What Does Affect the Correlation Among Evaluation Measures?

NICOLA FERRO, University of Padua

Information Retrieval (IR) is well-known for the great number of adopted evaluation measures, with new ones popping up more and more frequently. In this context, correlation analysis is the tool used to study the evaluation measures themselves and to let us understand if two measures rank systems similarly, if they grasp different aspects of system performances or actually reflect different user models, if a new measure is well motivated or not. To this end, the two most commonly used correlation coefficients are the Kendall's τ correlation and the AP correlation τ_{AP}.

The goal of the paper is to investigate the properties of the tool itself, i.e. correlation analysis, we use to study evaluation measures. In particular, we investigate three research questions about these two correlation coefficients: (i) what is the effect of the number of systems and topics? (ii) what is the effect of removing low performing systems? (iii) what is the effect of the experimental collections?

To answer these research questions, we propose a methodology based on *General Linear Mixed Model (GLMM)* and *ANalysis Of VAriance (ANOVA)* in order to isolate the effects of the number of topics, number of systems, and experimental collections and to let us observe expected correlation values, net from these effects, which are stable and reliable.

We learned that the effect of the number of topics is more prominent than the one of the number of systems. Even if it produces different absolute values, the effect of removing low performing systems does not seem to provide information substantially different from not removing them, especially when comparing a whole set of evaluation measures. Finally, we found out that both document corpora and topic sets affect the correlation among evaluation measures, being the effect of the latter more prominent. Moreover, there is a substantial interaction between evaluation measures, corpora and topic sets, meaning that the correlation between different evaluation measures can be substantially increased or decreased depending on the different corpora and topics at hand.

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

Additional Key Words and Phrases: correlation analysis, Kendall's tau correlation, AP correlation, evaluation measures, general linear mixed models (GLMM), analysis of variance (ANOVA), grid of points (GoP)

ACM Reference format:
Nicola Ferro. 2017. What Does Affect the Correlation Among Evaluation Measures?. *ACM Transactions on Information Systems* 0, 0, Article 0 (June 2017), 39 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Correlation analysis plays a central role in *Information Retrieval (IR)* evaluation where it is one of the tools we use to study properties and relationships among evaluation measures. When a new evaluation measure is proposed, correlation analysis is used to assess how the new measure ranks

Author's address: N. Ferro, Dept. Information Engineering, Via G. Gradenigo, 6/B, 35131 Padova, Italy; email: ferro@dei.unipd.it.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. 1046-8188/2017/6-ART0 $15.00
DOI: 0000001.0000001

IR systems with respect to the other existing measures and, thus, to understand whether it actually grasps different aspects of the systems and its introduction is somehow motivated [23, 32, 52, 61, 63, 81]. In this context, the most used correlation coefficients are the Kendall's tau correlation τ [41] and the AP correlation τ_{AP} [87].

Correlation analysis works as follows: let m_1 and m_2 be two evaluation measures; for example, let m_1 be *Average Precision (AP)*. Let M_1 and M_2 be two $t \times s$ matrices where each cell contains the performances on topic i of system j according to measures m_1 and m_2, respectively. Therefore, M_1 and M_2 represent the performances of s different systems (columns) over t topics (rows); for example, M_1 contains the AP score of each system on each topic. Let $\overline{M}_1$ and $\overline{M}_2$ be the column-wise averages of the two matrices; for example, $\overline{M}_1$ is a vector where each element is the *Mean Average Precision (MAP)* of a system. If you sort systems by their score in $\overline{M}_1$ and $\overline{M}_2$, you obtain two *Rankings of Systems (RoS)* corresponding to m_1 and m_2, respectively. A correlation coefficient like τ or τ_{AP} is then used to quantify how "close" these two RoS are; if the RoS produced by m_1 and m_2 are too "close", then m_1 and m_2 are actually measuring very similar aspects of the systems and they do not provide substantially different information.

The goal of the paper is to investigate the properties of correlation analysis, i.e. the tool itself we use to study evaluation measures. In particular, we carry out a thorough experimental study on the factors affecting the τ and τ_{AP} correlation coefficients and answer the following research questions

RQ1 What is the effect of the number of systems and topics?
RQ2 What is the effect of removing low performing systems?
RQ3 What is the effect of the experimental collections?

RQ1 stems from the observations of [50], who pointed out that

> when τ is used [...] the null hypothesis of discordance will be likely rejected when the sample size is large [i.e. the length of the ranking]. Because the samples are often large in IR, it is not surprising that the use of τ often supports concordance

and [14], who argued that

> the sample space we are actually interested in is the topic space [rather than the system space]: given that two rankings of systems are correlated over a particular set of topics, would they still be correlated, and would the correlation be as high, if run over a different set?

These remarks suggest that both the number of systems, i.e. the length of the RoS, and the number of topics, i.e. the sample space, may affect correlation. Therefore, we will investigate not only what is their individual impact on the correlation among evaluation measures but also if they have some joint effect.

RQ2 investigates the common wisdom in the IR field according to which, when conducting analyses, it is better to remove low performing systems, as done for example by [5, 55, 80].

Moreover, [68] highlighted that

> when comparing the way that test collections rank runs, if the range of scores assigned to each of the runs (being ranked) is wide, τ will tend to have a higher coefficient than if the range of scores is narrow.

It follows that, when you remove low performing systems, the range of system scores narrows down and τ goes down as well. However, apart from providing higher or lower numbers, do the different scores tell us something actually different? We will investigate this issue across joint distributions of different numbers of topics and systems.

RQ3 derives from the common practice of running experiments over different collections to seek for some sort of stable and consistent behavior. Therefore, we will investigate, across several

collections, what are the effects of the experimental collections, i.e. the effects of corpora and topic sets. *RQ3* explores the previous claim by [14] as well, since it also answers the question about what happens when you move from one topic set to another.

In order to answer these research questions, we rely on 7 different *Text REtrieval Conference (TREC)*[1] collections and, for each collection, we create a *Grid of Points (GoP)*[2] [29, 30], i.e. a set of system runs originating from all the possible combinations of the following components: 6 different stop lists, 6 types of stemmers, 7 flavors of *n*-grams, and 17 distinct IR models. These GoPs basically represent nearly all the state-of-the-art components which constitute the common denominator almost always present in any IR system for English retrieval.

We consider 8 different evaluation measures – namely, AP, P@10, Rprec, RBP, nDCG, nDCG@20, ERR, and Twist. We compute them over the set of created GoP and this originates an M_k matrix for each measure and GoP. We then properly sample these M_k matrices over topics and systems, we average them column-wise, and we compute the τ and τ_{AP} correlation coefficients between the RoS induced by each pair of evaluation measures. Finally, we use *General Linear Mixed Model (GLMM)* and *ANalysis Of VAriance (ANOVA)* [48, 58] to conduct the analyses needed to answer the above research questions.

The main contributions of the paper are:

- a robust methodology for analyzing the behaviour of correlation among evaluation measures, based on GoP and GLMM/ANOVA;
- the insights gained from answering *RQ1* to *RQ3* which show: (i) how the effect of the number of topics is more prominent than the number of systems; (ii) how removing low performing systems changes the absolute correlation scores but does not convey substantially different information when comparing a set of evaluation measures; (iii) how the experimental collections affects the correlation and, in particular, how the effect of topic sets is more prominent than the one of corpora;
- one of the most systematic studies of correlation among up-to-date evaluation measures ever conducted across many TREC collections. Indeed, as a side effect, the proposed methodology allows us to compute the expected correlation values among evaluation measures (Tables 4 and 9), net from the other effects, which prove to be stable and reliable.

The paper is organized as follows: Section 2 provides background information; Section 3 describes the experimental setup; Sections 4 to 6 answer the above research questions; finally, Section 7 draws some conclusions and discusses future work.

2 BACKGROUND

2.1 Kendall's Tau Correlation

Given two rankings X and Y, their Kendall's τ correlation [41] is given by

$$\tau(X, Y) = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}} \tag{1}$$

where P is the total number of concordant pairs (pairs that are ranked in the same order in both vectors) Q the total number of discordant pairs (pairs that are ranked in opposite order in the two vectors), T and U are the number of ties, respectively, in the first and in the second ranking.

[1] http://trec.nist.gov/
[2] http://gridofpoints.dei.unipd.it/

$\tau \in [-1, 1]$ where $\tau = 1$ indicates two perfectly concordant rankings, i.e. in the same order, $\tau = -1$ indicates two fully discordant rankings, i.e. in opposite order, and $\tau = 0$ means that 50% of the pairs are concordant and 50% discordant.

[76, 77] considers $\tau \geq 0.9$ as indication of equivalent rankings while $\tau < 0.8$ as an indication of noticeable changes in the rankings; these values have been then taken as de-facto reference thresholds by several authors, for example [15, 67, 86].

2.2 AP Correlation

AP correlation τ_{AP} [87] is a correlation coefficient inspired by Kendall's τ correlation but putting more emphasis on the order of the top ranks.

Given two m elements rankings X and Y, their AP correlation is given by

$$\tau_{AP}(Y, X) = \frac{2}{m-1} \sum_{i=2}^{m} \frac{C(i)}{i-1} - 1 \tag{2}$$

where $C(i)$ is the number of items above rank i in X and correctly ranked with respect to the item at rank i in Y, which acts as a reference. As τ, also $\tau_{AP} \in [-1, 1]$ with the same meaning.

Note that τ_{AP} is not symmetric and so, in general, $\tau_{AP}(Y, X) \neq \tau_{AP}(X, Y)$; for this reason, [87] proposed a symmetric version of τ_{AP} consisting of the average of $\tau_{AP}(Y, X)$ and $\tau_{AP}(X, Y)$. In the paper, we use the not symmetric version of τ_{AP}, where the first measure in a pair acts a reference[3].

Differently from τ, there are not (yet) de-facto reference thresholds for τ_{AP}.

2.3 Previous Work on Correlation Coefficients

Generally speaking, correlation analysis is not limited to comparing evaluation measures but it is also widely used in many other areas of IR as a means for comparing and assessing alternatives, ranging from Web page crawling [6] and ranking [7], simulation of implicit user feedback [85], resource selection in distributed IR [13], ranking queries by their difficulty [88], to measuring retrieval effectiveness itself [21, 22, 45, 84], just to name a few.

When it comes to evaluation, correlation analysys is also adopted to study inter-assessor agreement [76, 77] and quality of crowd-sourced relevance judgments [71]; it is employed in studying the effects of graded relevance judgments [40] and evaluation by highly relevant documents [24, 78]; it is utilized in investigating the impact of incomplete information in pools [8, 86], the robustness of measures to pool downsampling [23, 28, 60] and alternative pooling strategies [79].

Many authors have observed shortcomings in interpreting Kendall's τ scores: [72] have shown that correlations around 0.4–0.5 are achieved even when relevance is assigned to retrieved documents randomly; [50] noted its tendency to concordance as the length of the ranked lists increases; [68] pointed out how its values are affected by the range of the underlying scores; [14] posed the problem of its dependence on the adopted topic set.

One of the biggest limitations of Kendall's τ is its inability to weight swaps and ties: [69] proposed a generic weighting framework for Kendall's τ; [51] introduced a probabilistic instantiation of this framework and showed that τ_{AP} is also an instance of the same framework but more focussed on top ranks. [45] laid in the same framework and focussed on weights based on items relevance and similarity, while [22] applied penalty scores to weight the top ranks. [45] also proposed a generalization of the Spearman's footrule to compare rankings, extended also by [21] to ease a graphical comparison of rankings.

[3] We also conducted preliminary experiments using the symmetric version of τ_{AP} but they led to similar conclusions as those reported here for the not symmetric version.

τ_{AP} does not handle tied values: therefore, [71] suggested to uniformly sample over possible orders and to average the obtained τ_{AP} coefficients. [75] noted that breaking ties randomly can lead to some paradoxes and proposed a new weighting scheme to avoid this.

[34] proposed the τ_{GAP} coefficient which is top heavy, as τ_{AP}, and also considers the amplitude of the gap between two items in weighting a swap. [33] introduced a similar approach but applied to the case of the Pearson correlation coefficient.

[14] took a different approach to overcome the shortcomings of Kendall's τ and proposed the *rank distance* measure, which estimates the probability of observing a particular alternative ranking of systems given a baseline ranking based on measurements of system results over a sample of topics.

Finally, [84] brought in yet another angle to rank correlation by defining *Rank-Biased Overlap (RBO)*, a measure for infinite rankings based on a simple user model in which the user compares the overlap of two rankings at incrementally increasing depths and, after examining each depth, she/he has a fixed probability of stopping, modeled as a Bernoulli random variable; RBO is then calculated as the expected average overlap that the user observes in comparing the two lists.

2.4 Grid of Points

The idea of creating all the possible combinations of components has been proposed by [27], 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 about the effectiveness of the different components and their interaction; this would have called also for the identification of *suitable baselines* with respect to which all the comparisons have to be made.

More recently, the proliferation of open source IR systems [73] allowed researchers to run systematic experiments more easily. In this context, [74] conducted a vertical exploration of variations of BM25 and *Language Models (LMs)* while the "Open-Source Information Retrieval Reproducibility Challenge" [2] provided several reproducible baselines over TREC and *Conference and Labs of the Evaluation Forum (CLEF)*[4] collections. Overall, both these efforts put some points in the ideal GoP mentioned above.

[29–31] moved a step forward and created much finer-grained and systematic GoPs and paired them with a methodology based on GLMM and ANOVA in order to break down system performances into the contribution of their constituent components.

In this paper, we take yet another angle and we exploit the GoPs with a completely different purpose, i.e. investigating the correlation among evaluation measures rather than breaking down component contributions to overall system performances. The GoP are instrumental to this kind of analyses because they contain an order of magnitude more runs than even large tracks of an evaluation campaign; this allows us to extend the sample space and better explore *RQ1* and *RQ2*. Moreover, as far as *RQ3* is concerned, they give us more control because the same systems are used across different collections, reducing the variance due to the systems effect, which would be less controlled if you just use runs submitted to different tracks of an evaluation campaign. Finally, the new goal of studying correlation coefficients calls for completely different GLMM models and ANOVA analyses with respect to those of [29].

[4] http://www.clef-initiative.eu/

2.5 GLMM and ANOVA

A *General Linear Mixed Model (GLMM)* [48, 58] 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 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 which experimental condition 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). 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 this ANOVA one once you add as many X_{ij} predictors as many levels 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. In the following, we consider a confidence level $\alpha = 0.05$ to determine if a factor is statistically significant. For a detailed description on how to estimate GLMM model parameters and assess their statistical significance via ANOVA, please refer to [29, 48, 58].

The experimental design determines how you compute the model and how you estimate its parameters. In particular, it is possible to have *independent measures* designs where different subjects participate to different experimental conditions (factors) or *repeated measures* designs, 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.

2.6 Effect Size, Multiple Comparisons, and Power

We are not only interested in determining whether a factor effect is significant, but also which proportion of the variance is due to it, that is we need to estimate its *effect-size measure* or *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" [53, 64].

$$\hat{\omega}^2_{\langle fact \rangle} = \frac{df_{fact}(F_{fact} - 1)}{df_{fact}(F_{fact} - 1) + N} \tag{3}$$

is an unbiased estimator of the variance components associated with the sources of variation in the design, where 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 [58] 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*. $\hat{\omega}^2_{\langle fact \rangle}$ values could happen to be negative and in such cases they are considered as zero.

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. When performing multiple comparisons, the probability of committing a Type I error increases with the number of comparisons and we keep it controlled by applying the Tukey *Honestly Significant Difference (HSD)* test [37] with a significance level $\alpha = 0.05$. Tukey's method is used in ANOVA to create confidence intervals for all pairwise differences between factor levels, while controlling the family error rate; it is an effective method generally more powerful than other popular statistical methods like the Bonferroni one [48]. Two levels u and v of a factor are considered significantly different when

$$|t| = \frac{|\hat{\mu}_u - \hat{\mu}_v|}{\sqrt{MS_{error}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}} > \frac{1}{\sqrt{2}} q_{\alpha,k,N-k} \tag{4}$$

where $\hat{\mu}_u$ and $\hat{\mu}_v$ are the marginal means, i.e. the main effects, of the two factors; n_u and n_v are the sizes of levels u and v; $q_{\alpha,k,N-k}$ is the upper $100 * (1-\alpha)$th percentile of the studentized range distribution with parameter k and $N-k$ degrees of freedom; k is the number of levels in the factor and N is the total number of observations.

A *Type 2* 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 2 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 2 error rate.

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

$$\phi = \sqrt{N \cdot \frac{\hat{\omega}^2_{\langle fact \rangle}}{1 - \hat{\omega}^2_{\langle fact \rangle}}} \tag{5}$$

and we compare it with its tabulated values for a given Type 1 error rate α to determine β. In particular, we use the G*Power[5] software to compute the power of the conducted experiments.

3 EXPERIMENTAL SETUP

To ease reproducibility, the code for running the experiments is available at[6]: https://bitbucket.org/frrncl/tois-correlation.

