# Reproduce. Generalize. Extend. On Information Retrieval Evaluation without Relevance Judgments

KEVIN ROITERO, MARCO PASSON, GIUSEPPE SERRA, and STEFANO MIZZARO,
University of Udine

The evaluation of retrieval effectiveness by means of test collections is a commonly used methodology in the information retrieval field. Some researchers have addressed the quite fascinating research question of whether it is possible to evaluate effectiveness completely automatically, without human relevance assessments. Since human relevance assessment is one of the main costs of building a test collection, both in human time and money resources, this rather ambitious goal would have a practical impact. In this article, we reproduce the main results on evaluating information retrieval systems without relevance judgments; furthermore, we generalize such previous work to analyze the effect of test collections, evaluation metrics, and pool depth. We also expand the idea to semi-automatic evaluation and estimation of topic difficulty. Our results show that (i) previous work is overall reproducible, although some specific results are not; (ii) collection, metric, and pool depth impact the automatic evaluation of systems, which is anyway accurate in several cases; (iii) semi-automatic evaluation is an effective methodology; and (iv) automatic evaluation can (to some extent) be used to predict topic difficulty.

CCS Concepts: • **Information systems** → **Test collections**;

Additional Key Words and Phrases: Test collections, relevance judgments, reproducibility, topic difficulty, few topics, automatic retrieval evaluation

**ACM Reference format:**
Kevin Roitero, Marco Passon, Giuseppe Serra, and Stefano Mizzaro. 2018. Reproduce. Generalize. Extend. On Information Retrieval Evaluation without Relevance Judgments. *J. Data and Information Quality* 10, 3, Article 11 (September 2018), 32 pages.
https://doi.org/10.1145/3241064



## 1 INTRODUCTION

The evaluation of Information Retrieval (IR) systems by means of test collections allows researchers to evaluate, develop, and compare different retrieval systems or algorithms in a well-defined experimental setting. The test collection methodology, defined by Cleverdon with the creation of the Cranfield collections in the early 1960s (Cleverdon 1991), is currently carried on by many initiatives, like TREC, NTCIR, FIRE, CLEF, and so on, which model the evaluation as a competition. A typical test collection is composed by:

—a collection of documents;
—a set of description of information needs (often called *topics*);

Authors' addresses: K. Roitero, M. Passon, G. Serra, and S. Mizzaro, University of Udine, via delle Scienze 206, Udine, 33100, ITA; emails: {roitero.kevin, passon.marco}@spes.uniud.it, {giuseppe.serra, mizzaro}@uniud.it.

1936-1955/2018/09-ART11 $15.00
https://doi.org/10.1145/3241064

—a set of relevance judgments for a subset of topic-document pairs, made by experts and taken as a ground truth (usually called *qrels*).

To build a test collection is an expensive process: the TREC initiative required in 20 years more than 30 million dollars (Tassey et al. 2010). The most expensive part of building a test collection is to produce, for every topic, the relevance assessment for the documents retrieved by the retrieval systems participating in the competition. To reduce the effort of this process, it is common practice to pool a subset of the top 1,000 documents retrieved by each system; the relevance assessment is then performed only for the pooled documents (Voorhees and Harman 2000). The pooling method leads to reliable results in evaluating the effectiveness of retrieval systems (Zobel 1998), but even with the pooling strategy the cost required to produce the relevance judgments is still high.

Many researchers tried to reduce the effort of producing relevance assessments in several different ways. For example, Lu et al. (2016) and Alonso and Mizzaro (2009) proposed to crowd-source relevance judgments, Lipani et al. (2016) and Losada et al. (2017) developed novel pooling strategies to build test collections with a reduced number of judgments, Guiver et al. (2009), Robertson (2011), and Berto et al. (2013) studied the evaluation of IR systems using fewer topics, and many others tried to propose more sensitive and reliable evaluation metrics (Yilmaz and Aslam 2006).

A perhaps more extreme approach is to produce automatic relevance assessment, i.e., to evaluate the systems participating in a test collection initiative without any relevance judgments, in a completely automatic way (Soboroff et al. 2001; Spoerri 2007; Wu and Crestani 2003). In this article, we focus on this approach, and we pursue the threefold aim of reproducing the main previous results (Aim A1), generalize them to other collections, metrics, and pool depth (A2), and expand the approach to derive some insights on related problems not studied yet (A3). More in detail, our aims can be stated as follows:

- **A1.** To reproduce the main results of the notable works on automatic evaluation of retrieval systems, as well as present such results in a uniform way.
- **A2.** To generalize such previous work, in particular:
  - **A2a.** To analyze the effect of using further test collections, featuring different properties from those used in the original experiments;
  - **A2b.** To study the effect of different evaluation metrics; and
  - **A2c.** To study the effect of a shallow pool.
- **A3.** To expand the idea; in detail:
  - **A3a.** To experiment with a mixed approach, in which a part of the evaluation is automatic and a part of it is manual; and
  - **A3b.** To apply the same approach to a dual problem, the estimation of topic difficulty.[1] Both these ideas have not yet been explored, although they do seem quite natural in this context.

This article is structured as follows. In Section 2, we detail the background of evaluating IR systems without relevance judgments. In Section 3, we describe data, methods, and measures used in our experiments. In Section 4, we focus on A1 and reproduce some of the most important work on evaluating IR systems without relevance judgments. In Section 5, we turn to A2 and generalize some of the obtained results to other collections, including a more recent one, evaluation metrics, and a shallow pool. In Section 6, we address A3 by expanding the approach; we report results on semi-automatic approaches and on topic difficulty. Finally, in Section 7 we conclude and provide some directions for future work.

[1]The duality of topic difficulty and system effectiveness will be discussed in detail in Section 6.2.

## 2 BACKGROUND

In this section, we discuss the state-of-the-art of ranking and evaluating IR systems without relevance judgments. We first discuss the three main contributions that can be found in the literature and that will be the focus of this article (Sections 2.1, 2.2, and 2.3); then we also briefly comment on some related work (Section 2.4).

### 2.1 The Method by Soboroff et al.

The approach proposed by Soboroff et al. (2001) is the first work investigating the ranking of retrieval systems without human assessments. With almost 100 citations,[2] this paper is considered by the research community a strong baseline in this context.

Soboroff et al. start questioning what happens if relevant documents are chosen randomly from the pool, considering the hypothesis that relevant documents occur in the pool according to a defined probability distribution. To address this question, they design an experiment in which they estimate a probabilistic model that describes the occurrence of the relevant documents in the pool. Specifically, they choose to model relevant document occurrence with a normal distribution $\mathcal{N}(\mu, \sigma)$ that requires only two parameters to be estimated from queries, namely the mean percentage of relevant documents occurring in the pool $\mu$ and the standard deviation $\sigma$:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\mu_i,$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\mu_i - \mu)^2}{n-1}},$$

where $n$ is the number of topics of the test collection and $\mu_i$ is the percentage of relevant documents occurring in the pool for the $i$th topic.

Using this model, Soboroff et al. randomly sample a set of documents and labeled them as "relevant" to form a set of pseudo relevance judgments called *pseudo-qrels*, in three ways:

- —sampling the documents from the official qrels using $\mu$ and $\sigma$;
- —sampling the documents from the official qrels considering each topic separately, i.e., using $\mu_1, \ldots, \mu_n$ and $\sigma$. Soboroff et al. named this strategy "exact–fraction sampling";
- —sampling the documents from the pool at depth 100, including duplicate documents and using $\mu$ and $\sigma$. The rationale of this experiment is that the higher the number of systems that retrieve a document, the more it is likely to be relevant.

We focus on the first approach since it is the more realistic and because it does not include *a posteriori* knowledge (i.e., knowledge that can be obtained only after the human relevance assessments have been gathered), except for the mean and standard deviation parameters; furthermore, we provide an experiment in which we estimate the $\mu$ and $\sigma$ parameters to present a realistic "without relevance judgments" scenario.

Using the *pseudo-qrels*, Soboroff et al. present experimental evaluation on all the runs of the TREC-3, TREC-5, TREC-6, TREC-7, and TREC-8 collections. The effectiveness of the method is measured by (i) computing, on each topic, the Kendall's $\tau$ correlation between the ground truth of the official rank of the systems and the rank obtained on the basis of the pseudo-qrels and (ii) taking the mean $\tau$ over all topics. The highest mean $\tau$ value is of 0.487, obtained for the TREC-5 collection (more details on these results will be shown in Table 4 in Section 4).

[2]Source: http://dl.acm.org/citation.cfm?id=383961; April 21, 2018.

Although this approach achieves a reasonable performance in terms of correlation, the method fails mostly on top-ranked systems (i.e., the most interesting ones for the evaluation process), whose effectiveness is usually heavily underestimated.

Based on the work by Soboroff et al., Aslam and Savell (2003) propose a strategy to infer the similarity of retrieval systems by assessing the similarity of their retrieved documents. Considering two retrieval systems and their ranked lists, the measure is simply the ratio between the number of documents that they have in common divided by the total number of retrieved documents. Although the proposed measure is trivial and easy to compute, the authors show that it is able to achieve correlation values with the ground truth similar to those by Soboroff et al. They also observe that both methods are affected by a "tyranny of the masses" phenomenon: top-ranked systems (i.e., the systems that lower the correlation in Soboroff et al.'s experiments) are being punished for retrieving documents significantly different from the average systems in a competition. Therefore, Aslam and Savell observe that Soboroff et al.'s method assesses the retrieval systems more in terms of "popularity" than actual "performance."

This effect has been also investigated by Spoerri (2005), who remarks that "the potential relevance of a document increases exponentially as the number of systems retrieving it increases," calling it the Authority Effect. Spoerri also suggests that selecting only a single run per participant group[3] "would help to sharpen the signal and make the Authority Effect more dominant" (Spoerri 2007, page 1061).

### 2.2 The Method by Wu and Crestani

The work proposed by Wu and Crestani (2003) presents another approach for ranking retrieval systems without relevance judgment. This technique uses a measure called reference count, which is developed by the same authors within the data fusion context (Wu and Crestani 2002).

Specifically, suppose we have a topic and a set of retrieval system results on the same data collection, the reference count of a retrieval system result can be obtained as follows. Given the set of retrieved documents by a system (called *original documents*), the reference count is the sum of the occurrences of these documents in the results of all the other retrieval systems (called *reference documents*) up to number of retrieved document per topic, which is usually 1,000.

This approach is called by the authors Basic reference count. However, this technique does not take into account the different position of *reference documents* and the position of the *original document*. To overcome these limits, Wu and Crestani present four different variations by changing either or both of the aspects. The first variation (V1) assigns different weighs to *reference documents* based on their ranking positions. The second variation (V2) assigns different weights to *original documents* based on their ranking positions. The third variation (V3) consists of assigning different weights to both the *reference documents* and the *original document*. Finally, the fourth variation (V4) assigns different weights based on the *reference documents'* ranking positions and the *original document*'s normalized scores (instead of their ranking positions).

Wu and Crestani present their results using Spearman's average $r_S$ correlation values over the topics (the detail of their values over different TREC collections will be shown in Table 5, discussed in the following). The values are not directly comparable to Soboroff et al.'s, who use $\tau$; however, Wu and Crestani compute $r_S$ also for Soboroff et al.'s method and find that they obtain average $r_S$ correlation values that are lower than, or comparable with, Soboroff et al.'s. Wu and Crestani observe that the results are mostly based on two effects: (i) the overlap of the relevant and nonrelevant documents retrieved by the systems is quite different, as also shown by Lee (1997); (ii) there

[3]Initiatives like TREC allow participant groups to submit the results obtained with different variations of their system. One of them is called "run."

is a connection between reference count and the percentage of relevant documents in a ranked list of a system. Furthermore, they observe that their results are affected by the same problem of the top-ranked systems as in Soboroff et al. Again, this phenomenon is probably due to the fact that top-ranked systems are quite peculiar; in fact, they retrieve documents that not many other systems retrieve.

### 2.3 The Method by Spoerri

Spoerri (2007) proposes another method to rank retrieval systems without human relevance judgments. The proposed approach starts from the fact that retrieval systems tend to retrieve similar sets of relevant documents and dissimilar sets of non-relevant documents (Lee 1997). Spoerri's approach estimates the relative performance of multiple retrieval systems computing the *structure of overlap* between their retrieved documents. Specifically, for each system he counts the number of documents retrieved by it, which are also retrieved by a specific number of other systems; in Spoerri's method implementation this number is five. This process is called *random grouping*. More in detail, the "structure of overlap" is used to extrapolate two measures: Single% (S% in the following), the percentage of a system's documents not found by other systems, and AllFive% (A%), the percentage of a system's documents found by all five systems. A third measure, Single%−AllFive% (S-A%), is also computed as the arithmetic difference between the two previously found percentages. In his experiments, Spoerri builds the structure of overlap both when considering the top 1,000 documents retrieved by each system and a shallow pool, to study the effect of pool depth on the effectiveness of his method.

Analyzing these evaluation measures, the author shows that the percentage of a system's documents that are only found by it and not by other systems (S%) increases as the system retrieval quality decreases. For the structure of overlap to be computed, it is critical that a single run for each system participating in the track is included. In fact, although runs of a system participating in a track are different, they usually share the same technique and system architecture. Therefore, the structure of overlap could be affected by this dependence. Furthermore, the author in the article demonstrates that there exists an optimal number of retrieval systems and topics needed for building the structure of overlap.

