR models.

TREC 07-08 news search										
LUGs	**Power**	**AP**	**P@10**	**R-prec**	**RBP**	**nDCG@20**	**nDCG**	**ERR@20**	**ERR**	**Twist**
Stemmers	Stop Lists	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Stemmers	0.82	0.46	0.75	0.46	0.59	0.99	0.13	0.12	0.91
	IR Models	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SL×M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LUG×M	0.10	0.18	0.08	0.13	0.09	0.09	0.14	0.14	0.14
N-grams	Stop Lists	0.74	0.85	0.83	0.88	0.88	0.95	0.76	0.77	0.92
	N-grams	0.63	0.30	0.75	0.30	0.62	0.89	0.36	0.36	0.83
	IR Models	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SL×M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LUG×M	0.96	0.86	0.92	0.90	0.85	1.00	0.81	0.82	0.98
TREC 09-10 Web search										
Stemmers	Stop Lists	1.00	0.98	0.98	1.00	1.00	1.00	0.88	0.89	1.00
	Stemmers	0.24	0.09	0.13	0.08	0.11	0.64	0.07	0.07	0.16
	IR Models	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SL×M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LUG×M	0.11	0.12	0.12	0.10	0.12	0.15	0.09	0.09	0.11
N-grams	Stop Lists	0.71	0.74	0.65	0.78	0.79	0.99	0.53	0.54	0.89
	N-grams	0.99	0.93	0.98	0.98	0.96	1.00	0.87	0.88	1.00
	IR Models	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SL×M	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LUG×M	0.79	0.95	0.67	0.97	0.98	1.00	0.76	0.76	0.85

always performs significantly worse than no n-grams at all and it seems to be more harmful for Web search than for news search.

Regarding IR models, the top group contains `lemurtfidf`, a vector space model, and `inexpb2`, a DFR probabilistic model, which were, respectively, in the second and the first top group in the case of news search. If we look at the interaction between stop lists and IR models, we can observe the same list of models suffering from the lack of a stop list, although `bm25` is less impacted than in the previous case. The interaction plot between n-grams and IR models shows that there is no model benefiting from n-grams, unlike the news search case where some models benefited from using a 4- or 5- grams component.

7. Power analysis and measures analysis

Table 1 also reports the estimated ω^2 SOA for all the other considered measures; this emphasises the difference between the various measures when it comes to breaking down component effects.

In general, all the measures agree on which effects are not significant, for both the stemmers and the n-grams groups and both the news and Web search tasks, with the only exception being P@10 where the Stemmers*IR models interaction is significant for the stemmers group in the news search tasks, even though its effect size is almost negligible.

It is interesting to note that the effect sizes with nDCG and Twist are bigger than with the other measures; in particular, nDCG and Twist greatly enhance the impact of stop lists, IR models and their interaction as well as putting more emphasis on LUG, both stemmers and n-grams. On the other hand, ERR and ERR@20 show the opposite behaviour by reducing the effect size of these components with respect to AP; indeed, for these measures the effect size of stop lists is small and not medium, as it is with all the other measures, and the impact of LUG, both stemmers and n-grams, becomes almost negligible.

This suggests that not all the measures are equally good at detecting differences between the components and they have quite a diverse perception of what the internals of an IR system are.

Table 3: Summary of component and measure effects for TREC 07-08 and TREC 09-10. Each cell reports the estimated ω^2 SoA for the specified effects and, within parentheses, the p-value for those effects, and their statistical power. Not significant effects are in light gray; small size effects are in light blue; medium size effects are in blue; and, large size effects are in dark blue. High power (0.90-1.00) is highlighted in dark blue; good power (0.80-0.90) is in medium blue; low power (0.70-0.80) is in light blue; insufficient power (below 0.70) is in light gray.