[5] http://www.gpower.hhu.de/

[6] The code is based on the *MATlab Toolkit for Evaluation of information Retrieval Systems (MATTERS)* library available at: http://matters.dei.unipd.it/

3.1 Collections

We used the following standard and shared collections:

- *TREC Adhoc tracks* T07 *and* T08 [81, 82]: they 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 T07 and T08 provide 50 different topics, topic sets 351-400 and 401-450, respectively, with binary relevance judgments and a pool depth of 100 documents. 103 and 129 runs were submitted to T07 and T08, respectively;
- *TREC Web tracks* T09 *and* T10 [35, 36]: focus on a Web search task and adopt a corpus of 1.7M Web pages, i.e. the WT10g collection; both T09 and T10 are composed of 50 different topics, topic sets 451-500 and 501-550 respectively, with graded relevance judgments – i.e., not relevant, relevant and highly relevant – and a pool depth of 100 documents. 104 and 97 runs were submitted to T09 and T10, respectively;
- *TREC Terabyte tracks* T13, T14, *and* T15 [11, 18, 20]: focus on a Web search task and adopt a corpus of 125M Web pages, i.e. the GOV2 collection; T13, T14, and T15 are composed of 50 different topics, topic sets 701-750 (but just 49 are actually used), 751-800, and 801-850 respectively, with graded relevance judgments – i.e., not relevant, relevant and highly relevant – and a pool depths of 85, 100 and 50 documents respectively. 70, 58, and 80 runs were submitted to T13, T14, and T15, respectively.

3.2 Grid of Points

We consider three types of components of an IR system: stop list, *Lexical Unit Generator (LUG)* – either stemmer or *n*-gram – and IR model. We select a set of alternative implementations of each component and, by using the Terrier[7] open source system [47], we create a run for each system defined by combining the available components in all the possible ways. This produced a different GoP, i.e. a full set of runs, for each of the adopted collections – T07, T08, T09, T10, T13, T14, and T15.

The components we experiment are:

- *Stop list* (6 components): nostop, indri, lucene, snowball, smart, terrier;
- *LUG* (13 components): nolug, krovetz, lovins, porter, snowballPorter, weakPorter, 4grams, 5grams, 6grams, 7grams, 8grams, 9grams, 10grams;
- *Model* (17 components): bb2, bm25, dfiz, dfree, dirichletlm, dlh, dph, hiemstralm, ifb2, inb2, inl2, inexpb2, jskls, lemurtfidf, lgd, pl2, tfidf.

Stop lists differ in the number of composing terms: lucene has 33 terms, snowball has 174 terms, indri has 418 terms, smart has 571 terms, and terrier 733 terms.

Stemmers can be classified into aggressive and weak stemmers. lovins [46] is the most aggressive stemmer; Porter-based ones (porter, snowballPorter, and weakPorter) [54] are weaker than lovins; krovetz [43] is as aggressive as porter and weaker than lovins.

We consider seven different *n*-grams lengths ranging from $n = 4$ to $n = 10$ [49], to have a very extensive coverage of this component.

The models we employ are classified into the three main approaches currently adopted by search engines [57]: the vector space model [66] (tfidf and lemurtfidf), the probabilistic model – comprehending the bm25 model [56] and *Divergence From Randomness (DFR)* models [1] (bb2, dfiz, dfree, dlh, dph, ifb2, inb2, inl2, inexpb2, and pl2) – and *Language Models (LMs)* [89] (dirichletlm, hiemstralm, jskls, and lgd). For all the models, we considered their off-the-shelf implementation with default parameters.

[7] http://www.terrier.org/

Overall, we create GoPs consisting of $6 \times 13 \times 17 = 1,326$ system runs. They represent nearly all the state-of-the-art components which constitute the common denominator almost always present in any IR system for English retrieval and thus they are a good account of what can be found in many different operational settings.

Moreover, these GoPs are one order of magnitude bigger than the average number of runs submitted to the tracks listed above; this allows for a much deeper and systematic experimentation concerning the effects of the number of systems in *RQ1* and *RQ2*.

Finally, the GoPs computed on the different tracks are constituted by the same systems and, in the case of *RQ3*, this allows us to keep the system variance controlled when conducting analyses across tracks to study the collection effects. This would have not happened if we just used the original systems submitted to the above tracks, since they changed from year to year.

3.3 Measures

We evaluate the GoPs by employing 8 different evaluation measures: AP, P@10, Rprec, RBP, nDCG, nDCG@20, ERR[8], and Twist. Considering the overall 1,326 runs and 349 topics, this turns into more than 3.7 million data points under experimentation. In the following we provide a short description of each evaluation measure with references for further reading.

Average Precision (AP) [9] 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 [55].

Precision at Ten (P@10) [12] is the classic precision measure with cut-off at the first 10 retrieved documents.

Rprec is precision calculated with cut-off at the recall base – i.e., the total number of relevant documents for a given topic. It is an highly informative measure which shares with AP the geometric interpretation as approximation of the area under the recall-precision curve [3, 4].

Rank-Biased Precision (RBP) [52] 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 `T07` and `T08`. 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.

Normalized Discounted Cumulated Gain (nDCG) [39] 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 `T07` and `T08`, we calculate nDCG in a binary relevance setting by giving gain 0 to non-relevant documents and gain 5 to the relevant ones; whereas, for `T09`, `T10`, `T13`, `T14`, and `T15` we assign a weight 0 to non-relevant documents, 5 to the relevant ones and 10 to the highly relevant ones. Furthermore, we use a log_{10} discounting function, which accounts for a reasonably persistent user. nDCG is calculated up to the last relevant retrieved document, whereas nDCG@20 is calculated up to rank position 20.

Expected Reciprocal Rank (ERR) [16] is a measure defined for graded relevance judgments and for evaluating navigational intents. It is particularly top-heavy since it highly penalizes systems placing not-relevant documents in high positions. For ERR we used the same gains as for nDCG.

[8]Due to the strong top heaviness of ERR, ERR@20 produces more or less the same scores as ERR. Therefore, we left it out since it does not add any interesting contribution to correlation analysis.

Twist [32] 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.

Overall, IR evaluation measures embed (possibly quite) different user models and they constitute different ways of scoring systems according to the different viewpoints represented by their user model. These differences affect the correlation among the evaluation measures: for example, ERR is a much more top heavy measure than AP and this is reflected in their relatively low correlation.

3.4 Validation of the Grid of Points

Before proceeding in the experimentation, we validate the created GoPs in order to understand how representative are these GoPs of the data originally submitted to the TREC tracks under consideration.

In particular, for each evaluation measure and track, we investigate how close is the performance distribution of the original systems submitted to that TREC track to the performance distribution of the GoP systems on the same track. To quantify this "closeness" we use the *Kullback-Leibler Divergence (KLD)* [44] between the two performance distributions. In order to compute the KLD, we need the *Probability Density Function (PDF)* of the performance distributions, which we estimate by using a *Kernel Density Estimation (KDE)* [83] approach.

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

$$\hat{f}_X(x) = \frac{1}{mb} \sum_{i=1}^{m} K\left(\frac{x - X_i}{b}\right) \tag{6}$$

where b is a positive number called *bandwidth* or *window width*; $K(\cdot)$ is the *kernel* satisfying $\int_{-\infty}^{+\infty} K(x)dx = 1$. We use Gaussian kernel with a bandwidth $b = 0.015$.

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

$$D_{KL}\big(X\big\|Y\big) = \sum_x \ln\left(\frac{\hat{f}_X(x)}{\hat{f}_Y(x)}\right) \hat{f}_X(x) \tag{7}$$

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

As explained in [10], $D_{KL} \in [0, +\infty)$ denotes the information lost when Y is used to approximate X; in our context, it denotes the information lost when the GoP systems are used to "approximate" an original set of systems submitted to a TREC track. Therefore, 0 means that there is no loss of information and, in our context, that the original systems and the GoP ones are considered the same; $+\infty$ means that there is a full loss of information and, in our context, that the original systems and the GoP ones are considered completely different.

Figure 1 shows the estimated PDF plots of the GoP systems and the systems originally submitted to TREC. We show the plots in the case of AP and nDCG@20 and for the `T08`, `T10`, and `T14` tracks; the other evaluation measures and tracks exhibit a similar behaviour. We have chosen AP and nDCG@20 because they are two widely used and very well understood evaluation measures while the selected tracks represents one example for each possible corpus: `T08` for `TIPSTER`, `T10` for `WT10g`, and `T14` for `GOV2`. Table 1 reports the KLD for all the evaluation measures and all the tracks under experimentation.

Figure 1 and Table 1 show that the performance distribution of GoP systems is very close to the performance distribution of the original TREC systems. Therefore, GoP systems are representative

Table 1. KL divergence between the estimated PDF of the GoP systems and the original TREC systems for the different tracks and evaluation measures under experimentation.

Systems	AP	P@10	R-prec	RBP	nDCG	nDCG@20	ERR	Twist
T07 vs T07gop	0.0330	0.0141	0.0304	0.0248	0.0428	0.0392	0.0073	0.0229
T08 vs T08gop	0.0335	0.0178	0.0330	0.0196	0.0368	0.0127	0.0168	0.0370
T09 vs T09gop	0.0227	0.0389	0.0461	0.0553	0.0356	0.0362	0.0442	0.0329
T10 vs T10gop	0.0227	0.0123	0.0566	0.0274	0.0234	0.0156	0.0197	0.0261
T13 vs T13gop	0.0322	0.0147	0.0557	0.0140	0.1235	0.0297	0.0148	0.0687
T14 vs T14gop	0.0376	0.0085	0.0205	0.0108	0.0286	0.0170	0.0141	0.0310
T15 vs T15gop	0.0328	0.0156	0.0323	0.0239	0.0313	0.0171	0.0212	0.0438

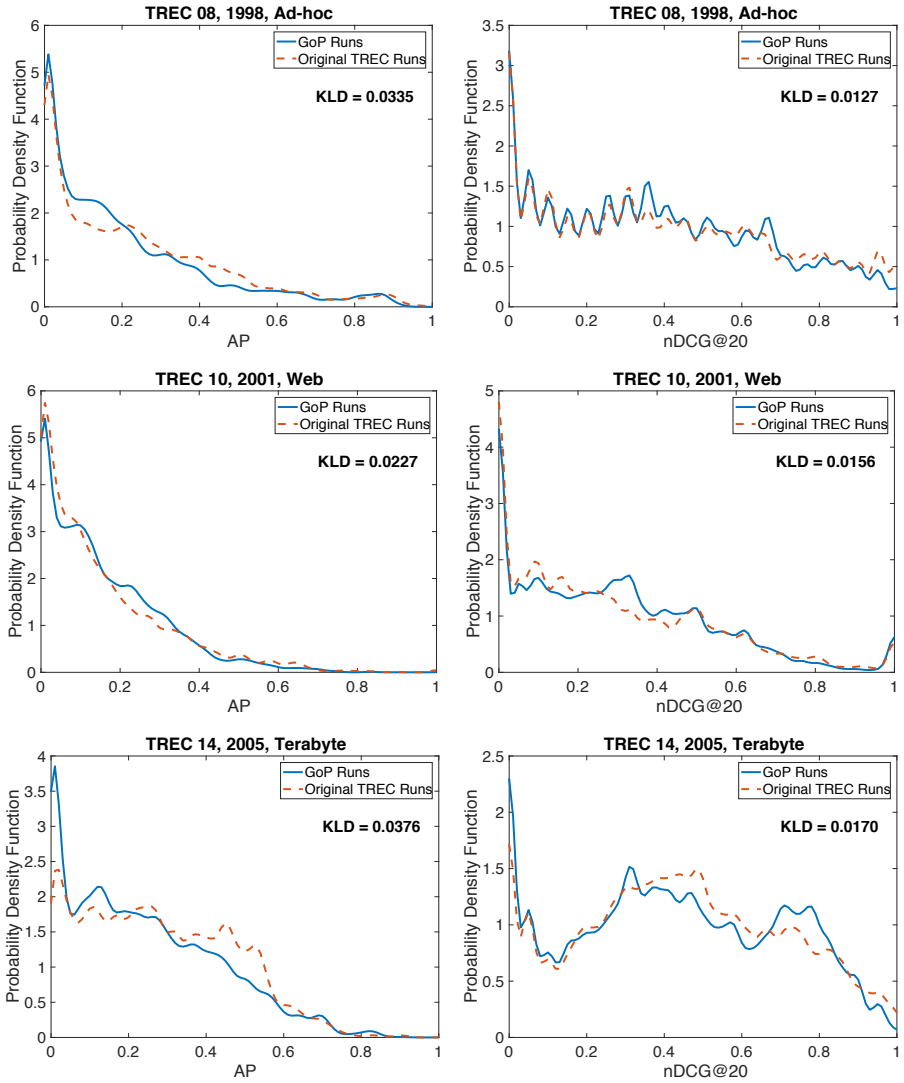

Fig. 1. Estimated PDF of the GoP systems (solid blue line) and the original TREC systems (dashed red line) and their KL divergence. On the left there is AP, on the right there is nDCG@20; the T08 (top), T10 (middle), and T14 (bottom) tracks are shown.

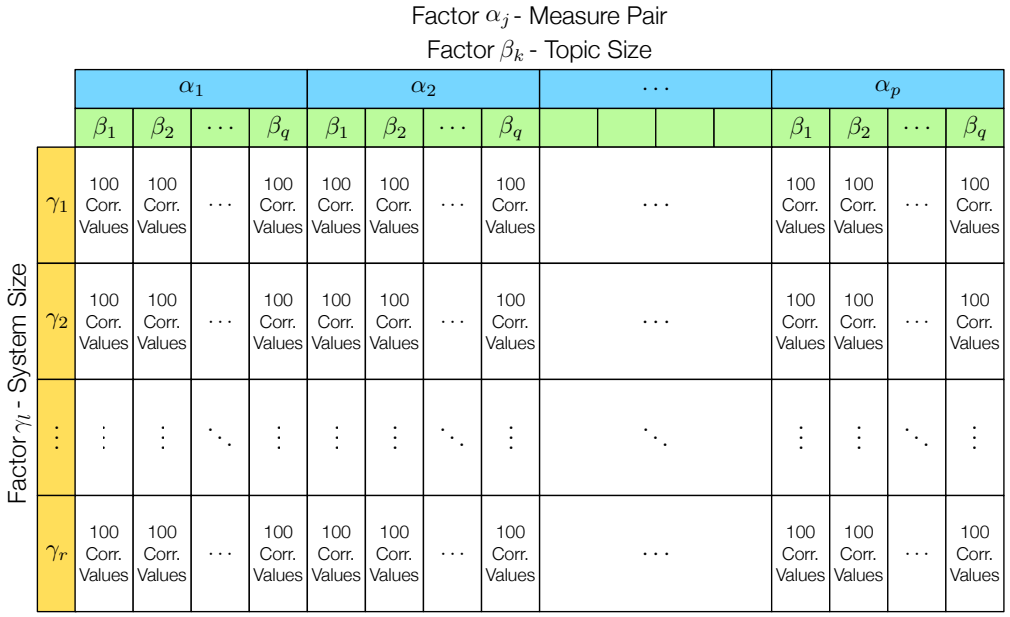

Fig. 2. Design of the GLMM model of equation (8) for the measure pair, topics size and system size effects.

of what happened in the selected TREC tracks but provide us with two key advantages. Firstly, for *RQ1* and *RQ2*, the GoPs contain at least one order of magnitude more systems than the original TREC tracks, allowing for a more fine-grained and extensive investigation. Secondly, *RQ3* is possible only using the GoP systems, since they are the same systems across all the tracks and this allows us to study the effect of corpora and topic sets.

4 EFFECT OF THE NUMBER OF SYSTEMS AND TOPICS

4.1 Methodology

We create a GoP merging the T13, T14, and T15 GoPs, called the T131415 GoP, and thus containing 149 topics and 1,326 runs. For each topic size $t \in T = \{10, 20, 30, 40, 50, 60, 70\}$ and system size $s \in S = \{10, 20, 50, 75, 100, 125, 150, 200, 250, 500\}$, we independently draw $H = 100$ random samples of t topics and $H = 100$ random samples of s systems from the T131415 GoP.

For each combination $(t, s) \in T \times S$ of topic and system sizes, we pick each sample $h = 1, \ldots, H$ of t topics and associate it with a corresponding, i.e. same index h, sample of s systems. In other terms, for each combination of topic and system sizes, we select t rows and s columns from the M_k matrices of the different evaluation measures and we repeat this operation $H = 100$ times. Therefore, we obtain M_k^h, $h = 1, \ldots, H$ matrices containing the performances of the s sampled systems over the t sampled topics according to the different evaluation measures and these matrices are then averaged column-wise $\overline{M}_k^h$.