In the experimental evaluation, Spoerri selects a subset of the runs submitted to the selected TREC tracks. Spoerri does not use multiple runs of a system, because these would have very strong structure of overlap compromising the comparison with the other selected systems. Therefore, for each participant group he selects one run, called *short-run*, according to the following criterion: the short-run is selected considering the one with the highest AP value among the runs that use automatic as Query Method and Title+Description as Topic Length (i.e., runs that did not use the Narrative field of the topic). Although the selection of the most effective short-run for each group allowed to run some experiments to analyze the effectiveness of his approach, it must be noticed that this kind of selection would be impossible in a real scenario without a pre-evaluation process.

Spoerri's approach achieves Spearman $r_S$ correlation values up to 0.96. However, this value is not comparable to Wu and Crestani's, since it is not computed by taking the mean $r_S$ over topics, but as the $r_S$ of predicted and real MAP values. The first and third variants (i.e., S% and S-A%) show the best effectiveness (details will be shown in Table 6). The results are also confirmed by using other metrics (i.e., P@100 and R@100).

### 2.4 Other Notable Examples of Evaluation without Relevance Judgments

Other proposals of studies about effectiveness evaluation without relevance judgments exist: for example, Aslam and Savell (2003)'s work can be exploited, and Nuray and Can (2003) is another viable alternative. However, in the rest of this article, we will focus on the three above-described

methods, which are somehow the most classical and well-known approaches. For the sake of completeness, we briefly describe some other studies that, rather than proposing a novel technique for ranking IR systems without relevance judgments, are more focused on analyzing causes and consequences of an automatic evaluation. Nuray and Can (2003) propose a new automatic evaluation technique that estimates the relative ranking of retrieval systems by computing a "popularity score" for each retrieved document, and then using trec_eval to compute a certain metric, and compare it to the human-based rankings. Experimental evaluation is performed on a web-like imperfect environment, in which all the documents are available at indexing time, but some of them cannot be accessible at retrieval time, e.g., because of server down issues. Yilmaz and Aslam (2006) propose new evaluation measures, which estimate the average precision (AP) with incomplete and imperfect relevance judgments. Hauff and de Jong (2010) compare Soboroff et al.'s approach directly to two evaluation methods that rely on incomplete manual relevance assessments: *statAP* and *MTC*. Experiments on the TREC's Million Query Tracks show that Soboroff et al.'s technique is able to achieve high correlation even in this scenario. Carterette and Soboroff (2010) study the effect and the robustness of the TREC Million Query track methods when some assessors make significant and systematic errors; their findings indicate that errors can have a large effect on the ranking of systems.

## 3 EXPERIMENTAL SETTING: DATA, METHODS, AND MEASURES

To pursue the aims of this article, namely, to reproduce and then to generalize as well as expand the above presented results, we run a battery of experiments. In this section, we describe the common features of those experiments, and in the rest of the article we will provide further details when needed, and the results.

### 3.1 Data

For reproducibility purposes, we use the datasets from the TREC[4] editions that have been used by at least one of the previous studies: the Ad Hoc tracks of TREC-3, TREC-5, TREC-6, TREC-7, TREC-8, and TREC-2001. Furthermore, to extend and generalize such results, we use some more datasets: besides again TREC-8, also, from other TREC editions, the Robust track of 2004 (R04), the TeraByte track of 2006 (TB06), and the Ad Hoc Web track of 2014 (WEB14). All the test collections used in this article are detailed in Table 1.

Concerning R04, we use 249 topics, removing topic 672, as described by Voorhees (2004, Section 1): "the TREC 2004 track used a set of 250 topics (one of which was subsequently dropped due to having no relevant documents)."

Concerning the TeraByte collection, we choose to investigate two versions of the original dataset (see Büttcher et al. (2006, Section 3.1): "Manual runs used only the 50 new topics; automatic runs used all 149 topics from 2004-2006"):

- —We select the subset of 61 runs that run over all 149 topics, as done by Roitero et al. (2017); this collection has no manual runs; we denote this dataset as TB06. Specifically, the dataset uses the 150 track topics and removes topic 703.
- —We consider all the participating runs (i.e., 80 runs, including the manual ones) that run over a common subset of 50 topics; we call this collection TB06M (M denotes the inclusion of the manual runs).

Concerning WEB14, we choose two different versions of the track to investigate the effect of the evaluation measures: this approach is detailed in Section 3.2.

For all the other datasets, we use the standard settings proposed in the track.

[4]See http://trec.nist.gov/.

Table 1. Test Collections Used for the Reproducibility Experiments in the Upper Part of the Table, and for the Other Experiments in the Lower Part

| Acronym | Track | Year | Topics | Runs | Used Topics | Manual Runs |
|---|---|---|---|---|---|---|
| TREC-3 | Ad Hoc | 1994 | 50 | 40 | 151-200 | 11 |
| TREC-5 | Ad Hoc | 1996 | 50 | 61 | 251-300 | 31 |
| TREC-6 | Ad Hoc | 1997 | 50 | 74 | 301-350 | 17 |
| TREC-7 | Ad Hoc | 1998 | 50 | 103 | 351-400 | 17 |
| TREC-8 | Ad Hoc | 1999 | 50 | 129 | 401-450 | 13 |
| TREC-01 | Ad Hoc | 2001 | 50 | 97 | 501-550 | 2 |
| TREC-8 | Ad Hoc | 1999 | 50 | 129 | 401-450 | 13 |
| TB06 | TeraByte | 2006 | 149 | 61 | 701-850† | 0 |
| TB06M | TeraByte | 2006 | 50 | 80 | 801-850 | 19 |
| R04 | Robust | 2004 | 249 | 110 | 301-450, 601-700† | 0 |
| WEB14 | Web | 2014 | 50 | 30 | 251-300 | 4 |

†Not all, see text.

Table 2. AP, MAP, and AAP for $n$ Topics and $m$ Systems (Adapted from Mizzaro and Robertson (2007))

| | $t_1$ | $\cdots$ | $t_n$ | MAP |
|---|---|---|---|---|
| $s_1$ | $\mathrm{AP}(s_1, t_1)$ | $\cdots$ | $\mathrm{AP}(s_1, t_n)$ | $\mathrm{MAP}(s_1)$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $s_m$ | $\mathrm{AP}(s_m, t_1)$ | $\cdots$ | $\mathrm{AP}(s_m, t_n)$ | $\mathrm{MAP}(s_m)$ |
| AAP | $\mathrm{AAP}(t_1)$ | $\cdots$ | $\mathrm{AAP}(t_n)$ | |

### 3.2 Evaluation Measures

The outcome of the TREC evaluation process can be represented as in Table 2: $s_i$ represents a system/run, $t_j$ represents a topic, $\mathrm{AP}(s_i, t_j)$ represents the effectiveness of the system $s_i$ on the topic $t_j$ according to an evaluation measure. AP is perhaps the most widely used metric; however, system effectiveness can be expressed by means of many other alternative metrics, like logAP, logitAP, NDCG, and so on. Since systems in TREC usually are required to retrieve 1,000 documents for each topic, we use the truncated versions of these metrics, e.g., we use AP @1000. To rank retrieval systems, a common approach is to average the performance over the set of topics according to a measure (e.g., MAP = Mean AP); thus,

$$\mathrm{MAP}(s_i) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{AP}(s_i, t_j). \tag{1}$$

More in detail, for TREC-8, R04, TB06, and TB06M we use as evaluation measure AP@1000; the official track measure for TB06 is AP@10.000 (which is very close to AP). We also present results in terms of GMAP, and logitAP (metrics detailed in Section 5.2). For WEB14 we use the official track measure NDCG. To present AP values for this dataset as well, we binarize WEB14 *qrels*. As it is usually done, we attempt two slightly different binarizations: in the first version we map the original relevance values −2 and 0 into not relevant and the values 1, 2, 3 into relevant; in the second version, we map the values −2, 0, and 1 into not relevant and the values 2 and 3 into relevant. In the following, we focus on the first binarization only, which incidentally provides better results in terms of the final correlation obtained. Note that selecting the binarization, which

Table 3. $\mu$ and $\sigma$ Values: Comparison for Reproducibility (Leftmost Four Columns) and Estimation (Two Rightmost Columns)

| | SNC original | | Our obtained values | | Estimated | |
|---|---|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| TREC-3 | 14.90 | .123 | 10.414 (14.902)[†] | .097 (.123)[†] | 23.15 | .110 |
| TREC-5 | 3.90 | .043 | 3.956 | .043 | 13.39 | .074 |
| TREC-6 | 6.32 | .067 | 6.351 | .067 | 10.13 | .062 |
| TREC-7 | 5.78 | .047 | 5.834 | .047 | 5.82 | .046 |
| TREC-8 | 5.35 | .048 | 5.497 | .048 | 3.60 | .038 |

[†]Pool built with all the participating runs at depth 100.

leads to higher correlation values, has no consequence in the experiment results since we are not competing against any baseline. The meaning of AAP (see the last row in Table 2) will be discussed in Section 6.2.

### 3.3 Methods Configuration

For brevity, in the following we denote the methods by Soboroff et al. (2001), Wu and Crestani (2003), and Spoerri (2007) with SNC, WUC, and SPO, respectively.

Considering SNC, we start estimating a probability distribution by randomly selecting relevant documents from *qrels* for building *pseudo-qrels*. First, based on the official *qrels* we compute the mean ($\mu$) and the standard deviation ($\sigma$) values of the percentage of relevant documents in the pool. Table 3 shows on the left side the comparison between SNC and our computed values of $\mu$ and $\sigma$ (the rightmost column is discussed in the following). Comparing these values with Soboroff et al.'s, we observe that we are able to reproduce the same $\mu$ and $\sigma$ values, apart from TREC-3. Our hypotheses is that, in this case, all participating runs have been used, in place of the official runs only (i.e., the ones that are selected to form the pool),[5] and indeed, when using this approach, we obtain the values in parentheses in the first row of the table, which perfectly match the original ones. For completeness and for (future) reproducibility, we also report the $\mu$ (and $\sigma$) values that we obtain for R04, TB06, TB06M, and WEB14: 5.12 (0.043), 13.39 (0.074), 8.98 (0.057), and 32.59 (0.145).

Once we have the estimated normal distribution (based on the mean and the standard deviation values computed before) we can build the *pseudo-qrels* by simply performing a random sampling on *qrels* based on this distribution; then using the official "trec_eval" software (version 9.0)[6] we compute an approximated AP (i.e., obtained with the sampled *qrels*) value for each run and each topic. Based on these values, we compared the approximated AP and MAP values (i.e., the ones originated from the *pseudo-qrels*) with the real AP and MAP values. To provide a realistic (i.e., without any post-evaluation knowledge) setting for Soboroff et al.'s work, we also estimate $\mu$ and $\sigma$ values using a best-fit interpolation with an order one polynomial trend-line, obtaining

$$\begin{aligned} \mu &= \frac{1133.3}{\text{no. runs}} - 5.1841, \\ \sigma &= 0.0037\mu + 0.0242. \end{aligned} \tag{2}$$

The two rightmost columns in Table 3 show the estimated values.

For WUC, we start by computing the so-called document "reference count" for each topic, run, and position of the rank; then we sum and normalize the reference count to compute the Basic,

[5]Note that a detailed list of the official runs is not provided by NIST.
[6]http://trec.nist.gov/trec_eval/.

V1, V2, V3, and V4 measures according to Wu and Crestani's definition. We consider all the runs submitted to the TREC tracks.

For SPO, we start by selecting the systems according to the selection method described in Section 2.3. Having the subset of systems, we compute the structure of overlap for all the runs; we compare then the percentages of overlap given by the structure of overlap with the real MAP values. As stated before, the structure of overlap is built forming random groupings of five retrieval systems and this structure is used to extrapolate S%, A%, and S-A% measures. Based on our experiments, we observe that by following the proposed selection method (see also Section 2.3) the Title+Description runs are often not enough to reach the number of runs selected by Spoerri. Most likely, Spoerri then included in his experiments some runs that have only Description as Topic Length; we follow this approach.[7]

To avoid noise and give stability to our results, we performed 20 repetitions for SNC and SPO, which have a non-deterministic part. In the following, we report the results obtained when averaging the AP and correlation values over the 20 repetitions.

### 3.4 Correlation Measures

In this article, we focus on reporting correlation values between the official system rank provided by TREC and the system rank obtained by the automatic evaluation methods. In the three methods that we reproduce, there is no homogeneity concerning the correlation coefficient used to compare the computed scores (representing the automatic evaluation of systems) with the real evaluation of systems (e.g., the real MAP values). Thus, to present the results in a homogeneous way, we report the correlation values using value- and rank-based correlations, as well as top-heavy correlation measures. More in detail, we use:

—Pearson's $\rho$, which measures linear correlation;
—Kendall's $\tau$, which measures rank correlation;
—Spearman's $r_S$, which measures rank correlation;
—Rank Biased Overlap (RBO) (Webber et al. 2010), a parametric rank correlation measure that is top-heavy, i.e., weights more the first positions of the rank. The rationale is that usually it is more important to correctly estimate the effectiveness of the top-ranked systems, i.e, the most effective ones. We choose to give the top 10% of the systems the 75% of the weight evaluation; thus we estimate the parameter $p$ for RBO as detailed by Webber et al. (2010, Eq. 21);
—AP *correlation* ($\tau_{AP}$), proposed by Yilmaz et al. (2008), a top-heavy rank correlation coefficient based on the AP measure.

We also present the results by means of scatterplots, which allow us to compare the real effectiveness with the effectiveness obtained by the three methods.