Lexical Units		TREC 07-08 News search		TREC 09-10 Web search	
		$\hat{\omega}^2$	power	$\hat{\omega}^2$	power
Stemmers	Stop Lists	0.0408 (0.00)	0.96	0.0264 (0.00)	0.82
	Stemmers	0.0067 (0.00)	0.24	0.0009 (0.00)	0.07
	IR Models	0.0277 (0.00)	1.00	0.0377 (0.00)	1.00
	Measures	0.4397 (0.00)	1.00	0.2597 (0.00)	1.00
	Stop Lists×Stemmers	0.0001 (0.00)	0.06	-0.0000 (0.91)	0.10
	Stop Lists×IR Models	0.0906 (0.00)	1.00	0.0598 (0.00)	1.00
	Stop Lists×Measures	0.0038 (0.00)	0.54	0.0014 (0.00)	0.17
	Stemmers×IR Models	0.0002 (0.00)	0.31	0.0001 (0.00)	0.07
	Stemmers×Measures	0.0004 (0.00)	0.09	0.0005 (0.00)	0.09
	IR Models×Measures	0.0038 (0.00)	0.13	0.0040 (0.00)	0.13
N-grams	Stop Lists	0.0154 (0.00)	0.54	0.0118 (0.00)	0.42
	N-grams	0.0053 (0.00)	0.23	0.0178 (0.00)	0.73
	IR Models	0.0267 (0.00)	1.00	0.0388 (0.00)	1.00
	Measures	0.3784 (0.00)	1.00	0.2245 (0.00)	1.00
	Stop Lists×N-grams	0.0008 (0.00)	0.10	0.0010 (0.00)	0.13
	Stop Lists×IR Models	0.0456 (0.00)	1.00	0.0339 (0.00)	1.00
	Stop Lists×Measures	0.0017 (0.00)	0.21	0.0007 (0.00)	0.10
	N-grams×IR Models	0.0035 (0.00)	0.83	0.0034 (0.00)	0.82
	N-grams×Measures	0.0013 (0.00)	0.20	0.0008 (0.00)	0.13
	IR Models×Measures	0.0031 (0.00)	0.81	0.0041 (0.00)	0.93

To further investigate this issue, we compute the statistical power using the G*Power tool[4]: Table 2 reports the statistical power of the main effects and the most interesting interaction effects. We can observe that, in general, we have good or high power in most of the cases, which ensures we are able to actually detect the effects we are looking for. In particular, the best and most robust measures in terms of statistical power are nDCG, Twist and AP, since they exhibit enough power in almost all the circumstances; on the other hand, the lower performing measures are ERR and ERR@20, closely followed by other top-heavy measures such as P@10, nDCG@20 and RBP.

More in detail, we can note how ERR and ERR@20 do not have enough power to detect the LUG effects, both stemmers and n-grams on both news and Web search tasks, and how they also lack some power for detecting stop lists effects when n-grams are employed. This confirms that ERR and its variants are not robust measures since they require more topics than other measures to detect reliable effect sizes (Sakai, 2016).

In general, stemmer effects are the most difficult to detect for most measures, the Web search task being more challenging than news search. It is generally thought that stemming is not particularly important for English retrieval, especially with respect to other European languages, but these analyses show that we may actually be in the situation of not properly detecting its effects due to the lack of power. Since p-values and significance are affected by the sample size and we would need to increase it to obtain enough power, it may be also worth reconsidering the stemmers * IR models interaction which might not be properly detected in the current settings.

To further investigate the impact of measures, we employ the four-way GLMM of Section 3.2 to carry out the analysis of IR system components by considering the measures as a factor contributing to explain the system variance. In Table 3 we report the estimated ω^2 SOA for all the main and interaction effects, the p-values for all the ANOVA four-way tests we conducted and the statistical power. This table presents the results divided by experimental collection and, within a collection, by LUG group.