Finally, for each pair of evaluation measures m_A and m_B and each sample $h = 1, \ldots, H$, we consider the RoS $\overline{M}_A^h$ and $\overline{M}_B^h$ produced by such pair and we compute the corresponding τ and τ_{AP} correlation coefficients.

Overall, for each combination $(t, s) \in T \times S$ of topic and system sizes and for each measure pair, this procedure originates $H = 100$ samples of correlation values for both τ and τ_{AP}.

As shown in Figure 2, this setup leads to a crossed design where subjects κ_i are the $H = 100$ samples for each combination (t, s) of topic and system sizes while factors α_j, β_k and γ_l correspond, respectively, to measure pairs, number of topics and number of systems, leading to the following GLMM model:

$$Y_{ijkl} = \underbrace{\mu.... + \kappa_i + \alpha_j + \beta_k + \gamma_l}_{\text{Main Effects}} + \underbrace{(\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}} \tag{8}$$

where:

- Y_{ijkl} is the correlation value, either τ or τ_{AP}, of the i-th subject in the j-th, k-th, and l-th factors;
- $\mu....$ is the grand mean;
- κ_i is the effect of the i-th subject, i.e. the $h = 1, \ldots, H$ samples for each (t, s) combination, where $\kappa_i = \mu_{i...} - \mu....$ and $\mu_{i...}$ is the mean of the i-th subject;
- $\alpha_j = \mu_{.j..} - \mu....$ is the effect of the j-th factor, i.e. measure pairs, where $\mu_{.j..}$ is the mean of the j-th factor. Considering that we are experimenting with 8 evaluation measures, there are $\binom{8}{2} = 28$ measure pairs, i.e. 28 levels for factor α_j;
- $\beta_k = \mu_{..k.} - \mu....$ is the effect of the k-th factor, i.e. number of topics, where $\mu_{..k.}$ is the mean of the k-th factor; there are $|T| = 7$ levels for factor β_k;
- $\gamma_l = \mu_{...l} - \mu....$ is the effect of the l-th factor, i.e. number of systems, where $\mu_{...l}$ is the mean of the l-th factor; there are $|S| = 10$ levels for factor γ_l;
- $(\alpha\beta)_{jk}$, $(\alpha\gamma)_{jl}$, and $(\beta\gamma)_{kl}$ are, respectively, the interactions between measures pairs and number of topics, measure pairs and number of systems, and number of topics and number of systems;
- 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.

Considering that there are 28 measure pairs, 7 topic sizes, 10 system sizes and that, for each combination of these factors, we use 100 subjects, overall this amounts to analyzing 196,000 correlation values for both τ and τ_{AP}.

4.2 Experimental Results

4.2.1 General Trends. Figure 3 shows the average correlation of the `AP vs nDCG@20` pair over the $H = 100$ samples for each (t, s) combination and the corresponding confidence interval; Kendall's τ correlation is drawn with a solid blue line while AP correlation τ_{AP} is drawn with a dashed green line. The other evaluation measure pairs exhibit a consistent behaviour with respect to the one of the `AP vs nDCG@20` pair, which we use here as an example to discuss the main trends because AP and nDCG@20 are two very widely used and well-understood measures. If in the table of Figure 2 we fix the level α_j corresponding to the `AP vs nDCG@20` pair, we can note that Figure 3 just plots the raw data contained in the cells under that level α_j, using a sub-plot for each topic size β_k level and, within each sub-plot, plotting the different system size γ_l levels on the x-axis.

As Figure 3 highlights, the number of topics affects both τ and τ_{AP}, since their average value increases as the number of topics increases. On the other hand, the number of systems exhibits less impact on the two correlation coefficients: indeed, apart from a small transient up to around 75-100 systems, the trend for both coefficients is somehow constant, especially when the number of topics increases. We can note how, in the transient phase, τ and τ_{AP} behave differently: τ tends to slightly increase before reaching stability while τ_{AP} manifest an initial decrease, sometimes followed by an increase, before getting more or less constant.

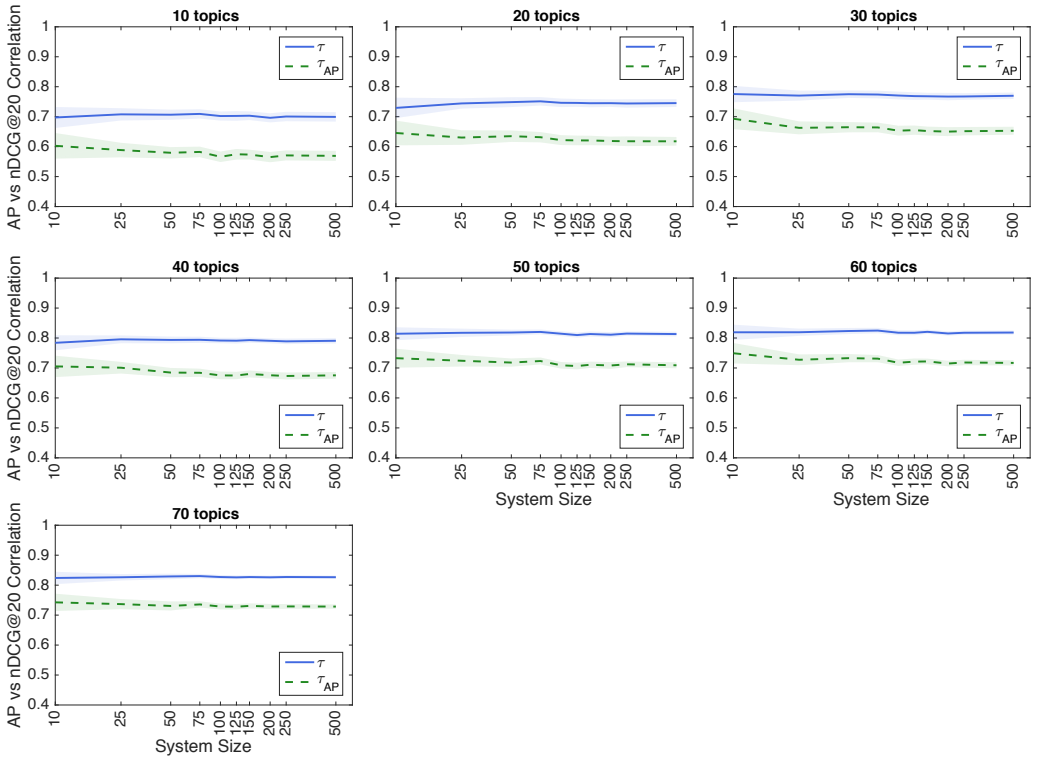

Fig. 3. AP vs nDCG@20 correlation on the T131415 GoP, averaged over $H = 100$ samples for each (t, s) combination, and confidence interval (shaded). Kendall's τ correlation is drawn with a solid blue line; AP correlation τ_{AP} is drawn with a dashed green line. Each plot shows the correlation for a given number of topics as the number of systems increases.

Figure 3 does not support the claim by [50], at least in the case of the correlation among evaluation measures, since the τ coefficient does not steadily increases as the sample size, i.e. the number of systems, increases. This represents a positive feature of the τ coefficient when used to study the correlation among evaluation measures, because it frees us from the concern of how many systems we are using and if this number would lead us to observe somehow biased correlation values.

When it comes to confidence intervals, lower number of topics and systems call for larger intervals, which is not surprising. However, τ generally exhibits smaller confidence intervals than τ_{AP}, especially for low number of topics. Moreover, τ seems to be a bit more effective than τ_{AP} in benefiting from the increasing number of topics and systems; indeed, correlation values get more stable and confidence intervals get smaller in a "faster" way for τ than for τ_{AP}.

4.2.2 GLMM and ANOVA Analysis. Tables 2 and 3 report the results of the ANOVA analyses on the GLMM model of equation (8) for τ and τ_{AP}, respectively.

The tables show that the main effects of the measure pair (α_j), topic size (β_k), and system size (γ_l) factors are all statistically significant for both τ and τ_{AP}.

Table 2. Kendall's τ correlation: ANOVA table for the GLMM model of equation (8), considering the measure pair, topic size, and system size effects on the T131415 GoP.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	Power
Subject	55.8432	99	0.5641	99.6711	0.0000		
Measure Pair	1,666.5150	27	61.7228	10,906.3664	0.0000	0.6004	1.0000
Topic Size	418.7689	6	69.7948	12,332.6910	0.0000	0.2740	1.0000
System Size	2.3875	9	0.2653	46.8753	3.64e-85	0.0021	0.9999
Measure Pair*Topic Size	33.1886	162	0.2049	36.2001	0.0000	0.0283	1.0000
Measure Pair*System Size	0.9346	243	0.0038	0.6796	1.0000	0.0000	0.8043
Topic Size*System Size	0.6136	54	0.0114	2.0079	1.69e-05	0.0002	0.5137
Error	1,105.8283	195,399	0.0057				
Total	3,284.0798	195,999					

Table 3. AP correlation τ_{AP}: ANOVA table for the GLMM model of equation (8), considering the measure pair, topic size, and system size effects on the T131415 GoP.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	Power
Subject	71.0670	99	0.7178	86.7603	0.0000		
Measure Pair	2,536.3318	27	93.9382	11,353.5200	0.0000	0.6100	1.0000
Topic Size	612.0528	6	102.0088	12,328.9432	0.0000	0.2740	1.0000
System Size	12.0979	9	1.3442	162.4638	4.20e-308	0.0074	1.0000
Measure Pair*Topic Size	26.9371	162	0.1662	20.0967	0.0000	0.0155	1.0000
Measure Pair*System Size	1.2495	243	0.0051	0.6214	1.0000	0.0000	0.7467
Topic Size*System Size	0.8735	54	0.0162	1.9550	3.44e-05	0.0002	0.4966
Error	1,616.7174	195,399	0,0083				
Total	4,877.3271	195,999					

The most prominent effect is the measure pair one, which is a large size effect in terms of $\hat{\omega}^2$, and it has almost the same size for both τ and τ_{AP}. The large portion of variance explained by the measure pair factor makes sense since correlation varies quite a lot from one measure pair to another one and it is what actually discriminates evaluation measures.

The second biggest effect is the topic size one, which again is a large size effect and it has the same size for both τ and τ_{AP}. This supports the previous observations about Figure 3 when we noted that the topic size is the most prominent factor influencing the correlation among evaluation measures. However, its size is slightly less half than the size of the measure pair effect, indicating that the correlation among evaluation measures is by far the dominating factor.

Finally, the system size effect, even if significant, is a very small size effect and we can consider it almost negligible; however, it should be noted that this effect is a little bit more than three times bigger for τ_{AP} than for τ. Overall, this sustains the observations made above about the smaller importance of the number of systems on the correlation among evaluation measures, with τ_{AP} being more sensitive to this factor than τ.

When it comes to the interaction between effects, for both τ and τ_{AP}, the measure pair and topic size $(\alpha\beta)_{jk}$ and the topic size and system size $(\beta\gamma)_{kl}$ interactions are statistically significant. On the other hand, the measure pair and system size $(\alpha\gamma)_{jl}$ interaction is not significant and this further stress the fact that the number of systems does not influence much the correlation among evaluation measures.

The interaction between measure pair and topic size $(\alpha\beta)_{jk}$ is a small size effect, about 1.8 times bigger for τ than for τ_{AP}, indicating that the former is slightly more influenced by it. This not only strengthens the importance of the number of topics on the correlation among evaluation measures but it also suggests that different evaluation measure pairs may interact differently with the number of topics, i.e. the correlation between certain measure pairs may be increased or decreased more than others by some number of topics. The topic size and system size interaction $(\beta\gamma)_{kl}$ is another

Table 4. Main effects of the τ and τ_{AP} correlation coefficients for each evaluation measure pair, net from the number of topics and systems effects on the T131415 GoP. These are the values plotted in Figure 4 on the left. Note that τ_{AP} is not symmetric and we used as reference the evaluation measure reported in the row.

		AP	P@10	R-prec	RBP	nDCG	nDCG@20	ERR	Twist
AP	τ	1.0000	0.7273	0.9087	0.7307	0.8923	0.7814	0.6216	0.8793
	τ_{AP}	1.0000	0.6064	0.8477	0.6160	0.8484	0.6741	0.4959	0.8125
P@10	τ	–	1.0000	0.7165	0.8720	0.6909	0.8221	0.6860	0.7126
	τ_{AP}	–	1.0000	0.5919	0.7787	0.5665	0.7098	0.5481	0.5864
R-prec	τ	–	–	1.0000	0.7197	0.8734	0.7736	0.6197	0.8690
	τ_{AP}	–	–	1.0000	0.6038	0.8002	0.6641	0.4912	0.7849
RBP	τ	–	–	–	1.0000	0.6894	0.8141	0.7550	0.7092
	τ_{AP}	–	–	–	1.0000	0.5710	0.7051	0.6322	0.5883
nDCG	τ	–	–	–	–	1.0000	0.7376	0.5892	0.8841
	τ_{AP}	–	–	–	–	1.0000	0.6276	0.4683	0.8125
nDCG@20	τ	–	–	–	–	–	1.0000	0.6754	0.7644
	τ_{AP}	–	–	–	–	–	1.0000	0.5381	0.6478
ERR	τ	–	–	–	–	–	–	1.0000	0.6057
	τ_{AP}	–	–	–	–	–	–	1.0000	0.4726
Twist	τ	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	1.0000

extremely small size effect that we can neglect, again indicating that the number of systems/topics tend to not influence each other.

As a final remark emerging from from Tables 2 and 3, we can note how consistent is the behavior of τ and τ_{AP}, which basically exhibit extremely close effect sizes for almost all the factors. Moreover, the statistical power is extremely high for all the significant effects, with the exception of the topic size and system size interaction which is slightly under-powered.

4.2.3 Main Effects Analysis. Figure 4 shows the main effects plot of the measure pair (on the left), topics size (on the middle) and system size (on the right) factors for both τ (top in blue) and τ_{AP} (bottom in green). The main effects plot graphs the response mean for each factor level connected by a dotted tiny line and, by means of this plot, we can easily determine the impact of the different levels of a factor. In the case of the topic size and system size factor, we also report the outcomes of the Tukey HSD test for the $t = 50$ topic size and the $s = 100$ system size, respectively: topic sizes and system size not significantly different from $t = 50$ and $s = 100$, i.e. those which are in the same group, are highlighted with a diamond.

Figure 4 on the left shows the expected values of τ and τ_{AP}, net from the effects of the number of topics and systems; these values are also reported in Table 4, which can be held as reference correlation values among these evaluation measures, distilled across a wide range of number of topics and systems. As anticipated in Section 2.2, τ_{AP} is not symmetric and we used as reference the evaluation measure reported in the row of Table 4 which corresponds to the first one in the labels of measure pair plot in Figure 4.

The effect of the number of topics (middle plot of Figure 4) is to increase the correlation score for both the correlation coefficients and the Tukey HSD plots show that the topic sizes tend to be significantly different from each other, apart from $t = 50$ and $= 60$ which belong to the same group. However, the actual increment due to the topic size slows down as the topic size increases. Indeed, the marginal means with respect to the topic size factor are: at 10 topics $\bar{\tau}^{t=10} = 0.6669$ and $\bar{\tau}_{AP}^{t=10} = 0.5441$; at 30 topics $\bar{\tau}^{t=30} = 0.7413$ and $\bar{\tau}_{AP}^{t=30} = 0.6268$ with a difference of 11.16% and 15.20%, respectively, with respect to 10 topics; at 50 topics $\bar{\tau}^{t=50} = 0.7908$ and $\bar{\tau}_{AP}^{t=50} = 0.6907$ with a difference of 6.68% and 10.19%, respectively, with respect to 30 topics; at 70 topics $\bar{\tau}^{t=70} = 0.8075$ and $\bar{\tau}_{AP}^{t=70} = 0.7132$ with a difference of 2.11% and 3.26%, respectively, with respect to 50 topics.