It has to be noted that while the SNC method actually predicts an effectiveness (e.g., MAP) value, the other two methods provide values that have a different meaning: SPO returns a percentage (a number between 0 and 100) that represents the structure of overlap between systems and WUC produces a (normalized) reference count value (a value between 0 and $+\infty$). Thus, we normalize SPO and WUC scores in three ways, as discussed in Section 5.2.1. Furthermore, we change the sign of the value returned by SPO method, to obtain positive correlations and to easily compare the results with SNC and WUC.

[7]To make our run selection process reproducible, we report the runs that we selected in the spreadsheet available at https://users.dimi.uniud.it/~kevin.roitero/OUTSIDE/Reproducibility_SI_EvalNoJudg (where we also include all the code used to carry out our experiments and some additional tables).

Table 4. Mean and Standard Deviation of $\tau$ Values: Comparison for Reproducibility

| | Original SNC | | Our obtained values† | | Orig. SNC dups | | Our obtained dups† ‡ | |
|---|---|---|---|---|---|---|---|---|
| | avg $\tau$ | std $\tau$ | avg $\tau$ | std $\tau$ | avg $\tau$ | std $\tau$ | avg $\tau$ | std $\tau$ |
| TREC-3 | .430 | .0312 | .401 (.411) | .0259 (.0276) | .482 | .0143 | .487 (.471) | .0113 (.0111) |
| TREC-5 | .487 | .0462 | .359 (.382) | .0766 (.0279) | .571 | .0107 | .421 (.409) | .0067 (.0043) |
| TREC-6 | .408 | .0354 | .391 (.387) | .0370 (.0299) | .491 | .0131 | .458 (.452) | .0052 (.0045) |
| TREC-7 | .369 | .0363 | .377 (.379) | .0474 (.0470) | .423 | .0091 | .446 (.446) | .0046 (.0046) |
| TREC-8 | .459 | .0340 | .460 (.444) | .0402 (.0567) | .543 | .0102 | .533 (.538) | .0043 (.0052) |

Original values from Soboroff et al. (2001, Tables 2 and 3) (first and third pair of columns) and our obtained values (second and fourth pairs of columns).
†The values in parentheses are those obtained with estimated $\mu$ and $\sigma$ (see formula Equation (2) and Table 3).
‡Pool built with all the participating runs at depth 100, using the duplicates.

Table 5. Average $r_S$ Correlation over the Topics of Each Collection: Comparison for Reproducibility

| | Original WUC | | | | | | Our obtained values | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Basic | V1 | V2 | V3 | V4 | RS† | Basic | V1 | V2 | V3 | V4 | RS† |
| TREC-3 | .246 | .248 | .548 | .567 | .587 | .628 | .513 | .522 | .512 | .504 | .283 | .629 |
| TREC-5 | .318 | .326 | .378 | .421 | .421 | .430 | .393 | .405 | .395 | .400 | .328 | .476 |
| TREC-6 | .309 | .316 | .371 | .383 | .384 | .436 | .442 | .451 | .485 | .497 | .498 | .522 |
| TREC-7 | .297 | .304 | .328 | .345 | .382 | .411 | .406 | .419 | .403 | .421 | .453 | .501 |
| TREC-01 | .279 | .288 | .377 | .401 | .413 | .463 | .449 | .460 | .448 | .459 | .443 | .571 |

Original values from Wu and Crestani (2003, Table 1) (left) and our obtained values (right).
†SNC, all runs, considering the duplicates and selecting randomly the 10% of the documents as relevant.

## 4 A1: REPRODUCE

We now turn to our first (and maybe most important) aim A1, namely to reproduce the results previously published in the literature.

### 4.1 Results

We first study the reproducibility of each of the three methods, and then address their comparison.

*4.1.1 SNC.* Table 4 compares the SNC original values of mean $\tau$ and standard deviation of $\tau$ with those that we have obtain when reproducing such method. The left side of the table shows the values obtained considering the pool without duplicates and right side of the table shows the values when considering duplicates. Remember that we compute the mean $\tau$ score obtained over 20 repetitions. As we can see from the table, apart from TREC-5, we obtain mean $\tau$ scores comparable to the ones of SNC, both when duplicates are considered and when they are not. We conjecture that the differences in the correlation values are caused by the possibility of selecting different runs to reproduce the original TREC pool.

*4.1.2 WUC.* Table 5 shows a similar comparison for the WUC method. Here the comparison is based on the average $r_S$ correlation values over the topics of each collection. This is of course different from SNC: we now focus on reproducing the results, thus we use the same correlation coefficients used in the original articles; we will provide a homogeneous comparison later. As we can see from the table, for almost all WUC variants (i.e., Basic, V1, V2, and V3) we obtain higher correlation values; concerning WUC V4, we obtain higher correlation values in three cases out of five. In the original article, the highest correlation value is always obtained with the variant V4; instead, we find that the best variant depends on the collection. Furthermore, in the original

article, the four variants have increasing correlation values (i.e., for each collection, we always have Basic < V1 < V2 < V3 < V4); on the contrary, the correlation values that we obtain are usually comparable across the different WUC variants.

We conjecture that the differences between our values and the original ones might be due to:

— A different normalization score, especially for V4 (i.e., the most effective version in the original article). Regarding V4, Wu and Crestani (2003, page 813) state: "V4 consists of using each document's normalized score," but a detailed explanation of the normalization process is not provided; for this reason we use each system "retrieval status values" (RSV) to compute each document normalized score. We also consider the rank of the document with similar (or worse) outcomes. Note that many different normalization formulas could have been used by original authors; we tried different variants: (i) normalize the RSV/rank in [0, 1], (ii) normalize the RSV/rank using the standard score $(x-\mu)/\sigma$), and (iii) the above normalization using WUC Basic, V1, V2, or V3 in place of the RSV/rank. We find that none of them leads to a successful reproducibility of original results. Furthermore, we find it surprising to not be able to reproduce WUC Basic version, since it does not require any normalization of the reference count score.

— A different approach used to compute correlations. Wu and Crestani (2003, page 813) state: "In Table 1 we present the results when all runs of participant systems are taken. Each item in the table is the mean Spearman rank correlation coefficient over 50 topics of that year." If we have a collection with $n$ topics, we denote with $\mathrm{AP}(t_i)$ the vector of official AP for all the runs of topic $i$ (i.e., a column of Table 2), and we use the subscripts $o$ and $w$ to differentiate, respectively, the official and the WUC measure, then Table 5 shows

$$\frac{1}{n}\sum_{i=1}^{n} r_S\left(\mathrm{AP}_o(t_i), \mathrm{AP}_w(t_i)\right).$$

We try different variations, namely:

$$\text{(i)}\ \frac{1}{n}\sum_{i=1}^{n} r_S\left(\mathrm{AP}_o(t_i), \mathrm{MAP}_w\right), \text{(ii)}\ \frac{1}{n}\sum_{i=1}^{n} r_S\left(\mathrm{MAP}_o, \mathrm{AP}_w(t_i)\right), \text{and (iii)}\ r_S\left(\mathrm{MAP}_o, \mathrm{MAP}_w\right).$$

None of them lead to an effective reproducibility. We also use our own implementation of Spearman rank $r_S$ correlation, which reflects the formula detailed by Wu and Crestani (2003, Eq. 4); also in this case the results do not vary from the previous ones, obtained using the official Python 3 *scipy.stats.spearmanr*[8] implementation.

— The number of runs used to compute the $r_S$ correlation score. Although Wu and Crestani (2003, Tab. 1) state: "Mean correlation coefficient for AP, all systems submitted to TREC," we try with a selection of systems. Also this attempt results in a failed reproducibility of the original scores.

We remark that it is unlikely that by doing something wrong we obtain higher correlation values than the original ones; one should expect exactly the opposite. We release the code used to compute WUC scores (see Footnote 7), which could be useful for future reproducibility, and for a correct implementation of the method described by Wu and Crestani.

Wu and Crestani also reported, for comparison, the $r_S$ correlation values that they obtained when reimplementing SNC (in a slightly different version, i.e., considering the duplicates and selecting randomly the 10% of the documents as relevant). Table 5 shows, in the two RS columns, the

[8] https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html.

Table 6. Average $r_S$ Correlation: Comparison for Reproducibility

| | Orig. SPO | | | Our obtained values | | | Orig. SPO top 50 | | | Our obtained top 50 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S% | A% | S-A% | S% | A% | S-A% | S% | A% | S-A% | S% | A% | S-A% |
| TREC-3 | .67 | – | .72 | .67 | .51 | .69 | .72 | – | .72 | .46 | .25 | .44 |
| TREC-6 | .79 | – | .78 | .69 | .40 | .68 | .86 | – | .88 | .80 | .44 | .79 |
| TREC-7 | .43 | – | .47 | .59 | .38 | .57 | .70 | – | .62 | .67 | .35 | .63 |
| TREC-8 | .89 | .88 | .95 | .79 | .48 | .76 | .92 | – | .95 | .82 | .55 | .80 |

Original values from Spoerri (2007, Table 1) (first and third groups of three columns) and our obtained values (second and fourth groups of three columns). The table shows on the left side the values obtained when using all 1,000 retrieved documents, and on the right side when using the top 50 ones. S%, A%, and S-A% values computed using one short-run per participant group (the one with highest MAP score). TREC-3, TREC-6, and TREC-7 correlations are estimated from Spoerri (2007, Fig. 5).

$r_S$ correlation values obtained by Wu and Crestani and by ourselves when we reimplement SNC following Wu and Crestani variation. We see that, also in this case, the values are in general different, apart from TREC-3 and TREC-5. Referring to Table 4, recall that our $\tau$ correlation values are similar to SNC original ones. This might depend, again, on one of the justifications detailed above.

*4.1.3 SPO.* Table 6 shows the SPO S%, A%, and S-A% scores obtained by Spoerri and by ourselves when considering one run (the short-run) per participant group; the table shows correlation values (averaged over the $N$ random grouping of $N$ short-runs) when considering all 1,000 retrieved documents (left) and only the top 50 retrieved (right). For completeness, the table also contains all our A% values, not always reported in the original article. We see that our computed values, for both top 50 and top 1,000, are comparable to Spoerri's, even if they are usually lower; there are two exceptions though: TREC-7 top 1,000 and TREC-3 top 50.

The differences between the correlation values might depend on two different factors: (i) the division of systems into the sets containing five systems is not trivial, and the implemented algorithm can affect the final result, and (ii) the selection of which run to include from each participating group has a major impact on the correlation values, and a list of the used runs is not available in the original article (we make available our list in the URL provided in Footnote 7).

Another important remark is that in the original work the version with the highest correlation value is S-A% for almost all collections, but according to our finding this is not always the case, especially when considering the top-50 rank positions.

*4.1.4 Comparison of SNC, WUC, and SPO.* As we have already noticed in Section 2, and as it is clear from the last three tables, the correlation values reported in the original articles are not directly comparable. To be able to compare in a more systematic and convenient way the three methods, in Table 7 we show the $\tau$ correlation obtained by us when reproducing the methods considering all the runs participating in the track and the top 1,000 documents retrieved for each topic. As we can see from the table, SNC dups always achieves the highest correlation values. This might be surprising, since it is the most trivial method.

Figure 1 shows some selected results as scatter plots: the $x$-axis shows the real MAP value, while the $y$-axis shows the score of the SNC, SPO, and WUC methods; each dot is a system; and automatic and manual runs are graphically different. In the plots, we also display the regression line as well as $\rho$, $\tau$, $r_S$, RBO, and $\tau_{AP}$ measures. Note that the correlation values are minimally different from the ones in Table 7; the table shows the average correlation obtained over the 20 repetitions, the figure shows instead the plot and the corresponding correlation values where each (M)AP is the average (M)AP obtained over the 20 repetitions.

Table 7. MAP Comparison between All State-of-the-Art Used Collections and All Methods Using $\tau$ Correlation

| | SNC | | WUC | | | | | SPO | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | qrel | dups | Basic | V1 | V2 | V3 | V4 | S% | A% | S-A% |
| TREC-3 | .401 | **.486** | .405 | .405 | .436 | .444 | .344 | .416 | .216 | .400 |
| TREC-5 | .359 | **.422** | .332 | .341 | .336 | .338 | .355 | .359 | .215 | .352 |
| TREC-6 | .391 | **.459** | .422 | .427 | .451 | .448 | .453 | .421 | .288 | .421 |
| TREC-7 | .378 | **.445** | .402 | .408 | .393 | .408 | .430 | .358 | .216 | .346 |
| TREC-8 | .460$^{\dagger}$ | **.532** | .466 | .480 | .386 | .391$^{\dagger}$ | .322 | .490$^{\dagger}$ | .322 | .475 |
| TREC-01 | .473 | **.625** | .537 | .545 | .582 | .582 | .554 | .521 | .347 | .520 |

We considered all the runs for each track, and the top 1,000 retrieved documents. The highest values for each collection are in bold.
$^{\dagger}$Shown in Figure 1 (see below).