[4] `http://www.gpower.hhu.de/`

The most noticeable phenomenon is the huge proportion of variance explained by the measures whose effect sizes are up to three orders of magnitude bigger than those of the IR components and that the interaction between components and measures is always significant even if we have a small size effect. Furthermore, this analysis allows us to point out the contribution of IR components and their interactions, without considering the influence of the evaluation measures. In this respect, the main trends observed in the previous sections are supported: for the news search task, stop lists yield the biggest effect, followed by IR models and stemmers or n-grams; and, for the Web search task, IR models yield the biggest effect followed by n-grams, stop lists and stemmers. It is interesting to note that, independently of the measure effects, almost all the second-order interactions are now significant even though of small/negligible size.

Regarding the power, we observe a loss of power with respect to the 3-way analysis of Sections 5 and 6. This gives us a feeling about how much the variability between measures in detecting component effects actually impacts our analyses and the conclusions we draw.

8. Conclusions and Future Work

We addressed a previously unsolved issue in the IR evaluation methodology and proposed an innovative solution to break-down the performances of an IR system into the effects of its components, in order to understand their contributions to the overall performances as well as their interactions.

To this end, we developed an analysis methodology consisting of two elements: a *Grid of Points (GoP)* created on standard experimental collections, where all the combinations of system components under examination are considered; and a GLMM to decompose the contribution of these components to the overall system variance, paired with some graphical tools to easily assess the main and interaction effects. Note that such a controlled experimental setting is typically not available in evaluation campaigns, such as TREC, where participating systems do not constitute a systematic sampling of all the possible combinations of components and often are not even described in such detail to know exactly what components have been used.

We conducted a thorough experimentation on TREC collections, considering two different tasks – news search and Web search – and adopting several evaluation measures. For each task, we created GoPs consisting of nearly all the state-of-the-art stop lists, stemmers and n-grams, and IR models available for English retrieval.

We found that the most prominent effects are those of stop lists and IR models, as well as their interactions, while stemmers and n-grams play a smaller role. As a general observation, we found that linguistic resources play a prominent role on English retrieval and that their interaction with IR models can have such a big influence to completely change the model performances.

We also highlighted the lack of some statistical power to properly detect the effects of stemmers and their interaction with IR models. This might be a consequence of the common knowledge that stemming does not have much impact on English retrieval. Nevertheless, these conclusions should be further validated with a sample size big enough to reliably detect these effects.

Finally, we have seen that measures explain a large portion of system variance and that different measures detect system and component effects differently. We conclude that not all the measures are suitable for all the cases, as for ERR and ERR@20 which almost fail to detect the stemmer effect. On the other hand, nDCG, Twist and AP proved themselves to be stable and robust, being able to detect components effect size with good or even high statistical power.

As far as future work is concerned, we plan to extend the proposed methodology in order to also be able to capture interaction between topics/systems and topics/components. Indeed, to estimate interaction effects, more replicates would be needed for each (topic, system) pair, as Robertson and Kanoulas (2012) simulated, and they are not possible in the present settings, since running the same system on the same topics more than once produces exactly the same results.

Furthermore, we will investigate how to apply the proposed methodology in the context of the analysis of *MultiLingual Information Access (MLIA)* components. We have already started to prepare GoPs in many European languages using CLEF collections (Ferro and Silvello, 2016b). To study them, we need to extend the current methodology to allow for comparisons across languages. Indeed, unlike the case handled in this paper, within each component family, we are actually experimenting with components that differ from language to language, e.g. an aggressive French stemmer is intrinsically different from a German one. As a consequence, this setting may require a GLMM which makes use of random effects instead of the fixed ones currently used for representing the various components.

List of Acronyms

ANOVA ANalysis Of VAriance

AP Average Precision

CLEF Conference and Labs of the Evaluation Forum

DCG Discounted Cumulated Gain

DFR Divergence From Randomness

ERR Expected Reciprocal Rank

GLMM General Linear Mixed Model

GoP Grid of Points

HSD Honestly Significant Difference

IR Information Retrieval

LUG Lexical Unit Generator

MAP Mean Average Precision

MLIA MultiLingual Information Access

nDCG Normalized Discounted Cumulated Gain

RBP Rank-Biased Precision

SOA Strength of Association

TREC Text REtrieval Conference

References

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

Büttcher, S., Clarke, C. L. A., and Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, Cambridge (MA), USA.