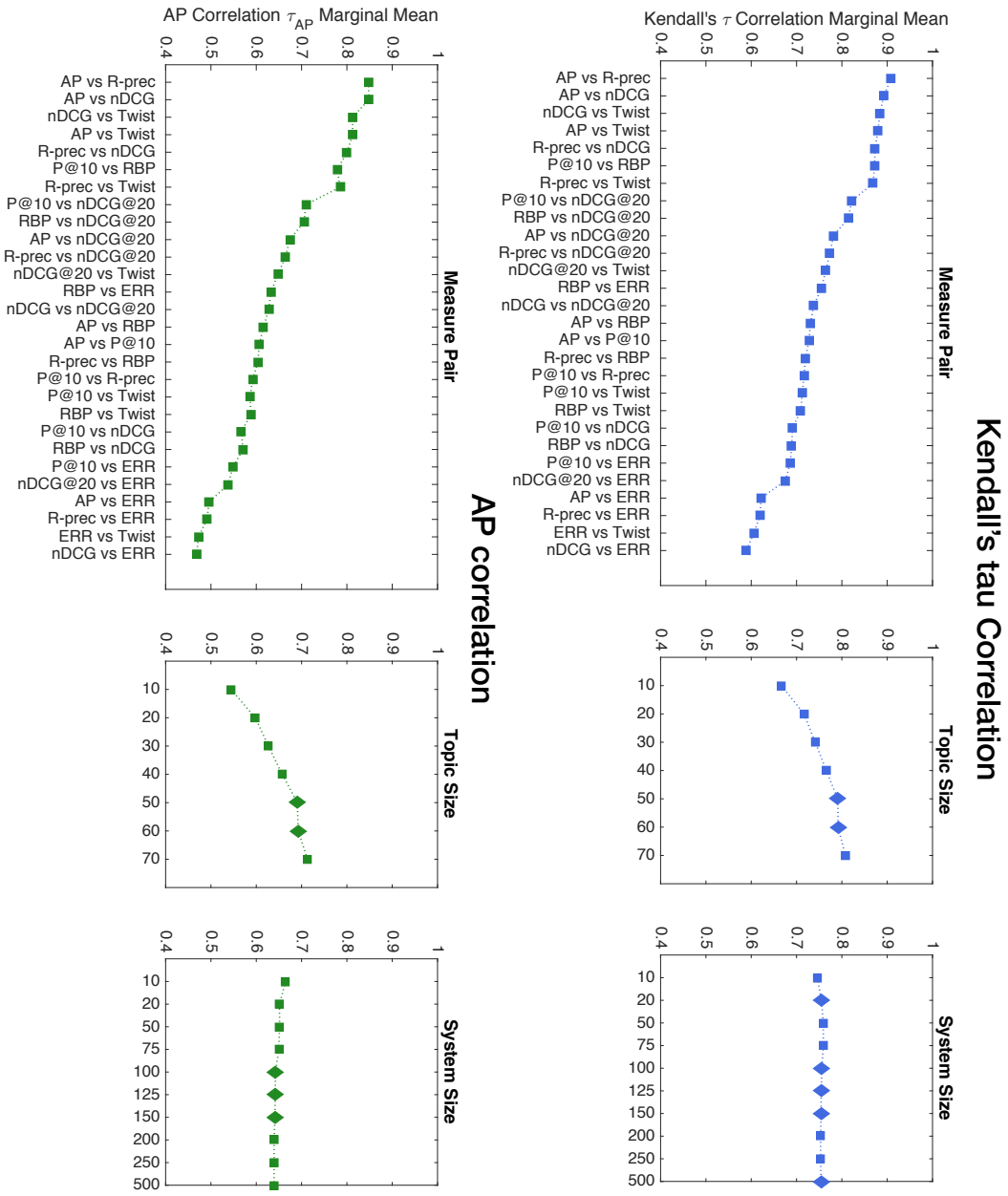

Fig. 4. Main effects for the measure pair (on the left), topics size (on the middle) and system size (on the right) factors on the T131415 GoP. Kendall's τ correlation is at the top in blue, AP correlation τ_{AP} is at the bottom in green. The topic size and system size plots report also the Tukey HSD comparison for the $t = 50$ topic size and the $s = 100$ system size, respectively. Topic sizes and system size not significantly different from $t = 50$ and $s = 100$ are highlighted with a diamond. Note that τ_{AP} is not symmetric and we used as reference the first evaluation measure in the labels of the measure pair plot. Also note that the figure is rotated and indications like left, middle and right all refer to when you rotate the figure to read it.

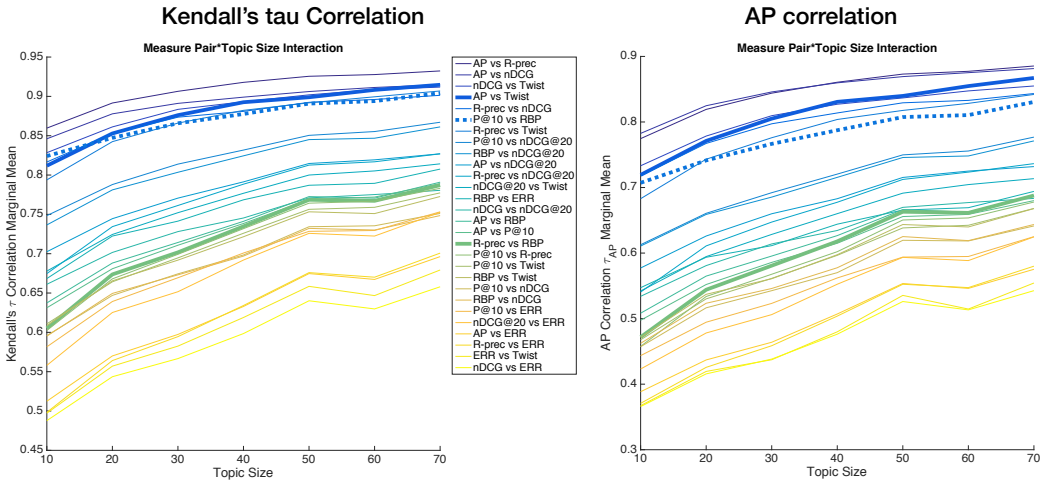

Fig. 5. Interaction effects for the Measure Pair*Topic Size factors on the T131415 GoP. Kendall's τ correlation is on the left, AP correlation τ_{AP} is on the right.

Overall, this suggests that the 50 topics typically found in a track of an evaluation campaign are at the beginning of the range where the effect of the number of topics stabilizes and thus they represent a reasonable choice, trading off the cost of topic and ground-truth creation. Moreover, these findings answer the question of [14] from a different angle, since they clearly show that the topic size matters a lot when studying the correlation among evaluation measures.

When it comes to the number of systems (right plot of Figure 4), we can see how τ is slightly less affected by them than τ_{AP} and that, after an initial small transient up to 75-100 systems where τ marginally increases and τ_{AP} marginally decreases, both τ and τ_{AP} becomes basically constant. This is also confirmed if you look at the marginal means with respect to the system size factor: at 10 systems $\bar{\tau}^{s=10} = 0.7458$ and $\bar{\tau}_{AP}^{s=10} = 0.6649$; at 50 systems $\bar{\tau}^{s=50} = 0.7588$ and $\bar{\tau}_{AP}^{s=50} = 0.6515$ with a difference of just 1.74% and 2.02%, respectively, with respect to 10 systems; at 100 systems $\bar{\tau}^{s=100} = 0.7558$ and $\bar{\tau}_{AP}^{s=100} = 0.6425$ with a difference of 0.40% and 1.38%, respectively, with respect to 50 topics; at 150 systems $\bar{\tau}^{s=150} = 0.7541$ and $\bar{\tau}_{AP}^{s=150} = 0.6410$ with a very small difference of only 0.22% and 0.23%, respectively, with respect to 100 systems; and similarly for even greater numbers of systems. So, even if the Tukey HSD test shows a few significant differences, we can see how these differences have a limited impact in practical terms, as also supported by the ANOVA Tables 2 and 3 which show how the system size is a significant but extremely small effect.

As previously noted, these results do not support the claim by [50], who suggested that τ increases with the length of the sample size, i.e. the number of systems in our case. However, the example of [50] consists of a list where the top 10 elements are in opposite order and kept fixed while all the others are concordant as the list gets longer: in this case, as the list length increases, τ increases. In our case, instead, we have swaps in all the positions of the RoS of the longer and longer lists and this motivates why we observe a different behavior than [50].

4.2.4 Interaction Effects Analysis. Figure 5 shows the interaction plots for the Measure Pair*Topic Size factor which, according to Tables 2 and 3 is the only significant interaction effect with a not-negligible effect size; τ is plotted on the left and τ_{AP} is plotted on the right. An interaction effects plot displays the levels of one factor on the X axis and has a separate line for the means of each

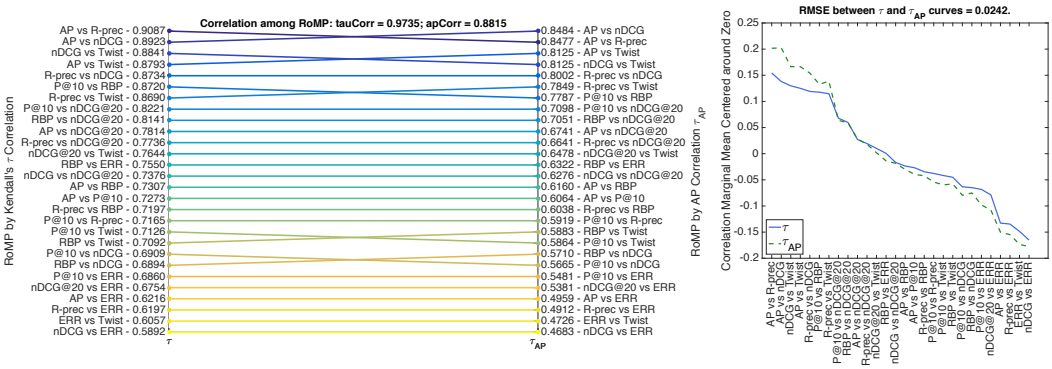

Fig. 6. Comparison of the main effects of measure pair factor for both τ and τ_{AP} on the T131415 GoP. On the left, RoMP are ordered by τ and τ_{AP}: horizontal and oblique lines indicate, respectively, concordance and discordance between the rankings. On the right, the same main effects shown in Figure 4 on the left but with means centered around zero, where τ is drawn with a solid blue line while τ_{AP} with a dashed green line.

level of the other factor on the Y axis; it allows us to understand whether the effect of one factor depends on the level of the other factor. 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.

We can see that in Figure 5, even if lines exhibit a common upward trend as the topic size increases, there are many non parallel lines and many crossings, denoting a good interaction between the two factors. For example, in the case of τ, `AP vs Twist` (solid thick blue line) is below `P@10 vs RBP` (dotted thick blue line) at 10 topics, they cross around 20 topics, and `P@10 vs RBP` gets over `AP vs Twist` from 30 topics onwards; we can note how the `AP vs Twist` slope is a bit steeper than the `P@10 vs RBP` one, indicating that `AP vs Twist` benefits more than `P@10 vs RBP` from the progressively increasing topic size; finally, in the case of τ_{AP} we can observe that `AP vs Twist` is always above `P@10 vs RBP` and, as in the case of τ, its slope is steeper than the `P@10 vs RBP` one, again denoting interaction between the two factors. As a further example, for both τ and τ_{AP}, `R-prec vs RBP` (solid thick green line) starts below many other measure pairs but it has a quite steep slope and, as the topic size increases, it crosses several of the measure pairs above it.

Overall, this confirms that the number of topics has a great impact on the correlation among evaluation measures and that, at different topic sizes, we may observe different behaviors for different measure pairs.

4.2.5 τ and τ_{AP} Comparison. It can be noted how, in Figure 4, the τ and τ_{AP} curves for the measure pair factor look very similar, provided that τ_{AP} has lower values than τ and it is translated towards the bottom. This suggest that, yet providing different correlation values, τ and τ_{AP} may convey somewhat similar information about the correlation among a set of evaluation measures.

To explore a bit more this aspect, we rank measure pairs by their τ and τ_{AP} scores reported in Table 4 and shown in Figure 4 on the left; this originates two *Rankings of Measure Pairs (RoMP)* and we investigate how close these two RoMP are, considering that the closer they are the more similar information they provide about the correlation among evaluation measures.

Figure 6 on the left shows a parallel coordinates plot [38] of the τ and τ_{AP} RoMP. The parallel coordinates plot is a visualization technique used to plot individual data elements across many dimensions; each of the dimensions corresponds to a vertical axis – in our case we have two axes,

one for τ and another for τ_{AP} – and each data element is displayed as a series of connected points along the dimensions/axes.

It emerges that there are many concordant pairs, that there are just 5 swaps between the two RoMP, and that these swaps are only local, i.e. they concern just two adjacent positions, as in the case of the `AP vs R-Prec` and `AP vs nDCG` pairs in the top ranks. To get a quantitative appreciation of how close these two RoMP are, we can adopt the Kendall's tau and AP correlation coefficients themselves but playing a different role: here we use them as an analysis tool and they are not the object of investigation, as it generally happens in this paper; to clearly indicate these different roles, we label them tauCorr and apCorr when they are used as an analysis tool while we label them τ and τ_{AP} when they are the object of investigation. We obtain that the correlation among the two RoMP is tauCorr = 0.9735 and apCorr = 0.8815, suggesting that τ and τ_{AP} rank evaluation measure pairs in a very similar manner. We can also note how the 3 swaps in the top ranks of the RoMP are penalized by apCorr, which is about 10% lower than tauCorr.

We look at how similar information τ and τ_{AP} provide about the correlation among evaluation measures also from another angle. We have observed that the τ and τ_{AP} curves for the measure pair factor, shown in Figure 4 on the left and reported in Table 4, look very similar even if they are somewhat translated. Therefore, to better appreciate this similarity, we center the mean of the τ and τ_{AP} curves around zero, i.e. we remove from each curve its mean across the different measure pairs, as shown in Figure 6. Finally, we quantify how close these two curves are by using the *Root Mean Square Error (RMSE)* [42], which is just RMSE = 0.0242, indicating very small differences.

Overall, these findings suggest that, if you consider a set of evaluation measures and you compare them across a large set of topic and system sizes, removing those effects, τ and τ_{AP} have different absolute values but they provide a quite consistent assessment of what the differences among these evaluation measures are.

5 EFFECT OF REMOVING LOW PERFORMING SYSTEMS

5.1 Methodology

When conducting experimentation, low performing systems are identified in various ways, according to what fits best with the goal of the experiment at hand. For example, [5, 55] consider only runs that retrieve at least 5 relevant and 5 highly relevant documents while [80] remove runs in the first quartile of MAP. Here, we take the same angle as [80] and we consider as low performing systems those falling in the first quartile of MAP, since this is a very commonly adopted approach.

Our goal is to understand whether the following two cases provide substantially different information about the correlation among evaluation measures: `allQ` case, where we use all the systems to compute the correlation scores; `no1Q` case, where we remove low performing systems, i.e. those falling in the first quartile of MAP, before computing the correlation scores.

We use the same dataset and experimental setup described in Section 4.1 and briefly recapped here: (i) generate $H = 100$ samples of t topics and s systems; (ii) for each sample, compute the performance of the systems over the topics with respect to all measures; (iii) for each sample, rank the systems by the performance measures, one RoS per measure; (iv) for each sample and for each pair of measures, compute the correlation of the RoS with both τ and τ_{AP}.

Therefore, for each combination $(t, s) \in T \times S$ of topic and system sizes, we consider:

(1) `allQ`: τ and τ_{AP} correlations scores, averaged over the $H = 100$ samples, where at step (iii) we use the whole RoS;
(2) `no1Q`: τ and τ_{AP} correlations scores, averaged over the $H = 100$ samples, where at step (iii) we remove first quartile systems from the RoS; we identify systems in the first quartile in terms of their MAP.

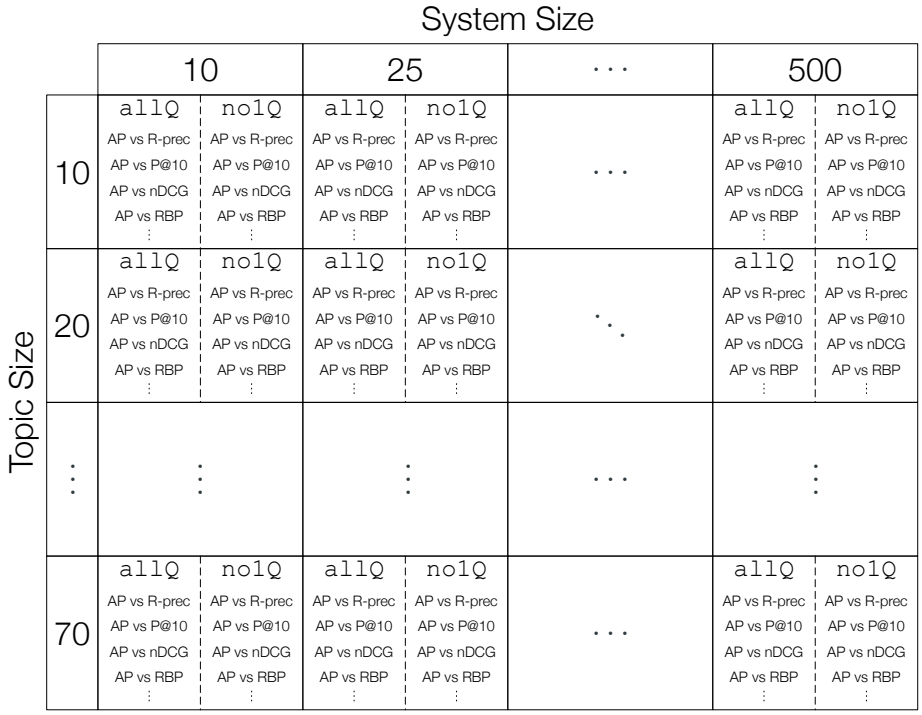

Fig. 7. Experimental design for investigating *RQ2*.