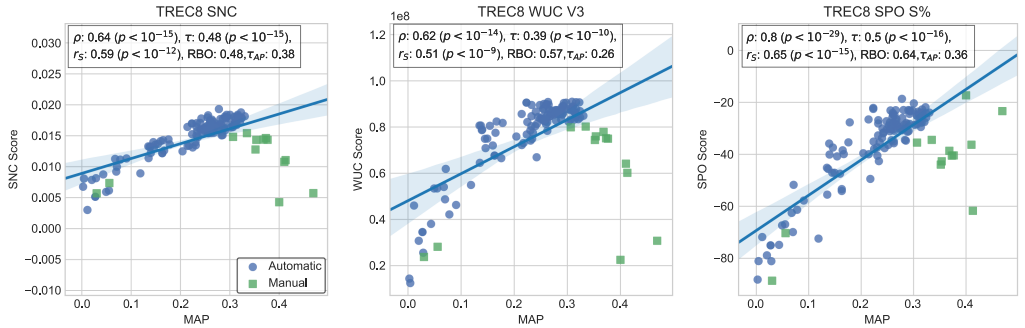



Fig. 1. Scatterplots for the three methods: SNC (left), WUC V3 (center), and SPO S% (right), on the TREC-8 collection. Each dot is a system/run. The *x*-axis shows the real MAP value; the *y*-axis shows the predicted MAP (for SNC) or a meaningful value of the specific method, i.e., a (normalized) count for WUC and a (negated) percentage for SPO.

For WUC and SPO, we selected the method variants featuring the best results, i.e., those having highest Kendall's correlation (V3 and S%, respectively). The V4 version of WUC and S-A% version of SPO have a very similar outcome. In the following, we report the WUC V3 and V4, and the SPO S%: they are almost always the best ones; when they are not the absolute best, they are very close to the best or even indistinguishable. An exhaustive appendix of our results can be found at the URL provided in Footnote 7.

Figure 1 shows that all three methods present an "inverse U" shape (less evident for SPO): all the methods fail in predicting the effectiveness of the most effective systems, which in this case (for TREC-8) are manual runs; we analyze the effect of the manual runs in Section 5.1; we can see that the methods are equally bad in predicting the top-ranked systems by the values of $\tau_{AP}$ and RBO, which are very similar for the three methods.

### 4.2 Discussion

Considering the results, we can make two main remarks:

(i) On reproducing, per se: reproducing the work is never easy, with all the methods (i.e., SNC, WUC, and SPO), for many different reasons: specific choices are often not described in the original articles, like the choice of including a system instead of another to form a subset of systems/run (see Sections 3.3 and 4.1); when the method includes a nondeterministic

process, an exact reproducibility is not possible. Furthermore, SNC is the only method which produces estimated AP values, the other ones produce a score, which can be of difficult interpretation.

(ii) On the results, the rather low correlation values, as well as the inverse U shape that can be seen in Figure 1, seem to discourage the use and further development of these approaches.

To investigate this latter remark, we generalize the three methods to other collections, other evaluation metrics, and a shallow pool, which we discuss next.

## 5 A2: GENERALIZE

We now turn to the second aim A2; more precisely, we generalize the previous work results to other test collections (Section 5.1), other evaluation metrics (Section 5.2), and a shallow pool (Section 5.3). We also briefly discuss these results (Section 5.4).

Before detailing the results, we make some remarks on generalizing the three methods: (i) SPO and WUC do not estimate AP values, thus we have to normalize the produced scores, which can be done in many different ways, as we discuss next; (ii) all WUC versions are calibrated considering the top 1,000 documents retrieved by each system, thus using other pool depths with WUC is not trivial; and (iii) SNC requires an estimation of the $\mu$ and $\sigma$ parameters: in the case of a new collection, these parameters can be estimated using different techniques, which may condition the effectiveness of this method. Furthermore, while SPO and WUC can in principle work for any relevance scale, SNC is restricted to the binary relevance scale.

### 5.1 A2a: Generalize to Other Collections

To generalize the results to other collections, we choose different TREC editions featuring a different number of systems/runs and of topics (see Table 1): TB06, TB06M, R04, and WEB14. Results on these other collections are shown in Figure 2. The charts in Figure 2 are organized by rows (each row represents a collection: TB06, TB06M, R04, and WEB14, respectively), and by columns (each column represents a different method: SNC, WUC, and SPO, respectively).

From the whole Figure 2 we notice that, on these collections, the correlations are usually higher for the SNC and WUC methods, whereas they are always lower for SPO. Furthermore, the three methods do not show a consistent behavior across various datasets: by comparing Figure 1 with Figure 2, and the rows of Figure 2 (i.e., considering different datasets but the same method), we notice that the performance of a single method is highly dependent on the particular dataset. On the contrary, the three methods are consistent within each dataset.

From the figure, we also see that the most effective systems are not penalized as they were in Figure 1: the inverse U shape of TREC-8 does not occur anymore. This is true for SNC and WUC, while SPO still penalizes the best systems, although in a less evident way. This behavior can be caused by the fact that, for TREC-8, the most effective systems are the manual runs, which are peculiar systems; in fact, the inverse U shape disappears if we imagine to remove all the manual runs from Figure 1. If we focus on the manual runs of TB06M and WEB14, their performance is still underestimated by the three methods, but, different from TREC-8, their computed score for the three methods is not similar (i.e., they do not "cluster" as in TREC-8); this can be caused by the fact that for TB06M and WEB14, the manual runs are not the most effective.

This hints that the three methods for effectiveness evaluation without relevance judgments, rather than failing in predicting the most effective systems, fail in providing a correct prediction of the effectiveness of manual systems. Also taking into account the work by Aslam and Savell (2003), this is probably due to the manual systems being "unpopular," both because they are

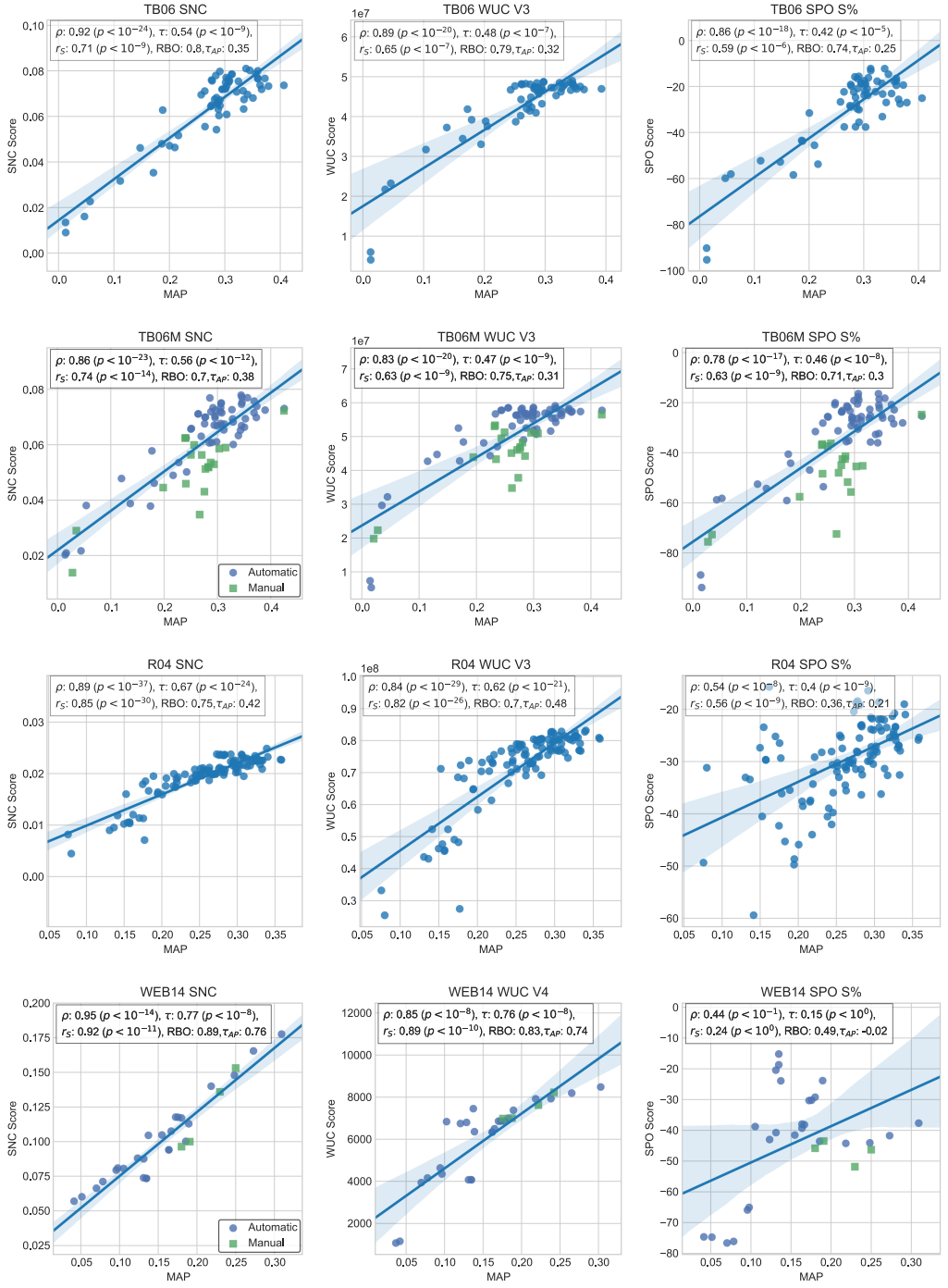



Fig. 2. Generalization to other collections. Scatterplots for the three methods SNC, WUC V3 (or V4), and SPO S% (on the first, second, and third columns, respectively) and four collections TB06, TB06M, R04, and WEB14 (on the first, second, third, and fourth rows, respectively). Compare with Figure 1.

intrinsically different from the automatic ones and because there are many more automatic than manual systems in each TREC edition (see Table 1).

### 5.2 A2b: Generalize to Other Evaluation Metrics

We now turn to Aim A2b: generalize to other evaluation metrics; more precisely, we generalize previous results to the GMAP (Section 5.2.1), logitAP (Section 5.2.2), and NDCG (Section 5.2.3) metrics.

*5.2.1 GMAP.* Geometric Mean Average Precision (GMAP), proposed by Robertson (2006), is the geometric mean of the AP values over the set of topics. Using the same notation as in Table 2 and in formula Equation (1):

$$\text{GMAP}(s_i) = \sqrt[n]{\prod_{j=1}^{n} \text{AP}(s_i, t_j)} = \exp\left(\frac{1}{n}\sum_{j=1}^{n} \ln(\text{AP}(s_i, t_j))\right), \tag{3}$$

where we use the natural logarithm (ln) and its inverse function (exp), and AP values of zero need to be replaced by a small $\epsilon$ value of $10^{-5}$ as done by Robertson. Even though the GMAP measure averages the system effectiveness over all the topics as MAP does (see formula Equation (1)), GMAP gives emphasis on the low values of the effectiveness measure (i.e., the bottom of the scale for an effectiveness measure). To better explain the concept, we report an example taken from Robertson (2006, page 81):

> GMAP treats a change in AP from 0.05 to 0.1 as having the same value as a change from 0.25 to 0.5. MAP would equate the former with a change from 0.25 to 0.3, regarding a change from 0.25 to 0.5 as five times larger.

Furthermore, as discussed by Fuhr (2017), when a researcher is interested in measuring relative changes, GMAP should be used; when absolute changes are studied, MAP should be used.

To compute the GMAP values, we tried different normalizations of the WUC and SPO scores: scores normalized in $[0, 1]$, scores normalized in $[0, 0.5]$ (since the AP scores are often in that range), scores normalized in $[\min_{i,j}(\text{AP}(s_i, t_j)), \max_{i,j}(\text{AP}(s_i, t_j))]$, and no normalization at all. To normalize the scores, first we transform the AP values into log AP scores; then, we normalize those scores per-collection; e.g., to normalize in $[0, 1]$ we consider the *min* and *max* value of the collection, (i.e., *max* and *min* values of Table 2, all runs and topics together); finally, we average the results over the set of topics. Results are all very similar, and in the following we use the first normalization.

Results of the generalization to the GMAP measure are shown in Table 8. By comparing the table to Figures 1 and 2, no qualitative differences emerge.

*5.2.2 logitAP.* logitAP is similar to GMAP, but operates using the logistic transformation of AP values. We compute logitAP as done by Robertson (2006):

$$\text{logitAP}(s_i, t_j) = \ln\left(\frac{\text{AP}(s_i, t_j)}{1 - \text{AP}(s_i, t_j)}\right).$$

Here again we use the natural logarithm and AP values of zero (and one) need to be replaced by a small $\epsilon$ value of $10^{-5}$ (and $1 - 10^{-5}$), as done by Robertson. Using Robertson's words (2006, page 132):

> Like the log transform, or equivalently like using the geometric mean GMAP, this pays attention to hard topics in a way that ordinary MAP does not.

Table 8. Generalization to Other Metrics: GMAP

| | SNC | | | | | WUC V3 | | | | | SPO S% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ |
| TREC-8 | .50 | .52 | .63 | .46 | .43 | .48 | .45 | .57 | .68 | .32 | .69 | .55 | .68 | .73 | .41 |
| TB06 | .88 | .51 | .67 | .76 | .37 | .87 | .46 | .62 | .75 | .31 | .84 | .43 | .59 | .72 | .30 |
| TB06M | .79 | .49 | .66 | .77 | .34 | .78 | .42 | .58 | .67 | .26 | .73 | .40 | .56 | .63 | .27 |
| R04 | .88 | .71 | .88 | .85 | .53 | .81 | .65 | .83 | .83 | .52 | .63 | .48 | .66 | .43 | .34 |
| WEB14 | .93 | .74 | .89 | .66 | .70 | .73 | .54 | .75 | .51 | .35 | $.35^{\#}$ | $.09^{\#}$ | $.19^{\#}$ | .49 | −.10 |

Correlations of Pearson's $\rho$, Kendall's $\tau$, Spearman's $r_S$, Rank Biased Overlap (RBO), and $\tau_{AP}$ for the three methods SNC, WUC V3, and SPO S% on the TREC-8, TB06, TB06M, R04, and WEB14 collections, for the GMAP metric. Compare with Figures 1 and 2.
$^{\#}p > 0.05$.
All the other values have $p < 0.01$.