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

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

Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

Croft, W. B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley, Reading (MA), USA.

Ferro, N. (2017). What Does Affect the Correlation Among Evaluation Measures? *ACM Transactions on Information Systems (TOIS)*.

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

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

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

Ferro, N. and Silvello, G. (2016b). The CLEF Monolingual Grid of Points. In Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016)*, pages 16–27. Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany.

Ferro, N., Silvello, G., Keskustalo, H., Pirkola, A., and Järvelin, K. (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):620–648.

Fuhr, N. (2010). IR between Science and Engineering, and the Role of Experimentation. In Agosti, M., Ferro, N., Peters, C., de Rijke, M., and Smeaton, A., editors, *Multilingual and Multimodal Information Access Evaluation. Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010)*, page 1. Lecture Notes in Computer Science (LNCS) 6360, Springer, Heidelberg, Germany.

Fuhr, N. (2012). Salton Award Lecture: Information Retrieval As Engineering Science. *SIGIR Forum*, 46(2):19–28.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Jayasinghe, G. K., Webber, W., Sanderson, M., Dharmasena, L. S., and Culpepper, J. S. (2015). Statistical comparisons of non-deterministic IR systems using two dimensional variance. *Information Processing & Management*, 51(5):677–694.

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

Lin, J., Crane, M., Trotman, A., Callan, J., Chattopadhyaya, I., Foley, J., Ingersoll, G., Macdonald, C., and Vigna, S. (2016). Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G. M., Hauff, C., and Silvello, G., editors, *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*, pages 357–368. Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany.

Macdonald, C., McCreadie, R., Santos, R. L. T., and Ounis, I. (2012). From Puppy to Maturity: Experiences in Developing Terrier. In Trotman, A., Clarke, C. L. A., Ounis, I., Culpepper, J. S., Cartright, M.-A., and Geva, S., editors, *Proc. SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 60–63.

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

Robertson, S. E. (1981). The methodology of information retrieval experiment. In Spärck Jones, K., editor, *Information Retrieval Experiment*, pages 9–31. Butterworths, London, United Kingdom.

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

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

Sakai, T. (2014). Statistical Reform in Information Retrieval? *SIGIR Forum*, 48(1):3–12.

Sakai, T. (2016). Topic set size design. *Information Retrieval*, 19(3):256–283.

Sanderson, M., Turpin, A., Zhang, Y., and Scholer, F. (2012). Differences in Effectiveness Across Sub-collections. In Chen, X., Lebanon, G., Wang, H., and Zaki, M. J., editors, *Proc. 21st International Conference on Information and Knowledge Management (CIKM 2012)*, pages 1965–1969. ACM Press, New York, USA.

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

Trotman, A., Clarke, C. L. A., Ounis, I., Culpepper, J. S., Cartright, M.-A., and Geva, S. (2012). Open Source Information Retrieval: a Report on the SIGIR 2012 Workshop. *ACM SIGIR Forum*, 46(2):95–101.

Trotman, A., Puurula, A., and Burgess, B. (2014). Improvements to BM25 and Language Models Examined. In Culpepper, J. S., Park, L., and Zuccon, G., editors, *Proc. 19th Australasian Document Computing Symposium (ADCS 2014)*, pages 58–65. ACM Press, New York, USA.

Zhai, C. and Lafferty, J. D. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information System (TOIS)*, 22(2):179–214.

Zhang, P., Dawei, S., Wang, J., and Hou, Y. (2014). Bias–variance analysis in estimating true query model for information retrieval. *Information Processing & Management*, 50(1):199–217.

Zobel, J. and Moffat, A. (1998). Exploring the Similarity Space. *SIGIR Forum*, 32(1):18–34.