Adopting the above procedure, as shown in Figure 7, we end up with 28 distinct τ and τ_{AP} average correlation scores, where 28 is the number of possible measure pairs, for the allQ and no1Q cases at each combination $(t, s) \in T \times S$ of topic and system sizes.

Now, for each cell of the table in Figure 7, i.e. for each combination $(t, s) \in T \times S$ of topic and system sizes, we need to understand whether the allQ and no1Q cases convey somewhat similar information about the correlation among evaluation measures. To this end, we reason in a way similar to what we did in Section 4.2.5 and we analyze the data as follows:

- we rank measure pairs by their allQ and no1Q correlation scores and we investigate how close these two allQ and no1Q RoMP are, considering that the closer they are the more similar information they provide about the correlation among evaluation measures. To get a quantitative appreciation of how close these two RoMP are, we use the Kendall's tau and AP correlation coefficients themselves but playing the role of an analysis tool; therefore, we label them tauCorr and apCorr in this case.
- we consider the allQ and no1Q curves, i.e. the plot of the values contained in a cell of the table in Figure 7, and we center the mean of the allQ and no1Q curves around zero, i.e. we remove from each curve its mean across the different measure pairs. This allows us to assess how close are the allQ and no1Q curves, once the vertical translation due to different absolute values has been removed. We quantify this "closeness" in terms of RMSE between the allQ and no1Q curves: the lower the RMSE scores, the more similar are the allQ and no1Q curves and the smaller the difference between removing or not removing low performing systems.

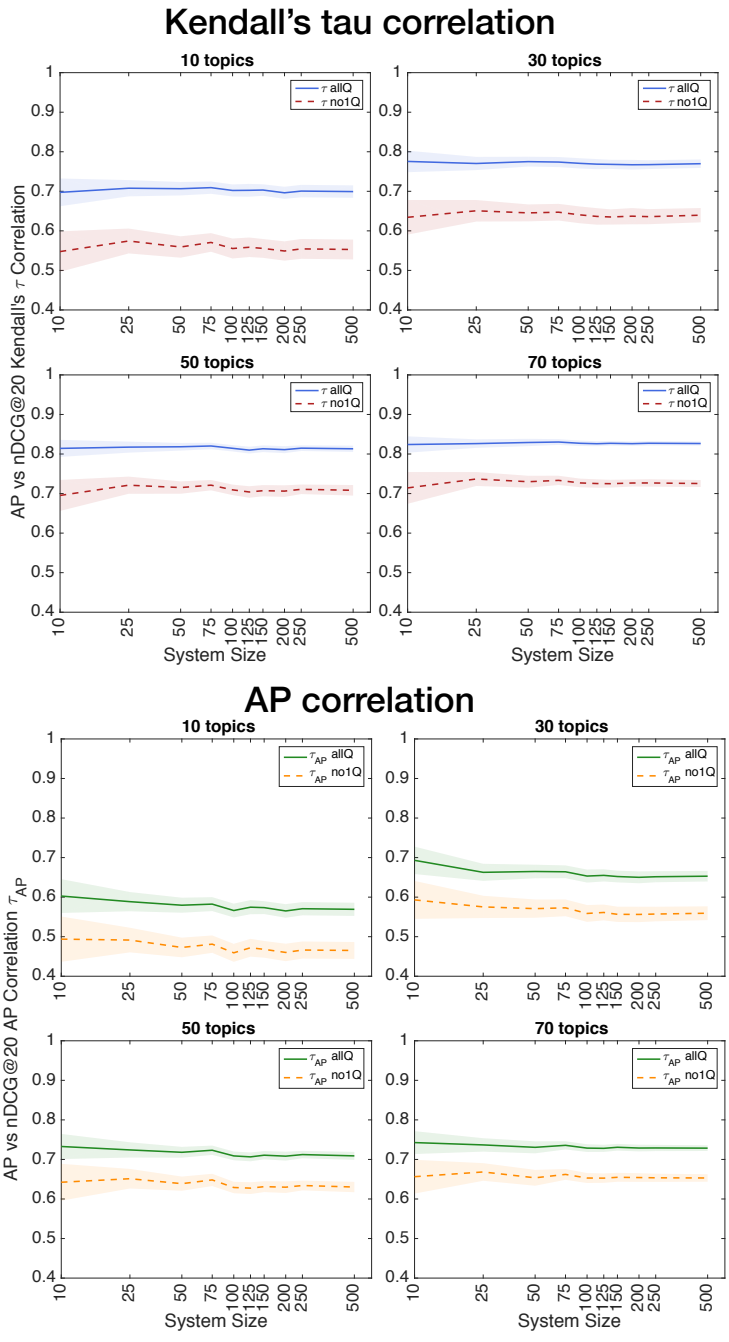

Fig. 8. AP vs nDCG@20 pair correlation on the T131415 GoP, averaged over $H = 100$ samples at the $t = 10, 30, 50, 70$ topic sizes, and confidence interval (shaded). Kendall's τ correlation is on the top: allQ is a solid blue line and no1Q is a dashed red line; AP correlation τ_{AP} is on the bottom: allQ is a solid green line and no1Q is a dashed orange line. Each plot shows the correlation for a given number of topics as the number of systems increases. The allQ curves are the same as those shown in the corresponding plots of Figure 3.

Table 5. Comparison of Kendall's τ correlation with and without removing first quartile systems on the T131415 GoP. For each (t, s) combination, the tauCorr and apCorr values measure how close the RoMP are with and without removing first quartile systems; the RMSE values quantify how close are the allQ and no1Q scores, once their mean over the set of evaluation measure pairs has been centered around zero.

			System Size									
			10	**25**	**50**	**75**	**100**	**125**	**150**	**200**	**250**	**500**
Topic Size	**10**	tauCorr	0.9312	0.9524	0.9577	0.9577	0.9577	0.9524	0.9524	0.9418	0.9418	0.9418
		apCorr	0.9075	0.9589	0.8937	0.9647	0.9617	0.9411	0.9415	0.9298	0.9233	0.9338
		RMSE	0.0452	0.0400	0.0419	0.0379	0.0416	0.0399	0.0419	0.0400	0.0414	0.0413
	20	tauCorr	0.9365	0.9788	0.9894	0.9894	0.9894	0.9841	0.9788	0.9841	0.9841	0.9894
		apCorr	0.9291	0.8966	0.9871	0.9899	0.9871	0.9830	0.9825	0.9832	0.9832	0.9871
		RMSE	0.0521	0.0406	0.0440	0.0463	0.0452	0.0471	0.0471	0.0465	0.0462	0.0464
	30	tauCorr	0.9524	0.9735	0.9947	0.9735	0.9894	0.9841	1.0000	0.9947	0.9841	0.9894
		apCorr	0.9264	0.9077	0.9815	0.9561	0.9903	0.9719	1.0000	0.9938	0.9714	0.9780
		RMSE	0.0566	0.0499	0.0497	0.0474	0.0493	0.0502	0.0513	0.0498	0.0494	0.0494
	40	tauCorr	0.9365	0.9841	0.9894	0.9788	0.9788	0.9841	0.9947	0.9841	0.9947	0.9841
		apCorr	0.9417	0.9744	0.9817	0.9830	0.9538	0.9785	0.9965	0.9571	0.9852	0.9785
		RMSE	0.0480	0.0428	0.0462	0.0454	0.0466	0.0467	0.0479	0.0457	0.0474	0.0473
	50	tauCorr	0.9683	0.9894	0.9788	0.9683	0.9947	0.9735	0.9841	0.9683	0.9735	0.9683
		apCorr	0.9306	0.9927	0.9788	0.9472	0.9954	0.9640	0.9858	0.9435	0.9643	0.9520
		RMSE	0.0440	0.0370	0.0385	0.0403	0.0423	0.0415	0.0415	0.0408	0.0415	0.0417
	60	tauCorr	0.9894	0.9841	0.9735	0.9788	0.9841	0.9894	0.9894	0.9735	0.9841	0.9894
		apCorr	0.9925	0.9671	0.9326	0.9628	0.9769	0.9806	0.9808	0.9632	0.9725	0.9802
		RMSE	0.0486	0.0416	0.0429	0.0453	0.0466	0.0456	0.0470	0.0462	0.0473	0.0468
	70	tauCorr	0.9577	0.9630	0.9524	0.9418	0.9418	0.9471	0.9683	0.9418	0.9577	0.9524
		apCorr	0.9547	0.9335	0.9337	0.9054	0.9027	0.9394	0.9281	0.9029	0.9142	0.9103
		RMSE	0.0417	0.0383	0.0421	0.0409	0.0421	0.0420	0.0433	0.0413	0.0428	0.0428

Table 6. Comparison of AP correlation τ_{AP} with and without removing first quartile systems on the T131415 GoP. For each (t, s) combination, the tauCorr and apCorr values measure how close the RoMP are with and without removing first quartile systems; the RMSE values quantify how close are the allQ and no1Q scores, once their mean over the set of evaluation measure pairs has been centered around zero.

			System Size									
			10	**25**	**50**	**75**	**100**	**125**	**150**	**200**	**250**	**500**
Topic Size	**10**	tauCorr	0.9471	0.9788	0.9735	0.9788	0.9788	0.9841	0.9735	0.9735	0.9683	0.9788
		apCorr	0.9645	0.9849	0.9768	0.9830	0.9709	0.9897	0.9718	0.9690	0.9775	0.9869
		RMSE	0.0306	0.0235	0.0243	0.0219	0.0231	0.0224	0.0231	0.0224	0.0230	0.0228
	20	tauCorr	0.9683	0.9788	0.9947	0.9947	0.9894	0.9894	0.9894	0.9947	0.9947	0.9947
		apCorr	0.9055	0.9681	0.9938	0.9943	0.9905	0.9911	0.9911	0.9938	0.9973	0.9973
		RMSE	0.0325	0.0252	0.0254	0.0269	0.0262	0.0268	0.0269	0.0264	0.0263	0.0263
	30	tauCorr	0.9577	1.0000	0.9894	0.9788	0.9788	0.9894	0.9841	0.9788	0.9894	0.9947
		apCorr	0.8900	1.0000	0.9916	0.9673	0.9051	0.9916	0.9741	0.9129	0.9213	0.9259
		RMSE	0.0360	0.0308	0.0297	0.0286	0.0287	0.0299	0.0299	0.0291	0.0292	0.0290
	40	tauCorr	0.9630	1.0000	0.9894	0.9788	0.9947	0.9894	0.9841	0.9841	1.0000	0.9841
		apCorr	0.8994	1.0000	0.9913	0.9098	0.9961	0.9213	0.9186	0.9191	1.0000	0.9193
		RMSE	0.0326	0.0273	0.0290	0.0280	0.0281	0.0283	0.0293	0.0278	0.0286	0.0285
	50	tauCorr	0.9788	0.9841	0.9947	0.9894	0.9947	0.9841	0.9894	0.9841	0.9947	0.9947
		apCorr	0.9061	0.9148	0.9259	0.9227	0.9259	0.9136	0.9210	0.9173	0.9259	0.9259
		RMSE	0.0287	0.0241	0.0247	0.0252	0.0261	0.0255	0.0257	0.0252	0.0254	0.0254
	60	tauCorr	0.9894	0.9947	1.0000	0.9841	0.9841	0.9947	0.9894	0.9841	0.9788	0.9894
		apCorr	0.9915	0.9973	1.0000	0.9186	0.9849	0.9973	0.9937	0.9158	0.9810	0.9232
		RMSE	0.0313	0.0266	0.0271	0.0278	0.0283	0.0279	0.0285	0.0280	0.0285	0.0281
	70	tauCorr	0.9788	0.9947	0.9788	0.9894	0.9788	0.9894	0.9841	0.9841	0.9841	0.9894
		apCorr	0.9157	0.9973	0.9030	0.9904	0.9124	0.9202	0.9174	0.9156	0.9159	0.9213
		RMSE	0.0285	0.0256	0.0271	0.0262	0.0265	0.0266	0.0267	0.0260	0.0267	0.0265

5.2 Experimental Results

5.2.1 General Trends. Figure 8 shows the correlation of the `AP vs nDCG@20` pair, averaged over the $H = 100$ samples, at the $t = 10, 30, 50, 70$ topic sizes[9] and the corresponding confidence interval; Kendall's τ correlation is drawn with a solid blue line (`allQ`) and a dashed red line (`no1Q`) while AP correlation τ_{AP} is drawn with a solid green line (`allQ`) and a dashed orange line (`no1Q`). Figure 8 basically plots the raw data contained in the cells of the table in Figure 7 for the rows corresponding to topic sizes $t = 10, 30, 50, 70$ across the columns of all the system sizes. The other evaluation measure pairs exhibit a consistent behaviour with respect to the one of the `AP vs nDCG@20` pair, which we use here as an example to discuss the main trends. Note that the `allQ` curves in the plots of Figure 8 are the same curves also shown in the corresponding plots of Figure 3.

[68] stated that the higher the range of scores of the ranked systems, the higher the correlation coefficient and Figure 8 basically generalizes their claim across many topic and system sizes. Indeed, we can observe that, for all the topics and system sizes, τ and τ_{AP} excluding first quartile systems (`no1Q`), i.e. reducing the range of the system scores, are consistently lower by their counterpart considering all the systems (`allQ`).

When can also note that the `no1Q` version of the curves tend to have bigger confidence intervals than the `allQ` one, especially at lower topic and system sizes. This makes sense when you consider that the `no1Q` version removes 25% of the systems and this particularly affects the lower topic and system sizes.

Finally, Figure 8 evidences how close is the behaviour of the `allQ` and `no1Q` curves, yet with different absolute values, for both τ and τ_{AP}.

5.2.2 Analysis. Figure 9 compares, for the $(t = 50, s = 100)$ combination, the `allQ` and `no1Q` RoMP as well as the `allQ` and `no1Q` curves. In other terms, it applies the analysis procedure described in Section 5.1 to the contents of the cell of the table in Figure 7 corresponding to $t = 50$ topics and $s = 100$ systems.

As shown in the parallel coordinates plot on the left, for both τ and τ_{AP}, the difference between `allQ` and `no1Q` consists of just one swap between two adjacent measure pairs, somehow in the middle for τ and at the top for τ_{AP}; as a consequence, in the case of τ, the correlations between the `allQ` and `no1Q` RoMP are tauCorr = 0.9947 and apCorr = 0.9954; in the case of τ_{AP}, the correlations between the `allQ` and `no1Q` rankings are tauCorr = 0.9947 and apCorr = 0.9259, a little lower due to the single swap happening at the top rank.

Figure 9 shows, on the right, the `allQ` and `no1Q` curves with mean centered around zero. We can observe how close they are, as also supported by the low RMSE which is RMSE = 0.0423 in the case of Kendall's τ correlation and RMSE = 0.0261 in the case of AP correlation τ_{AP}. We can also note that, for both τ and τ_{AP}, the `no1Q` curve is constantly above the `allQ` one for measure pairs in the top half of the RoMP while the opposite happens for measure pairs in the bottom half. This suggest that removing low performing systems somehow boosts more highly correlated measure pairs and narrows down the less correlated ones.

Finally, Table 5 (τ) and Table 6 (τ_{AP}) reports the results of the application of the analysis methodology described in Section 5.1 to each cell of the table show in Figure 7.

We can observe that the correlations between the `allQ` and `no1Q` RoMP are quite high, for all the possible topic and system sizes, being a bit lower just in case of very low numbers of topics and systems: they are typically over 0.9 for both τ (Table 5) and τ_{AP} (Table 6).

Overall, the grand mean across the cells of Tables 5 and 6 is: for Kendall's τ correlation tauCorr = 0.9723 and apCorr = 0.9580 and for AP correlation τ_{AP} tauCorr = 0.9852 and apCorr = 0.9538,

[9] The trends for the other topics are quite similar and we do not show them here for space reasons.