Table 9. Generalization to Other Metrics: logitAP

| | SNC | | | | | WUC V3 | | | | | SPO S% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ |
| TREC-8 | .34 | .52 | .63 | .46 | .42 | .33 | .42 | .55 | .69 | .30 | .58 | .55 | .69 | .71 | .43 |
| TB06 | .83 | .51 | .67 | .76 | .33 | .84 | .49 | .65 | .75 | .33 | .61 | .42 | .59 | .69 | .27 |
| TB06M | .74 | .51 | .68 | .77 | .37 | .67 | .42 | .57 | .66 | .26 | .55 | .43 | .59 | .61 | .26 |
| R04 | .85 | .71 | .88 | .84 | .53 | .82 | .64 | .82 | .81 | .50 | .37 | .30 | .37 | .32 | .15 |
| WEB14 | .94 | .75 | .90 | .66 | .69 | .71 | .49 | .68 | .51 | .39 | .47 | $.14^{\#}$ | $.21^{\#}$ | .49 | −.05 |

Table for the correlations of Pearson's $\rho$, Kendall's $\tau$, Spearman's $r_S$, and Rank Biased Overlap (RBO), for the three methods SNC, WUC V3, and SPO S% on the TREC-8, TB06, TB06M, R04, and WEB14 collections, for the logitAP metric. Compare with Figures 1 and 2 and with Table 8.
$^{\#}p > 0.05$.
All the other values have $p < 0.01$.

Results of the generalization to the logitAP measure are shown in Table 9. To compute the logitAP values, we normalized the WUC e SPO scores as done for the GMAP values: in this case, the normalization makes a difference, and performing no normalization at all leads to lower correlation values. The correlation values for SNC are almost identical to the ones obtained for GMAP (Table 8); the correlation values for WUC are comparable, especially when considering top-heavy correlation measures (RBO, and $\tau_{AP}$); the correlation values for SPO are more different, even though the similarity is still high when considering top-heavy correlation measures.

When comparing GMAP (Table 8), logitAP (Table 9), and the correlation values of Figure 2, no significant differences emerge.

*5.2.3 NDCG.* To generalize previous results to the NDCG metric we used the official NDCG@20 measure for the WEB14 collection. Results are shown in Figure 3. When we compare Figure 3 to Tables 8 (GMAP) and 9 (logitAP) as well as Figures 1 and 2 (TREC-8, R04, TB06, and WEB14 AP), we notice two visible differences: the correlation values are lower and the inverse U shape is not present. The most effective systems are not penalized, but in this case the correlation is more scattered. This behavior can be caused by the fact that having more than one relevance value (i.e., from 0 to 3), the error that can be made with a random assessment is much higher: for example, imagine a not relevant document (i.e., 0) assessed as highly relevant (i.e., 3) or vice versa.

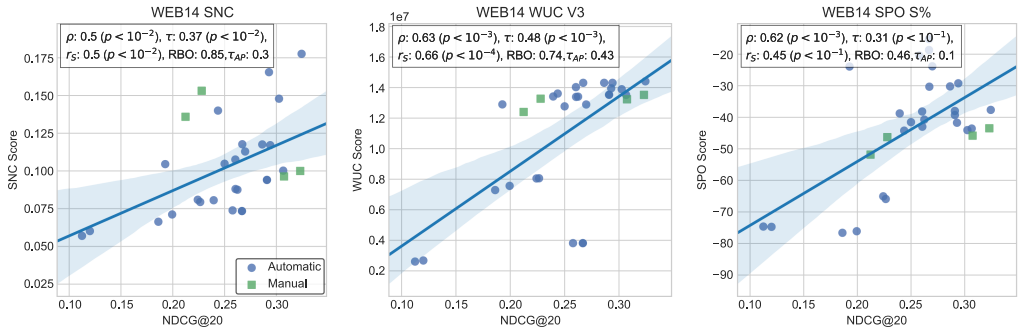



Fig. 3. Generalization to other metrics: NDCG. Scatterplots for the three methods SNC, WUC V3, SPO S% from left to right on the WEB14 collection, for the NDCG metric. Compare with Figures 1 and 2 and Tables 8 and 9.

Table 10. Generalization to a Shallow Pool: AP@20

| | **SNC**† | | | | | **WUC V3** | | | | | **SPO S%** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO | $\tau_{AP}$ |
| TREC-8 | .55 | .44 | .56 | .58 | .34 | .70 | .46 | .60 | .66 | .33 | .71 | .47 | .62 | .72 | .34 |
| TB06 | .78 | .29 | .43 | .60 | .14 | .83 | .36 | .52 | .86 | .21 | .82 | .34 | .49 | .83 | .16 |
| R04 | .78 | .51 | .70 | .74 | .29 | .90 | .65 | .82 | .79 | .45 | .90 | .65 | .82 | .80 | .42 |
| WEB14 | .53 | .43 | .56 | .88 | .28 | .83 | .59 | .74 | .92 | .36 | .80 | .49 | .64 | .93 | .27 |

Table for the correlations of Pearson's $\rho$, Kendall's $\tau$, Spearman's $r_S$, Rank Biased Overlap (RBO), and $\tau_{AP}$ for the three methods SNC, WUC V3, and SPO S% on the TREC-8, TB06, TB06M, R04, and WEB14 collections.
†All runs.
All the values have $p < 0.01$.

### 5.3 A2c: Generalize to a Shallow Pool

Since results from Section 5.2.3 are obtained with both a new metric (NDCG) and a shallow pool (20), we computed the results when considering the standard MAP metric with a shallow pool at depth 20 (i.e., AP@20). Table 10 shows the results. Comparing these results with the ones obtained considering the top 1,000 documents retrieved, we can see that the linear and rank correlations (i.e., $\rho$, $\tau$, and $r_S$) are generally lower for SNC, and higher or comparable for WUC (except for WEB14) as well as for SPO. This can be caused by the fact that a shallow pool results in low-quality qrels, which penalize SNC; on the contrary, WUC e SPO measures appear to be stable, probably because the top-ranked documents are more informative for providing an accurate final ranking of systems. For SPO, it is worth noticing that, for R04 and WEB14, the correlations obtained are significantly higher than the correlations obtained when considering all (i.e., 1,000) documents retrieved. When considering top-heavy correlations (i.e., RBO and $\tau_{AP}$), we notice that RBO values are generally comparable with the values obtained when considering the full set of documents, except, again, for WEB14 SNC and WUC V3 methods. The particular behavior of WEB14 can be explained by the fact that such a dataset is known to be rather incomplete due to shallow pools and low number of participants (Lu et al. 2016). Thus, our results suggest that considering a shallow pool does not penalize the SNC ranking of the most effective systems, which we might have supposed at first when considering linear and rank correlation values. When considering $\tau_{AP}$ correlations, we see that they are usually higher or comparable to the values obtained when considering the full set of documents, confirming previous findings.

### 5.4 Discussion

Considering all generalization experiments, we can make two general remarks. First, with the other collections TB06, TB06M, R04, and WEB14, correlation values are higher than TREC-8, and the inverse U shape that was very clear in TREC-8 (see Figure 1) disappears on the other collections: the automatic evaluation process can obtain a reasonable rank of IR systems in a completely automatic way. Second, although the choice of the metric can potentially impact the outcome of the methods for automatic evaluation of IR systems, this impact is clear only when using NDCG, i.e., a metric based on a different relevance scale (from binary to four levels).

We can also make some more specific remarks.

One issue that is worth considering in reproducibility is the choice of using the whole dataset or not (Ferro 2017; Ferro et al. 2016). In TREC result analysis, the choice is whether to use all of the dataset or only the top 75% of the most effective systems, as it is commonly done in the analysis of TREC data, see for example Voorhees and Buckley (2002). The comparison between the results restricting or not to the top 75% of most effective systems shows that there is no effect for the SNC method, while in general to consider the top 75% of runs leads to lower correlation values for the other two methods, i.e., SPO and WUC.

Another interesting issue concerning reproducibility is whether to distinguish between systems that retrieve all the documents for all the topics (e.g., 1,000 for each topic) or not. For all the analyzed collections, there are systems that do not retrieve all the documents for all the topics: an example is system READWARE for TREC-8. The result of including or not such systems depends on the method. For the SNC method, this corresponds simply to remove some points (i.e., systems) in the scatterplots. For the other two methods, i.e., SPO and WUC, the effect is twofold: some points are removed from the scatterplots, as well as the structure of overlap for SPO and the count value for WUC are recomputed; this results in lower correlation values for SPO and WUC. This is related to the well-known fact that "The quality of the pools is significantly enhanced by the presence of the recall oriented manual runs" (Voorhees and Harman 2000, page 8) that remarks the peculiarity and the effect of some runs that may do not retrieve all the documents for all the topics.

Finally, SNC is the only method which gives AP values so is the only one which can be generalized naturally, whereas WUC and SPO require normalizations that introduce some arbitrariness in the process.

## 6 A3: EXTEND

We now turn to the last aim of this article, A3, and address two novel research questions that, although they arise in a quite natural way in the context of this research, have been neglected so far.

### 6.1 A3a: Semi-Automatic Approaches

A research question that in our opinion is quite natural is what happens when a part of the evaluation is automatic and a part of it is manual, i.e., based on human relevance assessments (Aim A3a). In other terms, what happens when some values in the AP matrix of Table 2 are *artificial*, i.e., obtained by one of the three methods, and some others are *real*, i.e., obtained by means of human relevance assessments? In this section we focus on this issue.

*6.1.1 Injecting Columns.* We assume to work atomically on the topics: we do not work on individual cells of the AP matrix of Table 2 but on its columns, i.e. (since each column corresponds to a topic), on individual topics. We select some columns from the real matrix, and some others from the artificial one, i.e., we are downsampling the topics. Note that besides downsampling the topics in this manner, other alternative approaches could have been used. One might for instance

randomly sample the assessments. On the one hand, this is a convenient working assumption (and in future work we do plan to discuss the possible alternative downsampling approaches); on the other hand, however, this is also a reasonable strategy, since assessing the relevance of another document for the same topic costs less than assessing the relevance of a document for a new topic.

Therefore, instead of using all the $n$ columns of the artificial matrix only, we use $a \leq n$ columns from that matrix, and $b = n - a$ columns from the original real matrix. The $b$ columns are "injected" into the artificial matrix, and the MAP value is computed accordingly. In other terms, injecting means to fully evaluate, with human relevance assessments, specific topics.

The $b$ columns, or topics, to be injected can be selected in many ways. If the selection criterion of the $b$ columns does not depend on the real AP matrix, then this would correspond to a procedure that can be applied in practice: the artificial AP matrix is built completely automatically and, still in a completely automatic way (before any human relevance assessment takes place), some topics are selected to be injected (i.e., manually evaluated, by means of human relevance assessments). Conversely, if the selection criterion of the $b$ columns depends on the real AP matrix, then this would be useless in a practical evaluation setting, since the full results of the evaluation would be needed to do the selection according to such a criterion. However, it might also be interesting to study this case, as hidden properties of topics and evaluation in general might be revealed. Of course it would be possible to imagine also mixed or approximated selection criteria, or even to add real columns (maybe approximated to a pool depth) rather than using them to replace the artificial ones; we leave that for future work.

We take into account the following theoretical selection criteria:

(T1) Select the topics having higher (and lower as well) correlation between real AP and artificial AP. This means that we are injecting the real columns that are more (less) similar to the artificial ones. Although it can be expected that injecting the most similar columns will have a smaller effect than the less similar ones, this needs to be verified experimentally.

(T2) Select the topics having higher (lower) correlation between real AP and real MAP. This means that we are selecting the topics whose real columns are individually most (least) similar to the real MAP values, i.e., those individual topics that somehow better (worse) resemble the overall real evaluation. Those are the topics that provide a most (least) similar final ranking of IR systems, and that might be most (least) important to evaluate in an accurate way.

(T3) Select the topics having higher (lower) correlation between artificial AP and real MAP. This means that we are selecting the topics whose artificial columns are individually most (least) similar to the real MAP values, i.e., those topics that somehow, in the artificial evaluation, better (worse) resemble the overall real evaluation. Similar to the first criterion T1, injecting more similar columns is likely to have a smaller effect than a less similar one (replacing the latter would mean to "correct the errors" made by the automatic evaluation).

(T4) Select the Best (and Worst as well) possible columns according to the BestSub method presented by Guiver et al. (2009) on the real matrix. Guiver et al. described how to find "a few good (bad) topics," i.e., the topic subset of a given cardinality that evaluates the systems in the most (least) similar way to the full set of topics. This method would provide for any cardinality a subset of topics to be injected that, in the real matrix, better (worse) resemble the overall real evaluation. Different from the previous criteria, this works on topic sets rather than individual topics, thus, it can be expected to work better than previous methods. Also this conjecture needs to be verified experimentally.

(T5) Select the Best (Worst) possible column according to the HITS method of Robertson (2011) on the real matrix. This method computes for each topic its hubness, a measure of how much the topic is able to predict system effectiveness.