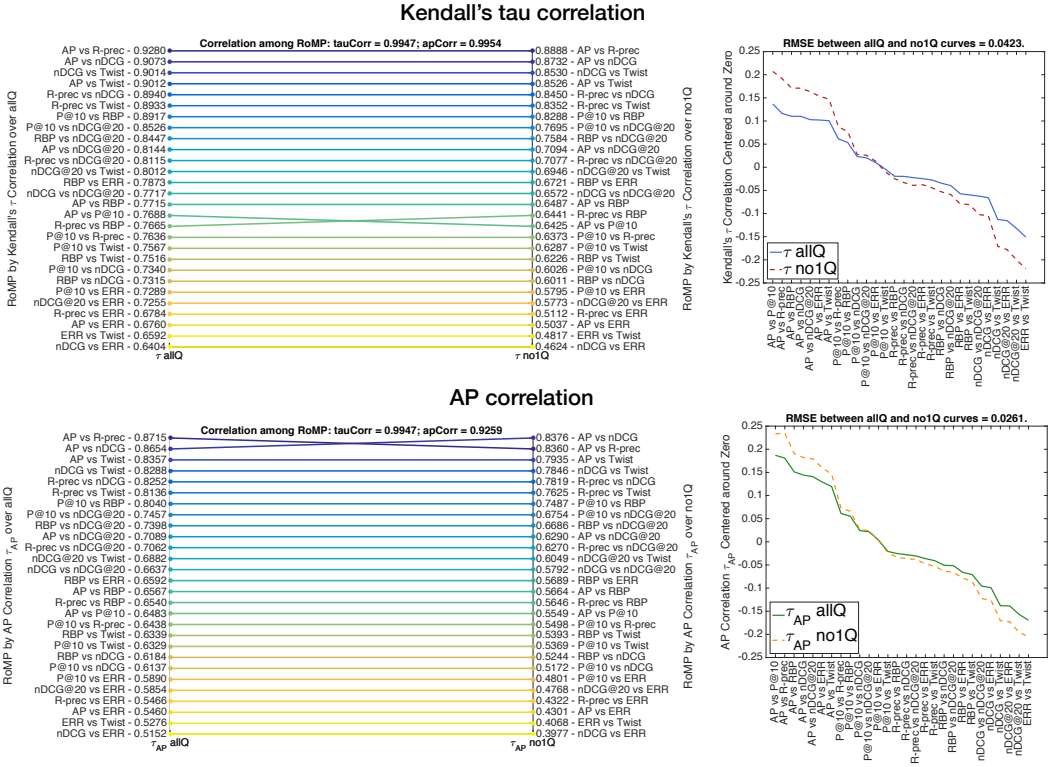

Fig. 9. Comparison of the `allQ` and `no1Q` τ and τ_{AP} versions for the $(t = 50, s = 100)$ level across all the pairs of evaluation measures. τ is on the top and τ_{AP} is on the bottom. On the left, the rankings of the evaluation measure pairs are ordered by the `allQ` and `no1Q` approaches: horizontal and oblique lines indicate, respectively, concordance and discordance between the rankings. On the right, the`allQ` and `no1Q` curves are shown but with means centered around zero.

indicating an almost perfect concordance between the `allQ` and `no1Q` cases. This is also supported by the RMSE which is typically low and whose grand mean is RMSE $= 0.0446$ for Kendall's τ correlation and RMSE $= 0.0271$ for AP correlation τ_{AP}

Overall, these analyses suggest that the `allQ` and `no1Q` approaches do not convey substantially different information when comparing a set of evaluation measures.

On the other hand removing first quartile systems, or lower performing systems according to some other criteria, may make the experiments more difficult to reproduce [25, 26] because of various reasons, e.g. overlooking to remove low performing systems, wrong implementation of the removal criterion, and so on.

Furthermore, these analyses highlight a serious issue concerning the difficulty in using and interpreting absolute thresholds, as pointed out also by [68]. The $\tau = 0.9$ threshold indicated by [76, 77], and then widely adopted by researchers, is well motivated and sensible when it comes to its interpretation of level above which we can consider rankings to be equivalent but its absolute value 0.9 is mostly bound to the specific experiments that have been conducted and to their setup.

For example, Figure 8 shows how for the `AP vs nDCG@20` pair, the difference between the `allQ` and the `no1Q` systems is to obtain a Kendall's τ score closer to or farther away from the 0.9 threshold;

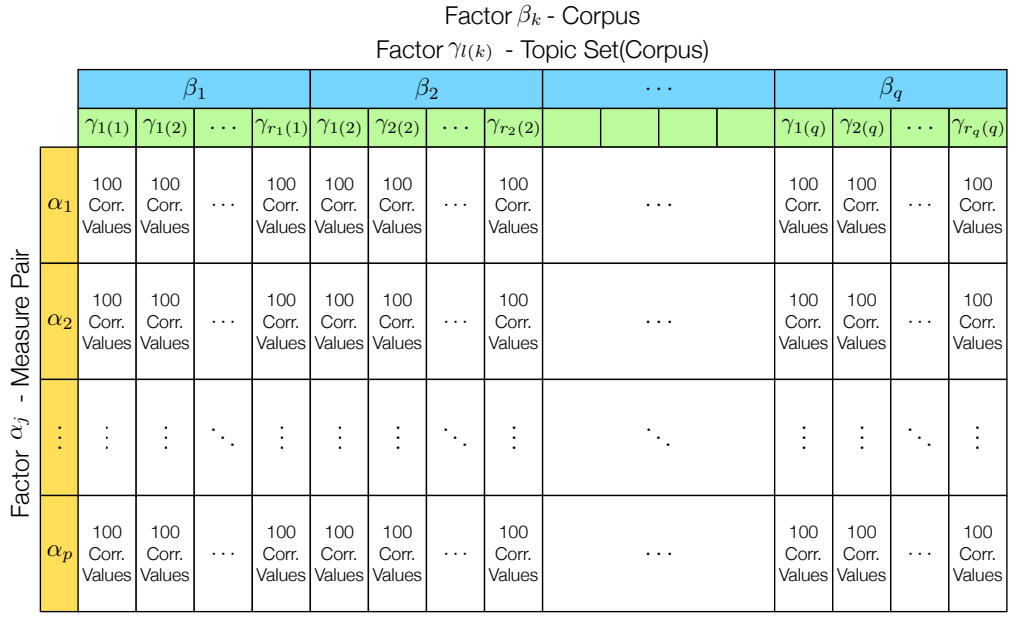

Fig. 10. Design of the GLMM model of equation (9) for the measure pair, corpus and topic set effects.

this, in turn, would make you reach quite different conclusions about the correlation between these two evaluation measures. And, if you consider that both the `allQ` and the `no1Q` choice does not substantially affect the positioning of the `AP vs nDCG@20` pair with respect to the others, this turns out to be a bit severe consequence of using absolute thresholds.

Therefore, for all these reasons, it should be carefully thought when it is really needed and beneficial to remove low performing systems and, instead, it is not detrimental of the reproducibility and ease of interpretation of the experimental results.

6 EFFECT OF EXPERIMENTAL COLLECTIONS

6.1 Methodology

To answer *RQ3*, we adopt an approach inspired by experimental design described in Section 4.1 but with changes to fit it to the case at hand. We create a GoP for each of the 7 adopted collections, from `T07` to `T15`. We use all the topics available in each collection, i.e. a topic size $t = 50$ for each collection except for `T13` which has $t = 49$, and we set a system size $s = 100$. This (t, s) combination represents the typical settings from a reasonably large track in an evaluation campaign.

We randomly draw $H = 100$ samples of $s = 100$ systems out of the $1,326$ possible systems and we run these system over the topics of all the collections from `T07` to `T15`. For each sample and each collection, we compute all the evaluation measures and this produces $H = 100$ RoS for each measure and collection. Finally, for each collection, we compute both τ and τ_{AP} over all the possible measure pairs for each of these $H = 100$ RoS.

This setup allows us to isolate the effects of the topic sets and corpora, since the same systems are evaluated over all the used GoPs.

We adopt the mixed design shown in Figure 10: subjects κ_i are the $H = 100$ samples for each combination of topics and systems; factors α_j and β_k correspond, respectively, to measure pairs

and document corpora; and, factor $\gamma_{l(k)}$, nested within factor β_k, represent topic sets. Factor $\gamma_{l(k)}$ is nested within factor β_k because the same corpus is used by more collections but each topic set is used just with one corpus: for example, T07 and T08 both use the TIPSTER corpus but they use two different topic sets, i.e. 351-400 and 401-450, respectively, as reported in Section 3.1. This leads to the following GLMM model:

$$Y_{ijkl} = \underbrace{\mu.... + \kappa_i + \alpha_j + \beta_k + \gamma_{l(k)}}_{\text{Main Effects}} + \underbrace{(\alpha\beta)_{jk} + (\alpha\gamma)_{jl(k)}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}} \tag{9}$$

where:

- Y_{ijkl} is the correlation value (either τ or τ_{AP}) of the i-th subject in the j-th, k-th, and l-th factors;
- $\mu....$ is the grand mean;
- κ_i is the effect of the i-th subject, i.e. the $h = 1, \ldots, H$ samples for each topics and system combination, where $\kappa_i = \mu_{i...} - \mu....$ and $\mu_{i...}$ is the mean of the i-th subject;
- $\alpha_j = \mu_{.j..} - \mu....$ is the effect of the j-th factor, i.e. measure pairs, where $\mu_{.j..}$ is the mean of the j-th factor. Considering that we are experimenting with 8 evaluation measures, there are $\binom{8}{2} = 28$ measure pairs, i.e. 28 levels for factor α_j;
- $\beta_k = \mu_{..k.} - \mu....$ is the effect of the k-th factor, i.e. corpora, where $\mu_{..k.}$ is the mean of the k-th factor; there are 3 levels for factor β_k, each one corresponding to a different corpus, namely TIPSTER, WT10g, and GOV2;
- $\gamma_{l(k)} = \mu_{...l(k)} - \mu....$ is the effect of the l-th factor, i.e. topic sets, where $\mu_{...l(k)}$ is the mean of the l-th factor; for TIPSTER and GOV2 there are 2 levels of factor $\gamma_{l(k)}$, corresponding to topics sets 351-400 and 401-450 for TIPSTER and topic sets 451-500 and 501-550 for WT10g; for GOV2 there are 3 levels of factor $\gamma_{l(k)}$, corresponding to topics sets 701-750, 751-800, and 801-850;
- $(\alpha\beta)_{jk}$ and $(\alpha\gamma)_{jl(k)}$ are, respectively, the interactions between measure pairs and corpora and between measure pairs and topic sets;
- 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.

Considering that there are 28 measure pairs, 2 corpora with 2 topic sets each and 1 corpus with 3 topic sets and that, for each combination of these factors, we use 100 subjects, overall this amounts to analyzing 19,600 correlation values for both τ and τ_{AP}.

6.2 Experimental Results

6.2.1 General Trends. Figure 11 shows the τ (top) and τ_{AP} (bottom) correlation values, averaged over the $H = 100$ samples, across the T07, T08, T09, T10, T13, T14, and T15 collections. On the left, you can see the actual average correlation values; on the right, there are the RoMP produced by these correlation values. Figure 11 on the left basically plots the raw data contained in the cells of the table in Figure 10, where each line in the plot corresponds to a row in the table and each collection on the x-axis in the plot corresponds to a column in the table identified by a (corpus, topic set) pair, e.g. T07 collection corresponds to the (TIPSTER, 351-400) column.

It clearly emerges from both the left and the right plots of Figure 11 that there is quite a big variation across the different collections and that it affects both τ and τ_{AP}. For example in the case of τ, the nDCG vs Twist pair (solid thick blue line) goes above and below the 0.9 threshold, making us to reach different conclusions about these two measures depending on the collection they are tested against. A similar example is the AP vs nDCG pair (dotted thick blue line) which

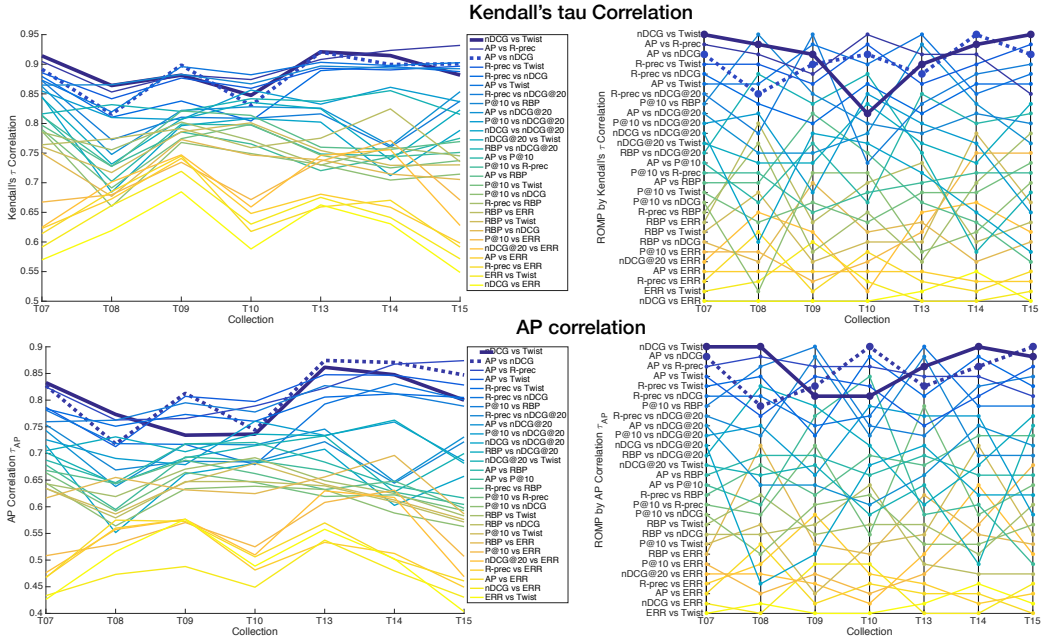

Fig. 11. Kendall's τ correlation (top) and AP correlation τ_{AP} (bottom) across the T07, T08, T09, T10, T13, T14, and T15 GoP. On the left, there are the correlation values, averaged over the $H = 100$ samples. On the right, there are the RoMP, ordered by their correlation on the T07 collection: the lines show how these ranks varies across the other collections.

goes above and under the 0.9 threshold many times and keeps crossing with the nDCG vs Twist pair. If we look at the parallel coordinates plots on the right, we can see how the changes in the correlation values affect the ranking of these two measures pairs, with nDCG vs Twist and AP vs nDCG becoming the top ranked pair depending on the collection.

Overall, this suggest that collections have an impact on the correlation among evaluation measures and *RQ3* investigates how much of this impact is due to the corpora and how much to the topic sets.

6.2.2 GLMM and ANOVA Analysis. Tables 7 and 8 report the results of the ANOVA analyses for the GLMM of equation (9) for τ and τ_{AP}, respectively: all the effects are statistically significant; they are all large size effects with the exception of the corpus effect which is a medium size one; and, the power is 1 for all the analyzed effects.

The most prominent effect is the measure pair one; as discussed also in Section 4.2, this makes sense since correlation values vary quite a lot from one measure pair to another one and it is what actually differentiates evaluation measures.

The corpus effect is the smallest one and, in the case of τ_{AP}, it is about half the size than in the case of τ, indicating that the former is less sensitive to the change of corpora. This might due to the fact that good systems are top ranked over different corpora and τ_{AP} focuses more on top ranked systems. You can also note how the topic set effect is about 2–3 times bigger than the corpus effect, indicating that it dominates.

Table 7. Kendall's τ correlation: ANOVA table for the GLMM of equation (9), considering measure pair, corpus and topic set effects on the T07, T08, T09, T10, T13, T14, and T15 GoPs.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	Power
Subject	5.1083	99	0.0516	96.2944	0.0000		
Measure Pair	114.9086	27	4.2559	7,942.3019	0.0000	0.9162	1.0000
Corpus	1.4975	2	0.7488	1,397.3473	0.0000	0.1247	1.0000
Topic Set(Corpus)	4.0247	4	1.0062	1,877.7154	0.0000	0.2769	1.0000
Measure Pair*Corpus	5.1302	54	0.0950	177.2957	0.0000	0.3269	1.0000
Measure Pair*Topic Set(Corpus)	13.8983	108	0.1287	240.1563	0.0000	0.5686	1.0000
Error	10.3446	19,305	0.0005				
Total	163.8230	19,599					

Table 8. AP correlation τ_{AP}: ANOVA table for the GLMM of equation (9), considering measure pair, corpus and topic set effects on the T07, T08, T09, T10, T13, T14, and T15 GoPs.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	Power
Subject	6.7980	99	0.0687	62.9948	0.0000		
Measure Pair	178.9458	27	6.6276	6,080.1540	0.0000	0.8933	1.0000
Corpus	1.4495	2	0.7247	664.8616	5.8e-280	0.0634	1.0000
Topic Set(Corpus)	5.5632	4	1.3908	1,275.9127	0.0000	0.2065	1.0000
Measure Pair*Corpus	10.0543	54	0.1862	170.8107	0.0000	0.3187	1.0000
Measure Pair*Topic Set(Corpus)	14.8491	108	0.1375	126.1342	0.0000	0.4081	1.0000
Error	21.0433	19,305	0.0011				
Total	253.0846	19,599					

Overall, this answers the question of [14] about the impact of changing the topic set: indeed, we experimented the same systems across different collections, i.e. corpora and topic sets, and topic sets demonstrate to influence correlation a lot. Moreover, if we compare $\hat{\omega}^2_{\langle \text{Topic Set(Corpus)} \rangle}$ in Tables 7 and 8 with $\hat{\omega}^2_{\langle \text{Topic Size} \rangle}$ in Tables 2 and 3, we can see how they have comparable effect sizes, indicating that these two different phenomena impact a lot the correlation among evaluation measures.

When it comes to the interaction effects, it is interesting to note how the measure pair and corpus interaction is about 3–5 times bigger than the corpus effect alone and that measure pair and topic set interaction is about 2 times bigger than topics set effect alone. This suggest that the variation we observed in Figure 11 is not only due to the corpus and topic set effects alone but also, and mostly, to the interaction these effects have with the evaluation measure pairs. Moreover, if we compare the Measure Pair*Topic Size interaction (Tables 2 and 3) with the Measure Pair*Topic Set(Corpus) interaction (Tables 7 and 8), we can observe how the former has a $\hat{\omega}$ about 25 times smaller than the latter, indicating that evaluation measure pairs interact more with the specific topics at hand rather than with a specific number of topics.

6.2.3 Main Effects Analysis. Figure 12 shows the main effects plot for both τ (in blue) and τ_{AP} (in green). On the left, it shows the expected values of τ and τ_{AP}, net from the effects of the corpora and topic sets; these values are also reported in Table 9, which can be held as reference correlation values, distilled across many experimental collections. As anticipated in Section 2.2, τ_{AP} is not symmetric and we used as reference the evaluation measure reported in the row of Table 9 which corresponds to the first one in the labels of measure pair plot in Figure 12.

The effect of the corpus (middle plot of Figure 12) varies: for τ there is an increase from the TIPSTER corpus, which basically represents a news retrieval task, to the WT10g one, which is a Web retrieval task, while there is a very slight decrease between WT10g and GOV2, which are both Web retrieval tasks, but the latter collection is about 2 orders of magnitude bigger. On the other hand,

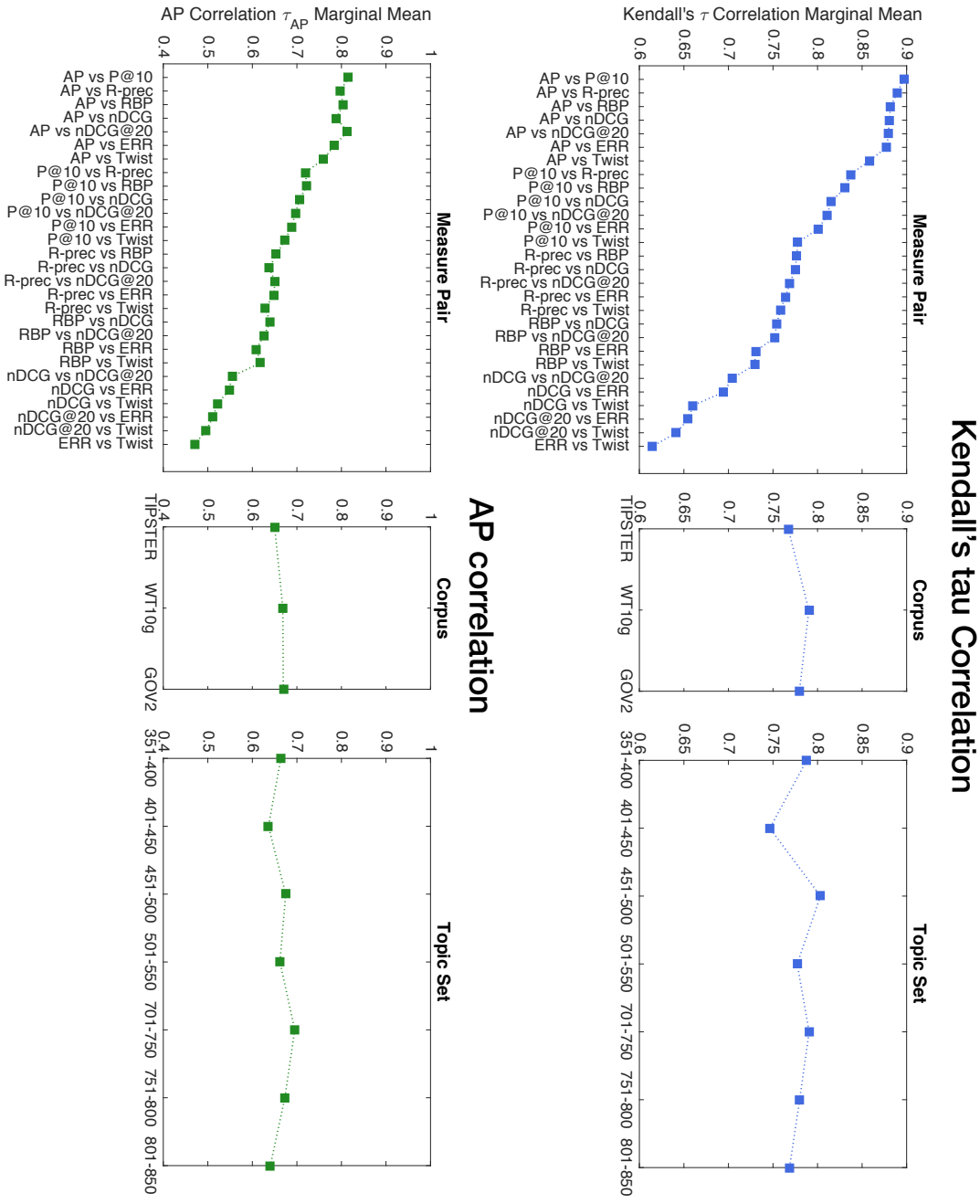

Fig. 12. Main effects for the measure pair (on the left), corpus (on the middle) and topic set (on the right) factors on the T07, T08, T09, T10, T13, T14, and T15 GoP. Kendall's τ correlation is at the top in blue, AP correlation τ_{AP} is at the bottom in green. Note that τ_{AP} is not symmetric and we used as reference the first evaluation measure in the labels of the measure pair plot. Also note that the figure is rotated and indications like left, middle and right all refer to when you rotate the figure to read it.

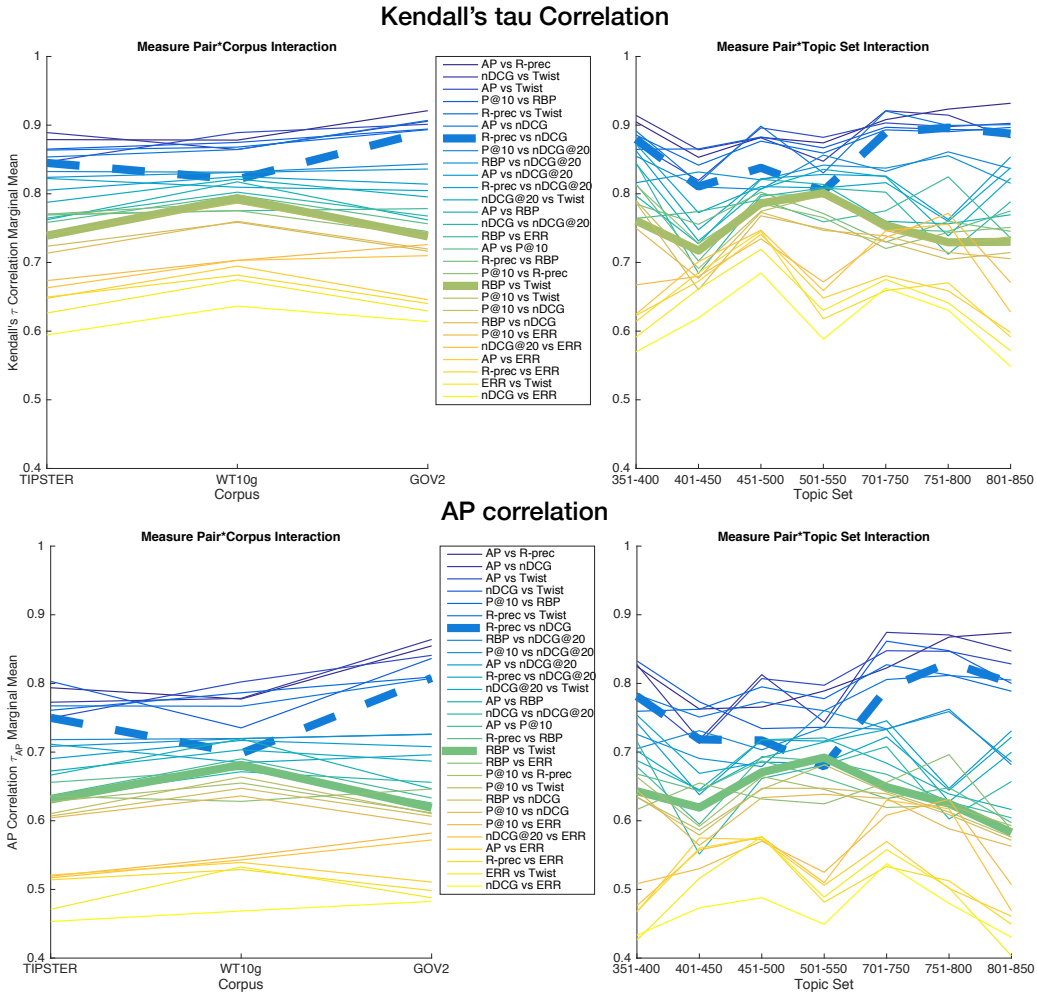

Fig. 13. Interaction effects for the Measure Pair*Corpus (on the left) and Measure Pair*Topic Set (on the right) factors. Kendall's τ correlation is at the top, AP correlation τ_{AP} is at the bottom.

τ_{AP} shows a very modest increase from TIPSTER to WT10g and no increase from WT10g to GOV2, being almost insensitive to different corpora. The effect of the topic set (right plot of Figure 12) is more pronounced and less regular.

Finally, we can note a trend observed also in Section 4.2.5: the τ and τ_{AP} curves look very similar, provided that τ_{AP} has lower values than τ and it is somehow translated towards the bottom. We applied the same analysis conducted in Section 4.2.5 to the experimental data used here and we drew the same conclusions about the similarities between τ and τ_{AP}; the detailed analysis is not reported here for space reasons.

6.2.4 Interaction Effects Analysis. Figure 13 shows the interaction plots for the Measure Pair*Corpus (on the left) and Measure Pair*Topic Set (on the right), where τ is at the top and τ_{AP} is at the bottom.

It can be noted how the interaction between measure pairs and corpora/topic sets is quite high, for both τ and τ_{AP}, and how differently each measure pair is increased or decreased by a given corpus or topic set.

For example, in the left plots, we can note how the TIPSTER and GOV2 corpora have somehow similar effects on the R-prec vs nDCG (dashed thick blue line) and RBP vs Twist (solid thick green line) while WT10g narrows down R-prec vs nDCG and boosts RBP vs Twist to the extent that they perform almost the same. In the right plots, we can see how much different topic sets influence these two measure pairs: while, in general, R-prec vs nDCG is higher than RBP vs Twist, topic set 501-550 narrows down R-prec vs nDCG and boost RBP vs Twist so much that they perform the same.

This further stresses the need for carefully taking into account the corpus and topic set effects because, together with their main effects, they can substantially affect the correlation between two evaluation measures, changing from collection to collection and making us draw possibly different conclusions about such measures.

6.2.5 Comparison between RQ1 and RQ3 Measure Pair Main Effects. We investigate the relationship between the main effects of the measure pair factor, i.e. the expected correlation values between evaluation measures, as computed for research questions *RQ1* and *RQ3* and reported, respectively, in Table 4 and Table 9. To this end, we adopt the same methodology used in Section 4.2.5 for comparing τ and τ_{AP}, i.e. we assess how close *RQ1* and *RQ3* rank evaluation measure pairs and how close are the *RQ1* and *RQ3* curves, once you have centered their mean around zero.

We are interested in understanding whether the two different kinds of analyses for *RQ1* and *RQ3*, defined by the GLMM models of equations (8) and (9), lead us to draw similar or different conclusions.

Figure 14 shows the comparison between the main effects of the measure pair factor for *RQ1* and *RQ3*. On the left, you can see the rankings of the evaluation measure pairs in *RQ1* and *RQ3* and it can be noted that they are reasonably similar. When considering Kendall's τ correlation, the correlations between the *RQ1* and *RQ3* RoMP are tauCorr = 0.9524 and apCorr = 0.8996, indicating they are quite similar but with some swaps in the top ranks. When considering AP correlation τ_{AP} , the correlations between the *RQ1* and *RQ3* RoMP are tauCorr = 0.9365 and apCorr = 0.8456, indicating they are quite similar again but with some more swaps in the top ranks.

On the right of Figure 14, we show the main effects of the measure pair factor for *RQ1* and *RQ3* but with means centered around zero. The figure further highlights how close the *RQ1* and *RQ3* curves are across the evaluation measure pairs with a very small RMSE = 0.0185, when considering Kendall's τ correlation, and RMSE = 0.0260, when considering AP correlation τ_{AP}.

Overall, these analyses suggest that both *RQ1* and *RQ3* are in agreement and provide stable and consistent results. Therefore, we can consider the expected correlation among measures reported in Tables 4 and 9 as a good approximation of the real (unknown) ones.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed three research questions about the Kendall's τ correlation and the AP correlation τ_{AP} coefficients among evaluation measures:

RQ1 What is the effect of the number of systems and topics?
RQ2 What is the effect of removing low performing systems?
RQ3 What is the effect of the experimental collections?

To conduct these analyses, we developed a methodology based on GLMM and ANOVA, which allowed us to break down and isolate the different effects, in order to appreciate their impact and

Table 9. Main effects of the τ and τ_{AP} correlation coefficients for each evaluation measure pair, net from the corpus and topic set effects, on the T07, T08, T09, T10, T13, T14, and T15 GoP. These are the values plotted in Figure 12 on the left. Note that τ_{AP} is not symmetric and we used as reference the evaluation measure reported in the row.

		AP	P@10	R-prec	RBP	nDCG	nDCG@20	ERR	Twist
AP	τ	1.0000	0.7688	0.8968	0.7779	0.8796	0.8147	0.6603	0.8821
	τ_{AP}	1.0000	0.6505	0.8151	0.6729	0.8134	0.7055	0.5218	0.8038
P@10	τ	–	1.0000	0.7581	0.8804	0.7305	0.8369	0.7045	0.7519
	τ_{AP}	–	1.0000	0.6286	0.7889	0.6092	0.7192	0.5543	0.6256
R-prec	τ	–	–	1.0000	0.7642	0.8579	0.8111	0.6547	0.8776
	τ_{AP}	–	–	1.0000	0.6484	0.7600	0.6972	0.5115	0.7843
RBP	τ	–	–	–	1.0000	0.7294	0.8312	0.7749	0.7542
	τ_{AP}	–	–	–	1.0000	0.6183	0.7219	0.6381	0.6402
nDCG	τ	–	–	–	–	1.0000	0.7767	0.6148	0.8892
	τ_{AP}	–	–	–	–	1.0000	0.6540	0.4702	0.7980
nDCG@20	τ	–	–	–	–	–	1.0000	0.6946	0.8007
	τ_{AP}	–	–	–	–	–	1.0000	0.5480	0.6874
ERR	τ	–	–	–	–	–	–	1.0000	0.6414
	τ_{AP}	–	–	–	–	–	–	1.0000	0.4959
Twist	τ	–	–	–	–	–	–	–	1.0000
	τ_{AP}	–	–	–	–	–	–	–	1.0000

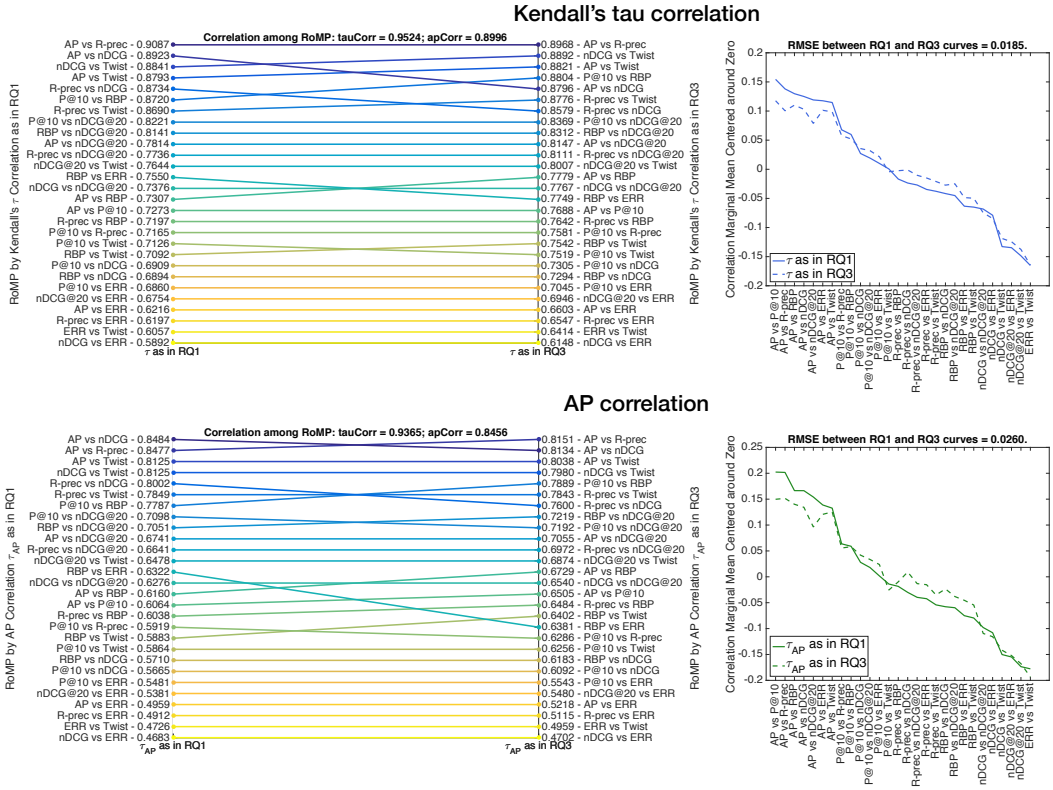

Fig. 14. Comparison of the main effects of measure pair factor for *RQ1* and *RQ3*. On the left, the rankings of the evaluation measure pairs are ordered as in *RQ1* and *RQ3*: horizontal and oblique lines indicate, respectively, concordance and discordance between the rankings. On the right, the same main effects shown in Figures 4 and 12 but with means centered around zero. τ is at the top and τ_{AP} is at the bottom.

to derive expected correlation values among evaluation measures, net from these effects, than can be taken as reference values.