Furthermore, we take into account the following practical selection criteria:

(P1) Select the topics randomly (repeating the experiment to avoid noise—we use 1,000 repetitions in the following).
(P2) Select the topics having higher (lower) correlation between artificial AP and artificial MAP. This means that we are injecting the artificial columns that are more (less) similar to the artificial MAP values, i.e., those individual topics that somehow better (worse) resemble the overall artificial evaluation.
(P3) Select the Best (Worst) possible columns according to the BestSub method on the artificial matrix.
(P4) Select the Best (Worst) possible column according to the HITS method, i.e., computing its hubness, on the artificial matrix.

Note that P2, P3, and P4 correspond to T3, T4, and T5 respectively, but the former ones can be used in practice. Furthermore, P1 (i.e., sample topics randomly) can be seen as a baseline both for theoretical and practical approaches.

Before turning to the results, we remark that with these experiments we are investigating two related but different things:

—It seems intuitive that by injecting real/correct values (i.e., substituting an SNC/SPO/WUC artificial column with a real one) all the three methods will be improved. Besides verifying this, we also study which is the column selection method that provides the best results (i.e., increases most the correlation values with the ground truth, that is the real MAP value).
—It is unclear whether any of the three methods can be exploited to improve the BestSub topic selection method by Guiver et al.. That is, by computing MAP values using not only the "few good topics" subset (i.e., just a few columns of the AP matrix), but using a complete matrix with the real AP values for the best topic subsets and the artificial AP values for the other topics, do we get a system evaluation/ranking that correlates better to the real MAP value computed using the full real matrix?

In addition, with the practical experiments, we are investigating whether a practical semi-automatic evaluation is possible; and in the case of an affirmative answer, which is the best selection criterion that should be adopted.

*6.1.2 Results.* We report results for SNC only in this section; the results for the other two methods are generally worse (even despite the normalization attempts), probably because SNC is the only one that predicts the actual AP values, and during the injection process this is probably critical. We leave further analyses of the other methods to future work, as well as a complete analysis of the variability across datasets. Figure 4 displays the result for SNC for TREC-8 and R04. We report the results by showing the same correlation charts used in the state-of-the-art work on topic set reduction, see for example Berto et al. (2013), Guiver et al. (2009), and Robertson (2011): the charts in the figure show on the $x$-axis the number of columns injected from the real table, and on the $y$-axis the $\tau$ correlation values between the MAP obtained using the matrix composed by artificial and injected real topics, and the real MAP obtained using the full set of real topics.

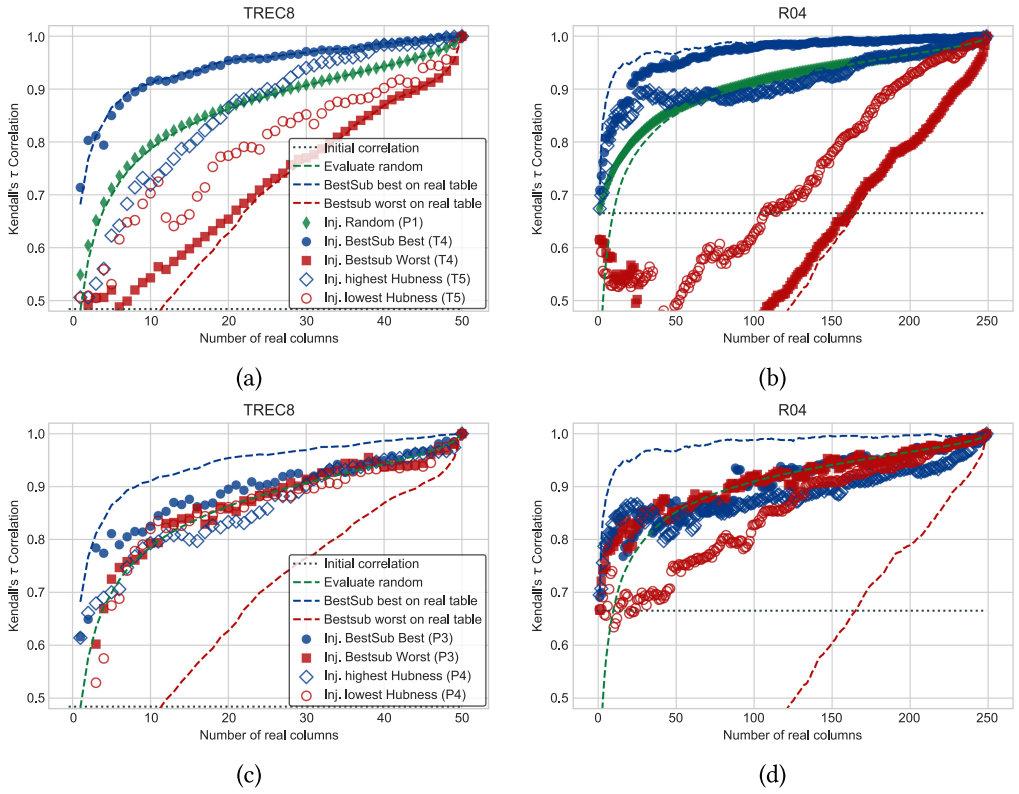



Fig. 4. Correlation curves, SNC for TREC-8 ((a) and (c)) and R04 ((b) and (d)), for the theoretical approaches T4 and T5 plus the random injection P1 ((a) and (b)), and the practical ones P3 and P4 ((c) and (d)); the horizontal black dotted line is the Kendall's correlation value for SNC method without injecting any column.

The series represent:

- —The Best/Worst series obtained by running Guiver et al.'s BestSub method on the original matrix, as well as the Average series, obtained by selecting topic subsets randomly (using 1,000 repetitions). These series are represented by the dashed lines in the plot. Note that in this case MAP is computed on a subset of topics, and not on a set composed of real and artificial topics, as it is for the other series.
- —The result of injecting topics/columns randomly, i.e., P1. This represents what happens when injecting topics without any strategy, thus it can be considered as a baseline.
- —The result of injecting the Best/Worst possible columns according to BestSub computed on the real matrix, i.e., T4 (in Figure 4(a) and (b)), or on the artificial matrix, i.e., P3 (in Figure 4(c) and (d)).
- —The result of injecting the Best/Worst possible column according to the HITS analysis computed on the on the real matrix, i.e., T5 (in Figure 4(a) and (b)), or on the artificial matrix, i.e., P4 (in Figure 4(c) and (d)).

The T1, T2, and T3 selection methods are not shown in the charts, since they have a similar behavior to the random topic injection.

Results of Figure 4(a) and (b) concern the theoretical selection criteria and show that:

—Perhaps surprisingly, the Best series are not improved by the topic injection, even when considering the theoretical Best possible columns (T4). This is valid on both datasets, and more evident in R04. Therefore, it is better to evaluate on a small subset of a few good topics rather than on a larger topic set obtained adding artificial SNC columns. Looking at injecting the Worst series, we remark that it does not decrease the correlation obtained by Worst BestSub.
—Although injecting random topics is a practical selection criterion (P1), it is shown in the charts in Figure 4(a) and (b) (also for a clearer graphical representation). Let us remark here that it does improve slightly the average subset of topics: injecting randomly is better than using a random subset of topics.
—Even though injecting the Best columns (T4) does not improve BestSub, the series are always well above the average series. Conversely, the Worst stays well below.
—The highest/lowest hubness selection criteria (T5), even if computed on the real topics, does not improve the random topic selection on TREC-8, at least not until the cardinality is higher than 20. Also for R04 the lowest hubness topic set is always well below the average, and up to cardinality 100 even below the horizontal dotted line: injecting low hubness topics is useless and, at least on R04, is even worse than not injecting topics at all. On the contrary, on R04 the highest hubness series is always above the average series for cardinalities up to 75, and it is even comparable to the BestSub selection criteria T4 for the low cardinalities.

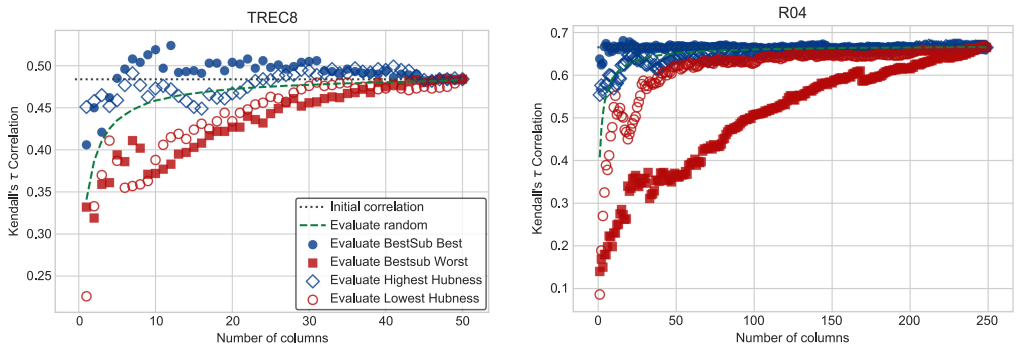
Figure 4(c) and (d) concern the corresponding practical selection criteria and show that:

—All the practical selection methods improve the artificial-only evaluation: there is almost no dot below the dotted line. This means that in the case of a semi-automatic evaluation, using practical approaches is always useful.
—As already mentioned, and shown in the Figure 4(a) and (b) charts, SNC does improve the Average BestSub series, although to a small extent: injecting topics, even randomly (P1), is still better than not injecting at all.
—The injection method does matter. Concerning P3, at low cardinalities the Best BestSub series tend to stay above random topic injection, and Worst BestSub series obtain almost always lower level of correlation than the random topic injection, for both datasets. Concerning P4, low hubness topics obtain almost always lower correlation values of both the random series and the random topic injection; on the contrary this is not true for the highest hubness topic, especially for the TREC-8 dataset.
—P2, not shown in the charts, has similar behavior to the random topic injection.

On a related issue, one might wonder what happens when combining automatic evaluation (we focus on SNC only in this analysis) with using fewer topics. In more detail, one could compare:

(i) the correlation between the real MAP and the SNC MAP, i.e., the MAP obtained averaging all the artificial AP values obtained by SNC (this is the the horizontal dotted line in Figure 4); with
(ii) the correlation between the real MAP and a "reduced" SNC MAP, i.e., a MAP obtained by considering only a limited number of topics and averaging only the corresponding artificial AP values obtained by SNC.

In other terms, one would use all the topics to run SNC, but then selects a subset of them to compute the artificial MAP. One might expect that when using fewer topics, the obtained correlation is lower that using all the topics; however, of course the selection can be done in different ways,




(a) (b)

Fig. 5. Correlation curves, SNC for TREC-8 (a) and R04 (b). The horizontal black dotted line is the Kendall's correlation value for SNC method without injecting any column. The other curves are obtained by using only a subset of topics (artificial AP values) to compute the correlation with the real MAP.

and an optimal selection of the best topics might lead to better results. We experiment with the selection criteria used above: random selection, the best and worst as found by BestSub, and high and low hubness.

Figure 5 shows the results, again on TREC-8 and R04: when using most selection methods, a subset of topics is always less effective than the whole set. However, when using the best subset of a few topics found by BestSub, the obtained correlation is higher than when using all topics. This is clearly the case for TREC-8 in the cardinality range 6–44, and it is less clear cut for R04, where anyway the subset of best topics is never worse than the full set.

Summarizing, our results indicate that, when the ground truth (i.e., the real AP matrix) is available, injecting topics does not help to improve the few good topics approach suggested by Guiver et al. (2009); this holds for any topic selection criterion we have tried. On the contrary, in the case of a semi-automatic evaluation, the Best columns selected from running the BestSub method on SNC matrix (P3) resemble the overall real evaluation better than not injecting topics at all, or injecting topics randomly.

These results show that there are some patterns in semi-automatic evaluation that are worth studying. Not only semi-automatic evaluation clearly improves the automatic only evaluation; although we have presented a limited sample of the results, it is clear that there are also important practical consequences. For example, looking at the cardinalities 10 for TREC-8 and 50 for R04 in Figure 4(c) and (d), we see that a quite high Kendall's $\tau$ correlations in the 0.8–0.9 range in system rankings can be obtained by using the SNC automatic evaluation method and "augmenting" it with the evaluation of only about 20% of the topics. This would be a practical approach to decrease the costs and resources in a test collection evaluation exercise. This needs to be further studied in future work. The results shown in Figure 5 hint at some promising future research directions as well.

### 6.2 A3b: Predicting Topic Difficulty

Another, last, very natural research question to be asked in this scenario is whether it is possible to automatically predict not only system effectiveness but also topic difficulty (see Aim A3b). First we discuss the problem in more detail, then we provide some background on the related issue of query difficulty prediction, then we present some more detailed motivations, and, finally, we describe our experiments and results.

*6.2.1 From System Effectiveness to the Dual Problem of Topic Ease.* In our context, automatically estimating topic ease corresponds to a sort of dual problem to estimating system effectiveness. To see why, let us go back to Table 2 and formula Equation (1). We can notice that while MAP represents a measure of system effectiveness, a dual measure of topic ease can be defined as proposed by Mizzaro and Robertson (2007):

$$\mathrm{AAP}(t_j) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{AP}(s_i, t_j) \tag{4}$$

(AAP stands for Average AP). Simply, in place of averaging the rows, one can average the columns.

Thus, in our context, we are now asking if the three methods can be used to predict not MAP of systems but rather AAP of topics. Although this seems a very natural research issue, it has not been addressed in the three studies that we have discussed at length so far, and by nobody else. Furthermore, the importance of this problem can be better understood by considering the related problem of query difficulty prediction.