An important part of this methodology is the use of GoPs, which represent nearly all the state-of-the-art components commonly used in English retrieval. These GoPs allow us both to have a much larger sample space than the one typically available with the runs submitted to an evaluation campaign and to keep the variance due to systems controlled, thus improving the experimental outcomes.

With respect to *RQ1*, we discovered that the number of topics impacts more than the number of systems and that the number of systems does not cause the correlation to steadily increase but it reaches a stable point quite quickly. The typical setting with 50 topics and about 100 systems produces good results and it is an effective tradeoff between the effort for topic and ground-truth creation and the quality of the results. We also observed that the behavior of τ and τ_{AP} is quite consistent when comparing a whole set of evaluation measures, yet producing different absolute correlation values.

When it comes to *RQ2*, we found out that removing the low performing systems does not convey information substantially different from not removing them, when you consider a whole set of evaluation measures. On the other hand, removing or not the low performing systems changes the absolute correlation values and this make the use of absolute thresholds problematic, possibly affecting the reproducibility and ease of interpretation of the experiments.

As far as *RQ3* is concerned, we observed that corpora and topic sets considerably affect correlation, with the latter effect being more prominent, and that there is quite a lot of interaction between correlation among evaluation measures and the corpus/topic set at hand, making the correlation values increase or decrease substantially. Moreover, compared to the outcomes of *RQ1*, we noted that the effect of the number of topics is comparable to the one of topic sets but the interaction between topic sets and measure pairs is much greater than the one between topic sizes and measure pair; this suggest that not only the number of topics matters but also which topics you actually use.

Finally, it is interesting to note how the main effects of the measure pair factor, i.e. the expected correlation values, determined with the GLMM of *RQ1* and *RQ3* agree each other and are quite consistent, yet being produced by different models over different data sets. This suggests that the expected correlation values reported in Tables 4 and 9 are a good approximation of the real ones.

Overall, this paper delivered two major outcomes: a methodology to investigate the properties of one of the tools, i.e. correlation analysis, we use to study evaluation measures and the findings about the correlation among evaluation measures obtained from the application of that methodology. Starting from these two results we envisage two future areas of work: one concerns the extension of the methodology itself in order to study further properties of the correlation among evaluation measures; the other concerns the application of the methodology developed here to other relevant tools we use to study evaluation measures.

With respect to the extension of the methodology, we plan to investigate how the different system components affect the correlation among evaluation measures. This interest stems from our previous work on breaking down the contribution of different components to the overall system performances [29] and from the open question "do different components induce somehow different correlation values?". In order to achieve this goal we will need to exploit the GoP in a different way, which allows us to group RoS originated by different types of IR components, and to develop a different ANOVA design to properly analyze such new data.

With respect to the application of this methodology to other tools, we intend to investigate *RQ1*, *RQ2*, and *RQ3* in the case of the *discriminative power* [59, 62], which is used to assess the degree to which an evaluation measure can detect differences between systems relative to other

evaluation measures. This kind of analysis is adopted by various authors to study the quality of evaluation measures [17, 19, 23, 32, 70]. Moreover, [65] suggests that a high discriminative power is a necessary condition for a good quality evaluation measures. Nevertheless, there is still no study on the factors affecting the discriminative power and the application of the methodology developed here will represent a first step in this direction.

ACKNOWLEDGMENTS

The author wishes to warmly thank Gianmaria Silvello for the great deal of help and effort put in preparing the *Grid of Points (GoP)* used for the analyses carried out in this paper. The author also thanks Gianfranco Bilardi for having made available the high performance computing facilities needed to produce the GoP utilized in the paper. Last but not least, the author sincerely thanks the associate editor and the anonymous reviewers for the thorough reviews and the challenging discussions which greatly helped in improving this paper.

REFERENCES

[1] G. Amati and C. J. van Rijsbergen. 2002. Probabilistic Models of Information Retrieval based on measuring the Divergence From Randomness. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 357–389.

[2] J. Arguello, M. Crane, F. Diaz, J. Lin, and A. Trotman. 2015. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum* 49, 2 (December 2015), 107–116.

[3] J. A. Aslam and E. Yilmaz. 2005. A Geometric Interpretation and Analysis of R-precision. In *Proc. 14th International Conference on Information and Knowledge Management (CIKM 2005)*, O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken (Eds.). ACM Press, New York, USA, 664–671.

[4] J. A. Aslam, E. Yilmaz, and V. Pavlu. 2005. A Geometric Interpretation of R-precision and Its Correlation with Average Precision. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait (Eds.). ACM Press, New York, USA, 573–574.

[5] J. A. Aslam, E. Yilmaz, and V. Pavlu. 2005. The Maximum Entropy Method for Analyzing Retrieval Measures. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait (Eds.). ACM Press, New York, USA, 27–34.

[6] R. Baeza-Yaetes, C. Castillo, M. Marin, and A. Rodriguez. 2005. Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering. In *Proc. 14th International Conference on World Wide Web (WWW 2005)*, A. Ellis, T. Hagino, F. Douglis, and P. Raghavan (Eds.). ACM Press, New York, USA, 864–872.

[7] A. Z. Broder, R. Lempel, F. Maghoul, and J. Pedersen. 2006. Efficient PageRank approximation via graph aggregation. *Information Retrieval* 9, 2 (March 2006), 123–138.

[8] C. Buckley and E. M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, M. Sanderson, K. Järvelin, J. Allan, and P. Bruza (Eds.). ACM Press, New York, USA, 25–32.

[9] 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.

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

[11] 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.

[12] 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.

[13] J. Callan and M. Connell. 2001. Query-Based Sampling of Text Databases. *ACM Transactions on Information Systems (TOIS)* 19, 2 (April 2001), 97–130.

[14] B. A. Carterette. 2009. On Rank Correlation and the Distance Between Rankings. In *Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel (Eds.). ACM Press, New York, USA, 436–443.

[15] B. A. Carterette and J. Allan. 2005. Incremental Test Collections. In *Proc. 14th International Conference on Information and Knowledge Management (CIKM 2005)*, O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken (Eds.). ACM Press, New York, USA, 680–687.

[16] 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.

[17] A. Chuklin, P. Serdyukov, and M. de Rijke. 2013. Click Model-Based Information Retrieval Metrics. In *Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*, G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai (Eds.). ACM Press, New York, USA, 493–502.

[18] 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.

[19] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. 2011. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proc. 4th ACM International Conference on Web Searching and Data Mining (WSDM 2011)*, I. King, W. Nejdl, and H. Li (Eds.). ACM Press, New York, USA, 84–75.

[20] 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.

[21] G. M. Di Nunzio and G. Silvello. 2015. A Graphical View of Distance Between Rankings: The Point and Area Measure. In *Proc. 6th Italian Information Retrieval Workshop (IIR 2015)*, P. Boldi, R. Perego, and F. Sebastiani (Eds.). CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1404/.

[22] R. Fagin, R. Kumar, and D. Sivakumar. 2003. Comparing top k lists. *SIAM Journal on Discrete Mathematics* 17, 1 (2003), 134–160.

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

[24] M. Ferrante, N. Ferro, and M. Maistro. 2014. Rethinking How to Extend Average Precision to Graded Relevance. In *Information Access Evaluation – Multilinguality, Multimodality, and Interaction. Proceedings of the Fifth International Conference of the CLEF Initiative (CLEF 2014)*, E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms (Eds.). Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany, 19–30.

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

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

[27] 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.

[28] N. Ferro and G. Silvello. 2015. Rank-Biased Precision Reloaded: Reproducibility and Generalization. In *Advances in Information Retrieval. Proc. 37th European Conference on IR Research (ECIR 2015)*, N. Fuhr, A. Rauber, G. Kazai, and A. Hanbury (Eds.). Lecture Notes in Computer Science (LNCS) 9022, Springer, Heidelberg, Germany, 768–780.

[29] 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.

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

[31] N. Ferro and G. Silvello. 2017. Towards an Anatomy of IR System Component Performances. *Journal of the American Society for Information Science and Technology (JASIST)* (2017).

[32] 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.

[33] N. Gao, M. Bagdouri, and D. W. Oard. 2016. Pearson Rank: A Head-Weighted Gap-Sensitive Score-Based Correlation Coefficient. 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, 941–944.

[34] N. Gao and D. W. Oard. 2015. A Head-Weighted Gap-Sensitive Correlation Coefficient. In *Proc. 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, R. Baeza-Yates, M. Lalmas, A. Moffat, and B. Ribeiro-Neto (Eds.). ACM Press, New York, USA, 799–802.

[35] 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.

[36] 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.

[37] Y. Hochberg and A. C. Tamhane. 1987. *Multiple Comparison Procedures.* John Wiley & Sons, USA.

[38] A. Inselberg. 2009. *Parallel Coordinates. Visual Multidimensional Geometry and Its Applications.* Springer-Verlag, New York, USA.

[39] 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.

[40] J. Kekäläinen. 2005. Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. *Information Processing & Management* 41, 5 (September 2005), 1019–1033.

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

[42] J. F. Kenney and E. S. Keeping. 1954. *Mathematics of Statistics – Part One* (3rd ed.). D. Van Nostrand Company, Princeton, USA.

[43] R. Krovetz. 2000. Viewing morphology as an inference process. *Artificial Intelligence* 118, 1–2 (April 2000), 277–294.

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

[45] R. Kumar and S. Vassilvitskii. 2010. Generalized Distances between Rankings. In *Proc. 19th International Conference on World Wide Web (WWW 2010)*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti (Eds.). ACM Press, New York, USA, 571–580.

[46] J. B. Lovins. 1971. Error Evaluation for Stemming Algorithms as Clustering Algorithms. *Journal of the American Society for Information Science (JASIS)* 22, 1 (January/February 1971), 28–40.

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

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

[49] P. McNamee and J. Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7, 1-2 (January 2004), 73–97.

[50] M. Melucci. 2007. On Rank Correlation in Information Retrieval Evaluation. *SIGIR Forum* 41, 1 (June 2007), 18–33.

[51] M. Melucci. 2009. Weighted Rank Correlation in Information Retrieval Evaluation. In *Information Retrieval Technology – Proc. 5th Asia Information Retrieval Symposium (AIRS 2009)*, G. G. Lee, D. Song, C.-Y. Lin, A. Aizawa, K. Kuriyama, M. Yoshioka, and T. Sakai (Eds.). Lecture Notes in Computer Science (LNCS) 5839, Springer, Heidelberg, Germany, 75–86.

[52] 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.

[53] 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.

[54] M. F. Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (July 1980), 130–137.

[55] 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.

[56] S. E. Robertson and U. Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval (FnTIR)* 3, 4 (2009), 333–389.

[57] T. Roelleke. 2013. *Information Retrieval Models. Foundations and Relationships.* Morgan & Claypool Publishers, USA.

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

[59] T. Sakai. 2006. Evaluating Evaluation Metrics based on the Bootstrap. In *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, E. N. Efthimiadis, S. Dumais, D. Hawking, and K. Järvelin (Eds.). ACM Press, New York, USA, 525–532.

[60] T. Sakai. 2007. Alternatives to Bpref. 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, 71–78.

[61] T. Sakai. 2007. On the reliability of information retrieval metrics based on graded relevance. *Information Processing & Management* 42, 2 (March 2007), 531–548.

[62] T. Sakai. 2012. Evaluation with Informational and Navigational Intents. In *Proc. 21st International Conference on World Wide Web (WWW 2012)*, A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab (Eds.). ACM Press, New York, USA, 499–508.

[63] 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.

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

[65] T. Sakai, N. Craswell, R. Song, S. E. Robertson, Z. Dou, and C.-Y. Lin. 2010. Simple Evaluation Metrics for Diversified Search Results. In *Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010)*, T. Sakai, M. Sanderson, and W. Webber (Eds.). National Institute of Informatics, Tokyo, Japan, 42–50.

[66] G. Salton and M. J. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, USA.

[67] M. Sanderson and H. Joho. 2004. Forming Test Collections with No System Pooling. In *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, M. Sanderson, K. Järvelin, J. Allan, and P. Bruza (Eds.). ACM Press, New York, USA, 33–40.

[68] M. Sanderson and I. Soboroff. 2007. Problems with Kendall's Tau. 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, 839–840.

[69] G. S. Shieh. 1998. A weighted Kendall's tau statistic. *Statistics & Probability Letters* 39, 1 (July 1998), 17–24.

[70] M. D. Smucker and C. L. A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. 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, 95–104.

[71] M. D. Smucker, G. Kazai, and M. Lease. 2014. Overview of the TREC 2013 Crowdsourcing Track. In *The Twenty-Second Text REtrieval Conference Proceedings (TREC 2013)*, E. M. Voorhees (Ed.). National Institute of Standards and Technology (NIST), Special Publication 500-302, Washington, USA.

[72] I. Soboroff, C. Nicholas, and P. Cahan. 2001. Ranking Retrieval Systems without Relevance Judgments. In *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel (Eds.). ACM Press, New York, USA, 66–73.

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

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

[75] S. Vigna. 2015. A Weighted Correlation Index for Rankings with Ties. In *Proc. 24th International Conference on World Wide Web (WWW 2015)*, A. Gangemi, S. Leonardi, A. Panconesi, K. Gummadi, and C. Zhai (Eds.). ACM Press, New York, USA, 1166–1176.

[76] E. M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel (Eds.). ACM Press, New York, USA, 315–323.

[77] E. M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management* 36, 5 (September 2000), 697–716.

[78] E. M. Voorhees. 2001. Evaluation by Highly Relevant Documents. In *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel (Eds.). ACM Press, New York, USA, 74–82.

[79] E. M. Voorhees. 2014. The Effect of Sampling Strategy on Inferred Measures. In *Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, and K. Järvelin (Eds.). ACM Press, New York, USA, 1119–1122.

[80] 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.

[81] E. M. Voorhees and D. K. Harman. 1998. Overview of the Seventh Text REtrieval Conference (TREC-7). In *The Seventh Text REtrieval Conference (TREC-7)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and

Technology (NIST), Special Publication 500-242, Washington, USA, 1–24.

[82] 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.

[83] M. P. Wand and M. C. Jones. 1995. *Kernel Smoothing*. Chapman and Hall/CRC, USA.

[84] W. Webber, A. Moffat, and J. Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Transactions on Information Systems (TOIS)* 4, 28 (November 2010), 20:1–20:38.

[85] R. W. White, I. Ruthven, J. M. Jose, and C. J. van Rijsbergen. 2005. Evaluating Implicit Feedback Models Using Searcher Simulations. *ACM Transactions on Information Systems (TOIS)* 23, 3 (July 2005), 325–361.

[86] E. Yilmaz and J. A. Aslam. 2006. Estimating Average Precision With Incomplete and Imperfect Judgments. In *Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006)*, P. S. Yu, V. Tsotras, E. A. Fox, and C.-B. Liu (Eds.). ACM Press, New York, USA, 102–111.

[87] E. Yilmaz, J. A. Aslam, and S. E. Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani (Eds.). ACM Press, New York, USA, 587–594.

[88] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. 2005. Learning to Estimate Query Difficulty – Learning to Estimate Query Difficulty Including Applications to Missing Content Detection and Distributed Information Retrieval. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait (Eds.). ACM Press, New York, USA, 512–519.

[89] C. Zhai. 2008. Statistical Language Models for Information Retrieval. A Critical Review. *Foundations and Trends in Information Retrieval (FnTIR)* 2, 3 (2008), 137–213.

Received 1 September 2016; revised 19 December 2016; revised 1 April 2017; accepted 9 June 2017