*6.2.2 Background on Predicting Query Difficulty.* Predicting query difficulty is an important research issue. The knowledge that a query is going to be difficult might be exploited by a system, that could adopt appropriate countermeasures. The Reliable Information Access Workshop (Harman and Buckley 2004, 2009) has been the first large-scale study aimed at understanding query variability and difficulty. The many approaches that have been developed can be classified as pre-retrieval and post retrieval, depending on when the prediction takes place. The pre-retrieval approaches (Carmel and Yom-Tov 2010; Hauff et al. 2008; Sparck Jones 1988) are more practical, but the correlation of the predicted query difficulty with the ground truth is rather weak. Pre-retrieval methods can be based on statistical or linguistic approaches (Mothe and Tanguy 2005); some results on combining pre-retrieval techniques can be found in the work by Bashir (2014).

The post-retrieval approaches exploit the results of a retrieval phase, and tend to provide slightly higher correlations with the ground truth, but they are less practical. Some works using post-retrieval features can be found in the work by Shtok et al. (2012). Carmel and Yom-Tov (2010) discuss post-retrieval prediction methods using various measures like clarity, robustness, and score distribution analysis. Pre- and post-retrieval approaches can be combined in various ways, as detailed by Carmel and Yom-Tov (2010), that also provide a complete review of estimating topic difficulty as well as propose a general model for query difficulty together with some practical applications.

Other approaches have been tested. For example, Chifu et al. (2017) and Mizzaro and Mothe (2016) use human prediction (crowdsourcing) in predicting topic difficulty. Buckley (2004) proposes a measure (called AnchorMap) to compute similarity between ranked document lists retrieved by systems; this measure allows a categorization of topics into easy and difficult ones.

It is important to remark that, although query difficulty prediction is an interesting research issue, the state-of-the-art is such that no satisfying solution is available yet: when measuring the correlation between predicted and actual difficulty, the best methods reach Pearson correlation values around 0.5 (Carmel and Yom-Tov 2010; Zhao et al. 2008).

*6.2.3 Topic Ease + Query Difficulty = Topic Difficulty.* The previous brief analysis of the literature on query difficulty highlights that the important issue is topic *difficulty*, not *ease*: it is important to understand which are the difficult topics (on which the current systems can be improved), rather than the easy topics (on which the state of the art is already satisfactory).

We also need to understand that topic and query difficulty, although related, are different. In query difficulty prediction, usually the AP of a single system is being predicted, rather than the AAP over a set of systems. However, the AAP measure is an interesting alternative (Mizzaro and

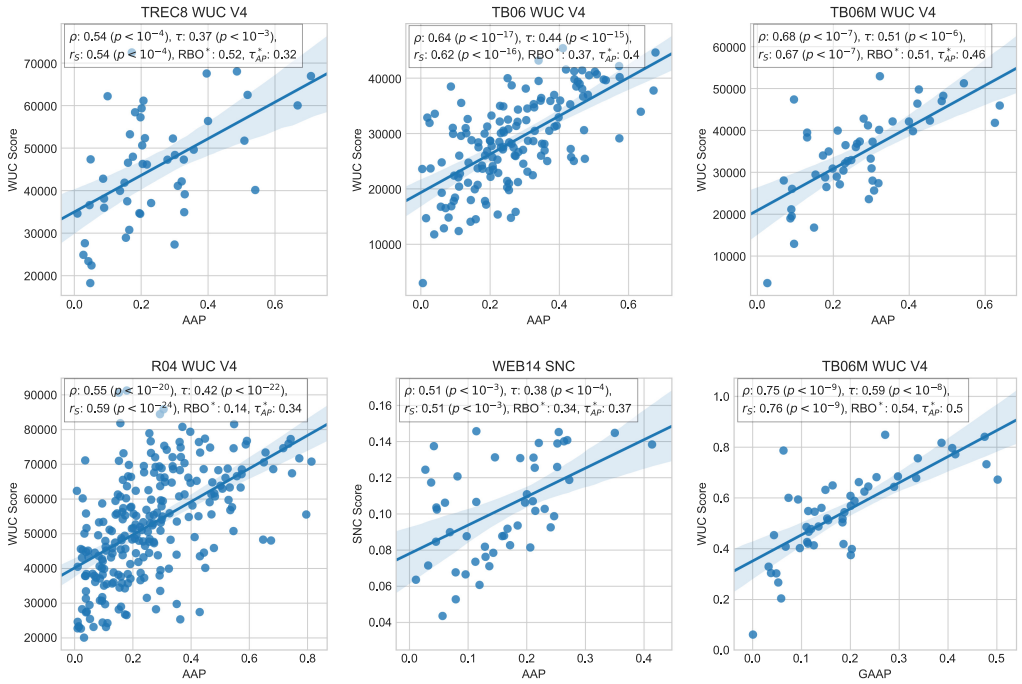



Fig. 6. AAP scatterplots for some methods and collections as indicated. Different from previous scatterplots, here each dot is a topic. In the bottom right scatterplot, GAAP is used in place of AAP and, here and in the following, GAAP values have been normalized as it has been done for GMAP (see Section 5.2.1).

Robertson 2007). Studying the difficulty of a topic rather than a query makes sense also given the recent result by Thomas et al. (2017) who find that "task difficulty" would be a more reliable notion than "query difficulty." Finally, although, of course, the AAP of a topic will depend on the systems participating to the evaluation exercise, the measure seems quite stable. For example, there are 50 common topics between R04 and TREC-8; the AAP values computed using the two different systems populations feature high correlations (Pearson's $\rho$ correlation is 0.99 and Kendall's $\tau$ is 0.89).

In other terms, what we are proposing is a post-retrieval topic difficulty prediction method that, at least in principle (i.e., not taking efficiency into account), could be used in practice. Before turning to the results, let us notice that, given the unsatisfactory results obtained by state of the art query difficulty predictors (correlations higher than $\rho = 0.5$ are difficult to obtain), this seems an interesting and promising approach.

*6.2.4 Experimental Results on Predicting Topic Difficulty.* At first sight, results seem not exciting. Figure 6 shows the AAP scatterplots for the five collections and some selected methods (we report the methods with the highest correlations: usually WUC V4, with the exception of the fifth chart—on WEB14 SNC is slightly better). Different from the scatterplots presented so far in this article, here each dot is a topic, not a system, and the axes represent real and predicted AAP, not MAP. The last (bottom right) chart in figure uses, in place of AAP, a slightly different metric, Geometric AAP, that can be defined as

$$\mathrm{GAAP}(t_j) = \sqrt[m]{\prod_{i=1}^{m} \mathrm{AP}(s_i, t_j)} = \exp\left(\frac{1}{m}\sum_{i=1}^{m} \ln(\mathrm{AP}(s_i, t_j))\right)$$

Table 11. AAP Correlations; GAAP in the Last Row

| | SNC | | | | | WUC V4 | | | | | SPO S-A% | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | $r_S$ | RBO* | $\tau^*_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO* | $\tau^*_{AP}$ | $\rho$ | $\tau$ | $r_S$ | RBO* | $\tau^*_{AP}$ |
| TREC-8 | .26$^{\#}$ | .24$^{+}$ | .35$^{+}$ | .44 | .18 | .54 | .37 | .54 | .52 | .32 | .33$^{+}$ | .30 | .42 | .46 | .25 |
| TB06 | .51 | .38 | .53 | .14 | .35 | .64 | .44 | .62 | .37 | .40 | .53 | .37 | .51 | .24 | .36 |
| TB06M | .47 | .30 | .42 | .43 | .27 | .68 | .51 | .67 | .51 | .46 | .44 | .31 | .44 | .41 | .28 |
| R04 | .26 | .24 | .36 | .09 | .20 | .55 | .42 | .59 | .14 | .34 | .23 | .21 | .31 | .06 | .20 |
| WEB14 | .51 | .38 | .51 | .34 | .37 | .52 | .35 | .49 | .20 | .32 | .32$^{+}$ | .34 | .44 | .09 | .34 |
| **GAAP** | | | | | | | | | | | | | | | |
| TB06M | .58 | .44 | .58 | .47 | .41 | .75 | .59 | .76 | .54 | .50 | .54 | .44 | .59 | .41 | .42 |

$^{+}p < 0.05$.
$^{\#}p > 0.05$.
All the other values have $p < 0.01$.

(GAAP is to AAP as GMAP is to MAP, compare also to formula Equation (3)). GAAP emphasizes more the difficult/low end of the topic difficulty scale, which seems the interesting one if one wants to work on difficult topics.[9] For the same reasons, in Figure 6 and in the following, we report RBO* and $\tau^*_{AP}$, which are the bottom-heavy versions of the top-heavy RBO and $\tau_{AP}$, computed simply by reversing the order of the vectors to give more weight to the difficult topics. Table 11 shows all the correlation values for the five collections, AAP (GAAP in one case), and the three methods (we selected the overall most effective method variants; V3 and V4 have similar correlation values).

The best results in topic difficulty prediction are obtained for TB06M with WUC V4 (see the two right most charts in figure and the values in the table); these values are even higher when GAAP is used. In these cases the correlation values are comparable to, if not even higher than, the state-of-the-art, which is around 0.5. However, when comparing the scatterplots and the values with the previous ones, the most visible difference is that correlations are much lower: topic ease seems less predictable than system effectiveness.

However, this is not the whole story. While it is important to understand which system is the best (the most effective one, having the higher effectiveness value), for the topics it is rather important to understand which are the difficult ones (the ones having the lowest effectiveness values), since on these alternative strategies can be used to improve effectiveness. Some further analysis is shown in Figure 7 that shows, for each scatterplot of Figure 6, one boxplot chart where the topics are grouped into quartiles according to their real AAP (GAAP) value. So, the $x$-axis represents the quartiles, the $y$-axis is still the topic ease predicted by the method, the dots are still topics (the small horizontal variations on the dots is just a random jitter for graphical reasons, to avoid overlapping dots), and the boxes summarize the distribution of the predicted values for each quartile, with the horizontal line corresponding to the median. For all six charts, with just one exception, the median of each quartile is lower than the subsequent ones. This means that if the difficulty of topics is measured by categorizing them into the four difficulty categories, the three methods are reasonably good in predicting it (although, of course, the increasing median is not a sufficient condition to make the four different classes fully separable). Furthermore, we ran an unpaired $t$-test using the Bonferroni's correction (Dunn 1961): whereas for the four charts about TREC-8, TB06M, and WEB14, the differences between adjacent quartiles are mostly not significant (also because only 50 topics occur in those charts, and of course even fewer in each quartile), the differences between the first and the third quartile, for the two charts about TB06 and R04, are significant.

[9]It is also possible to use logitAP in place of ln(AP), but in our experiments we did not find any significant difference.

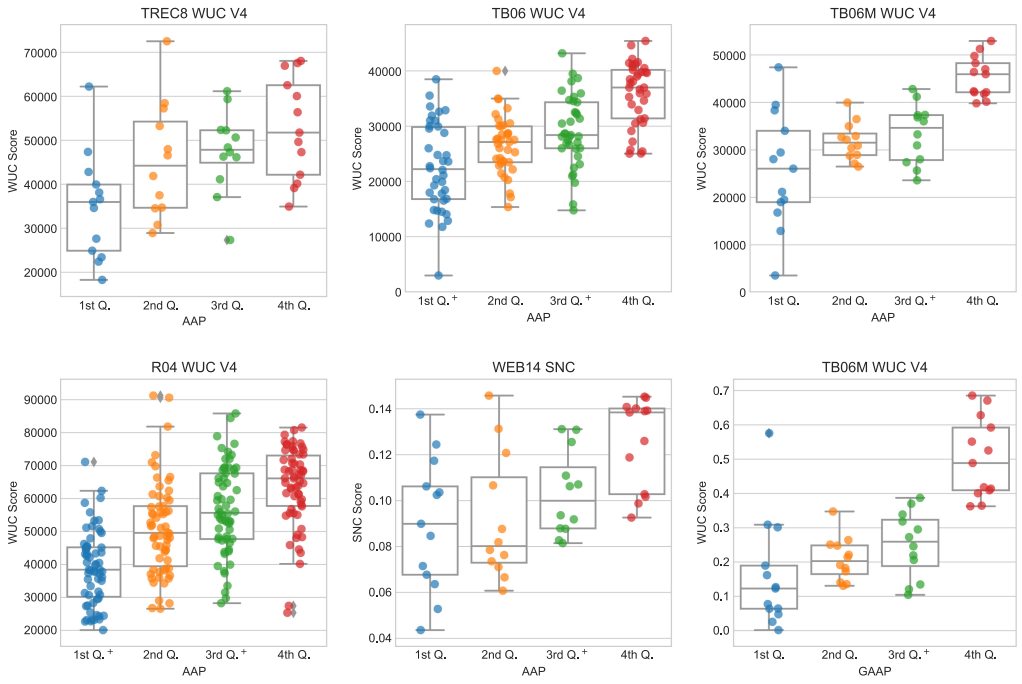



Fig. 7. AAP boxplots for the scatterplots of Figure 6 divided into quartiles according to AAP (or GAAP); statistical significance computed considering Bonferroni's correction.

Figure 8 presents yet another analysis. The six charts are again those corresponding to the charts in Figures 6 and 7 but, in this figure, the topics are grouped into quartiles accordingly to the difficulty predicted by the SNC/WUC/SPO method, rather than their real difficulty. What is interesting here is that for each method the topics that are in the first quartile turn out to be indeed difficult ones also accordingly to their exact AAP value, with very few exceptions. Although this is also true for some topics in the second, third, and fourth quartiles (especially for R04 and TB06, but note that the number of topics is much higher for those collections), this last result can have immediate practical applications. We could build an IR system that adopts some countermeasures (e.g., perform an automatic query reformulation, ask further information to the user, and so on) for the difficult topics. And we can reliably predict which topics are difficult by selecting those in the first quartile according to each of the three methods. If the countermeasures are effective, this would lead to a more effective IR system.

In other terms, for topic difficulty prediction, precision seems more important than recall: if a topic prediction system fails to recognize that a topic is difficult, no harm is done; conversely, if an IR system adopts some countermeasure on a topic that is easy but is wrongly predicted as false, this would likely decrease retrieval effectiveness. The three methods are indeed precision-oriented ones. To state it in yet another way, going back to Figure 6, what is important is that no topics, or at least very few of them, are in the bottom-right part of the charts, which is indeed the case: the vast majority of dots are in the top left triangular part.

## 7 CONCLUSIONS AND FUTURE WORK

In this article, we set to reproduce the most important work on automatic evaluation of IR effectiveness, i.e., evaluation without human relevance judgments. Instead of only reproducing the work, we provide a fourfold contribution:

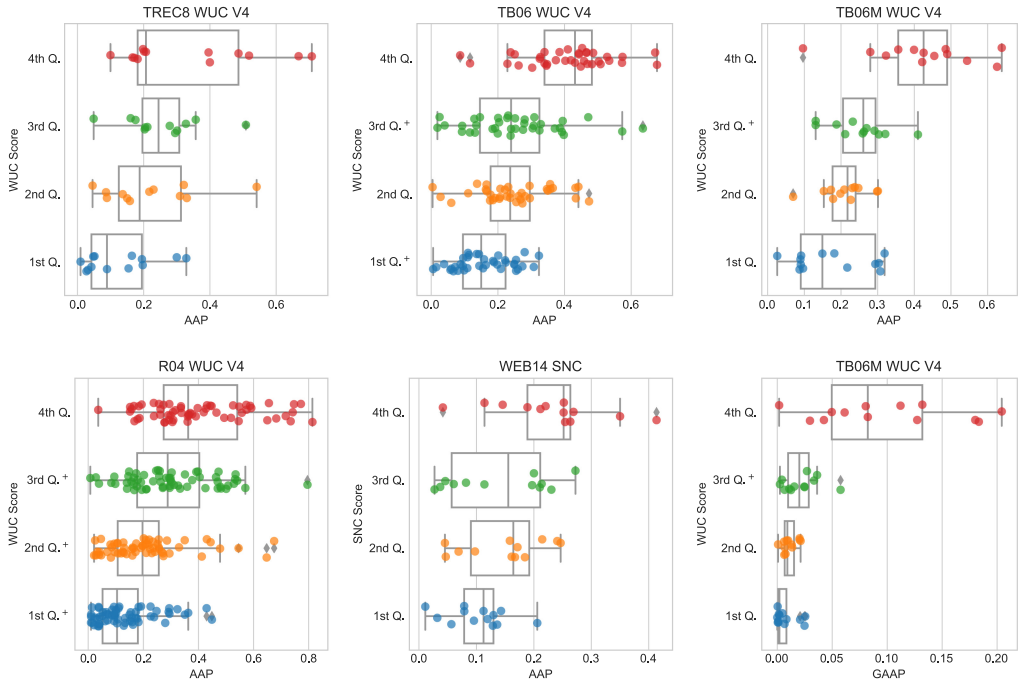



Fig. 8. AAP boxplots for the scatterplots of Figure 6 divided into quartiles according to SNC/SPO/WUC score. Statistical significance computed considering Bonferroni's correction.

(i) we succeeded in reproducing the main results of three previous similar studies, with only some minor caveats, which we discussed in the respective sections; we released all the code used to carry out the experiments; different from the original works, we focused on future reproducibility, and we detailed all the parameters required to implement each method;
(ii) we presented the results in a uniform way;
(iii) we generalized those results to other test collections, evaluation metrics, and a shallow pool; and
(iv) we expanded those results, obtaining two practical strategies that seem effective to, respectively, decrease the costs involved in test-collection-based evaluation, and improve retrieval effectiveness on difficult topics.

A general lesson learned of methodological nature is that we believe that this is the right attitude in a reproducibility setting: not only simply reproducing, but also providing a uniform representation; in our case this lead naturally to generalization, and was also inspiring to obtain the apparently very interesting results (iv).

This article also suggests some future work, some of which has already been mentioned above and is not repeated here. The results (iv) will need further verifications on other datasets, fine tuning of the methods, and further experiments on other similar methods. We have not tried to combine the three methods, which is surely a natural attempt. Injecting strategies to improve AAP estimation, instead of MAP, are less straightforward but can be devised, perhaps at the individual AP level. Combinations of injecting and shallow pool-based (i.e., approximated metric computation) methods can also be devised.

## ACKNOWLEDGMENTS

We want to acknowledge the help obtained by the reviewers, who provided long, detailed, and insightful comments, always with a constructive attitude. This article has greatly improved thanks to their feedback.

## REFERENCES

Omar Alonso and Stefano Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Vol. 15. 16.

Javed A. Aslam and Robert Savell. 2003. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR'03)*. ACM, New York, NY, 361–362. DOI: https://doi.org/10.1145/860435.860501

Shariq Bashir. 2014. Combining pre-retrieval query quality predictors using genetic programming. *Applied Intelligence* 40, 3 (April 2014), 525–535. DOI: https://doi.org/10.1007/s10489-013-0475-z

Andrea Berto, Stefano Mizzaro, and Stephen Robertson. 2013. On using fewer topics in information retrieval evaluations. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval (ICTIR'13)*. ACM, New York, NY, Article 9, 8 pages. DOI: https://doi.org/10.1145/2499178.2499184

Chris Buckley. 2004. Topic prediction based on comparative retrieval rankings. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM, New York, NY, 506–507. DOI: https://doi.org/10.1145/1008992.1009093

Stefan Büttcher, Charles L. A. Clarke, and Ian Soboroff. 2006. The TREC 2006 terabyte track. In *TREC*, Vol. 6. 39. http://trec.nist.gov/pubs/trec15/papers/TERA06.OVERVIEW.pdf.

David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval* (1st ed.). Morgan and Claypool Publishers.

Ben Carterette and Ian Soboroff. 2010. The effect of assessor error on IR system evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 539–546. DOI: https://doi.org/10.1145/1835449.1835540

Adrian-Gabriel Chifu, Sébastien Déjean, Stefano Mizzaro, and Josiane Mothe. 2017. *Human-Based Query Difficulty Prediction*. Springer International Publishing, Cham, 343–356. DOI: https://doi.org/10.1007/978-3-319-56608-5_27

Cyril W. Cleverdon. 1991. The significance of the cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'91)*. ACM, New York, NY, 3–12. DOI: https://doi.org/10.1145/122860.122861

Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.

Nicola Ferro. 2017. Reproducibility challenges in information retrieval evaluation. *J. Data and Information Quality* 8, 2, Article 8 (Jan. 2017), 4 pages. DOI: https://doi.org/10.1145/3020206

Nicola Ferro, Norbert Fuhr, Kalervo Järvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. 2016. Increasing reproducibility in IR: Findings from the dagstuhl seminar on reproducibility of data-oriented experiments in e-science. In *ACM SIGIR Forum*, Vol. 50. ACM, 68–82.

Norbert Fuhr. 2017. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum* 51, 3 (2017), 32–41. DOI: https://doi.org/10.1145/3190580.3190586

John Guiver, Stefano Mizzaro, and Stephen Robertson. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Trans. Inf. Syst.* 27, 4, Article 21 (Nov. 2009), 26 pages. DOI: https://doi.org/10.1145/1629096.1629099

Donna Harman and Chris Buckley. 2004. The NRRC reliable information access (RIA) workshop. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM, New York, NY, 528–529. DOI: https://doi.org/10.1145/1008992.1009104

Donna Harman and Chris Buckley. 2009. Overview of the reliable information access workshop. *Inf. Retr.* 12, 6 (Dec. 2009), 615–641. DOI: https://doi.org/10.1007/s10791-009-9101-4

Claudia Hauff and Franciska de Jong. 2010. Retrieval system evaluation: Automatic evaluation versus incomplete judgments. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 863–864. DOI: https://doi.org/10.1145/1835449.1835654

Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, New York, NY, USA, 1419–1420. DOI: https://doi.org/10.1145/1458082.1458311

Joon Ho Lee. 1997. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*. ACM, New York, NY, 267–276. DOI: https://doi.org/10.1145/258525.258587

Aldo Lipani, Guido Zuccon, Mihai Lupu, Bevan Koopman, and Allan Hanbury. 2016. The impact of fixed-cost pooling strategies on test collection bias. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR'16)*. ACM, New York, NY, 105–108. DOI: https://doi.org/10.1145/2970398.2970429

David E. Losada, Javier Parapar, and Alvaro Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management* 53, 5 (2017), 1005–1025. DOI: https://doi.org/10.1016/j.ipm.2017.04.005

Xiaolu Lu, Alistair Moffat, and J. Shane Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Inf. Retr.* 19, 4 (Aug. 2016), 416–445. DOI: https://doi.org/10.1007/s10791-016-9282-6

Stefano Mizzaro and Josiane Mothe. 2016. Why do you think this query is difficult? A user study on human query prediction. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1073–1076.

Stefano Mizzaro and Stephen Robertson. 2007. Hits hits TREC: Exploring IR evaluation results with network analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. ACM, New York, NY, 479–486. DOI: https://doi.org/10.1145/1277741.1277824

Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *ACM Conference on Research and Development in Information Retrieval, SIGIR, Predicting Query Difficulty-Methods and Applications Workshop*. 7–10.

Rabia Nuray and Fazli Can. 2003. Automatic ranking of retrieval systems in imperfect environments. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR'03)*. ACM, New York, NY, 379–380. DOI: https://doi.org/10.1145/860435.860510

Stephen Robertson. 2006. On GMAP: And other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. ACM, New York, NY, 78–83. DOI: https://doi.org/10.1145/1183614.1183630

Stephen Robertson. 2011. On the contributions of topics to system evaluation. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval - Volume 6611 (ECIR 2011)*. Springer-Verlag New York, Inc., New York, NY, 129–140. DOI: https://doi.org/10.1007/978-3-642-20161-5_14

Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. 2017. Do easy topics predict effectiveness better than difficult topics? In *Advances in Information Retrieval: 39th ECIR 2017*. 605–611.

Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.* 30, 2, Article 11 (May 2012), 35 pages. DOI: https://doi.org/10.1145/2180868.2180873

Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*. ACM, New York, NY, 66–73. DOI: https://doi.org/10.1145/383952.383961

Karen Sparck Jones. 1988. A statistical interpretation of term specificity and its application in retrieval. In *Document Retrieval Systems*. Taylor Graham Publishing, London, UK, 132–142. http://dl.acm.org/citation.cfm?id=106765.106782

Anselm Spoerri. 2005. How the overlap between the search results of different retrieval systems correlates with document relevance. *Proceedings of the American Society for Information Science and Technology* 42, 1 (2005). DOI: https://doi.org/10.1002/meet.14504201175

Anselm Spoerri. 2007. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Processing & Management* 43, 4 (2007), 1059–1070. DOI: https://doi.org/10.1016/j.ipm.2006.09.009

Gregory Tassey, Brent R. Rowe, Dallas W. Wood, Albert N. Link, and Diglio A. Simoni. 2010. Economic impact assessment of NIST's text REtrieval conference (TREC) program. *Report prepared for National Institute of Technology (NIST)* (2010). https://trec.nist.gov/pubs/2010.economic.impact.pdf.

Paul Thomas, Falk Scholer, Peter Bailey, and Alistair Moffat. 2017. Tasks, queries, and rankers in pre-retrieval performance prediction. In *Proceedings of the 22nd Australasian Document Computing Symposium (ADCS'17)*. ACM, New York, NY, Article 11, 4 pages. DOI: https://doi.org/10.1145/3166072.3166079

Ellen M. Voorhees. 2004. Overview of the TREC 2004 robust track. In *TREC*, Vol. 4. http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf.

Ellen M. Voorhees and Chris Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*. ACM, New York, NY, 316–323. DOI: https://doi.org/10.1145/564376.564432

Ellen M. Voorhees and Donna Harman. 2000. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*. 1–24. http://trec.nist.gov/pubs/trec23/papers/overview-web.pdf.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (Nov. 2010), 38 pages. DOI: https://doi.org/10.1145/1852102.1852106

Shengli Wu and Fabio Crestani. 2002. Data fusion with estimated weights. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02)*. ACM, New York, NY, 648–651. DOI: https://doi.org/10.1145/584792.584908

Shengli Wu and Fabio Crestani. 2003. Methods for ranking information retrieval systems without relevance judgments. In *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC'03)*. ACM, New York, NY, 811–816. DOI: https://doi.org/10.1145/952532.952693

Emine Yilmaz and Javed A. Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. ACM, New York, NY, 102–111. DOI: https://doi.org/10.1145/1183614.1183633

Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. ACM, New York, NY, 587–594. DOI: https://doi.org/10.1145/1390334.1390435

Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR'08)*. Springer-Verlag, Berlin, 52–64. http://dl.acm.org/citation.cfm?id=1793274.1793285.

Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*. ACM, New York, NY, 307–314. DOI: https://doi.org/10.1145/290941.291014

Received October 2017; revised April 2018; accepted July 2018