Information Retrieval Journal
https://doi.org/10.1007/s10791-019-09357-w

Fewer topics? A million topics? Both?! On topics subsets in test collections

Kevin Roitero[1] · J. Shane Culpepper[2] · Mark Sanderson[2] · Falk Scholer[2] · Stefano Mizzaro[1]

Received: 23 November 2017 / Accepted: 26 April 2019

Abstract
When evaluating IR run effectiveness using a test collection, a key question is: What search topics should be used? We explore what happens to measurement accuracy when the number of topics in a test collection is reduced, using the Million Query 2007, TeraByte 2006, and Robust 2004 TREC collections, which all feature more than 50 topics, something that has not been examined in past work. Our analysis finds that a subset of topics can be found that is as accurate as the full topic set at ranking runs. Further, we show that the size of the subset, relative to the full topic set, can be substantially smaller than was shown in past work. We also study the topic subsets in the context of the power of statistical significance tests. We find that there is a trade off with using such sets in that significant results may be missed, but the loss of statistical significance is much smaller than when selecting random subsets. We also find topic subsets that can result in a low accuracy test collection, even when the number of queries in the subset is quite large. These negatively correlated subsets suggest we still lack good methodologies which provide stability guarantees on topic selection in new collections. Finally, we examine whether clustering of topics is an appropriate strategy to find and characterize good topic subsets. Our results contribute to the understanding of information retrieval effectiveness evaluation, and offer insights for the construction of test collections.

Keywords Retrieval evaluation · Few topics · Statistical significance · Topic clustering

✉ Stefano Mizzaro
mizzaro@uniud.it

Kevin Roitero
roitero.kevin@spes.uniud.it

J. Shane Culpepper
shane.culpepper@rmit.edu.au

Mark Sanderson
mark.sanderson@rmit.edu.au

Falk Scholer
falk.scholer@rmit.edu.au

[1] Department of Maths, Computer Science, and Physics, University of Udine, Udine, Italy

[2] School of Science, RMIT, Melbourne, Australia

Published online: 08 May 2019

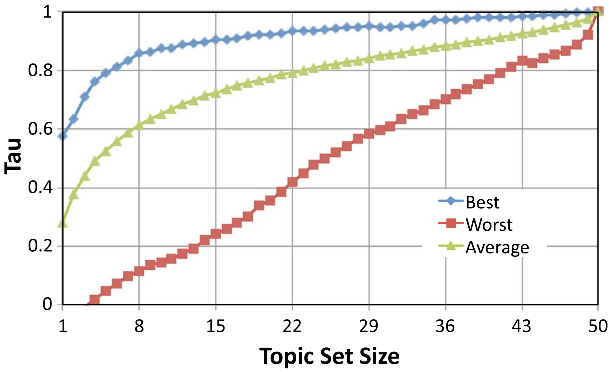

Fig. 1 Kendall's τ correlation curves for AH99, adapted from Guiver et al. (2009, Figure 2)

1 Introduction and background

When evaluating the effectiveness of Information Retrieval (IR) systems, the design of the measurement process has been examined by researchers from many 'angles': e.g. the consistency of relevance judgments; the means of minimizing judgments while maintaining measurement accuracy; and the best formula for measuring effectiveness. One aspect—the number and type of queries (*topics* in TREC terminology) needed in order to measure reliably—has been discussed less often. In general, there has been a trend in test collection construction of increasing the number of topics, but without much consideration of the benefits of such an approach. In many areas of measurement via sampling, it is generally accepted that there are diminishing returns from increasing the sample size (Bartlett et al. 2001). Beyond a certain point, improvements in measurement accuracy are small and the cost of creating the sample becomes prohibitive. We are not aware of work in IR that establishes if such an optimal sample size exists.

Other work has been conducted on whether smaller topic sets (*subsets*) could be used in a test collection, examining early TREC ad hoc collections (Guiver et al. 2009; Robertson 2011; Berto et al. 2013), and the 2009 Million Query (MQ) Track (Carterette et al. 2009a, b). These approaches, in general, ask how similarly a set of retrieval runs are ranked when using such a subset versus a full set of topics. Note that in these experiments, the full set of topics is taken to be the *ground truth*. The similarity of the two rankings is measured using Kendall's Tau (henceforth, τ). Figure 1 shows an example result from this work, taken from Guiver et al. (2009). On the x-axis are topic subsets of increasing cardinality, the y-axis measures τ. Three types of subset are shown for each cardinality:

- Best—the subset of a given cardinality that results in a ranking that is closes to the ranking of runs using the full topic set;
- Average—the average τ of all the topic subsets examined;
- Worst—the topic subset that results in a ranking that is furthest from the ranking of runs from the full topic set.[1]

[1] Guiver et al. (2009) use the terminology Best/Average/Worst, and we adopt it in this paper in order to be consistent with past work.

The best correlation curve shows that even when using a topic subset of cardinality 6, a relatively high τ (> 0.8) can be found. The curve for the average topic set reaches a τ of 0.8 at cardinality 22. The generality of this basic result was questioned by Robertson (2011), and revisited again by Berto et al. (2013) with results that confirmed the original conclusions.[2] Carterette et al. (2009a) conducted similar experiments though only measuring the average. However, they also examined different topic types, which will be discussed later.

There are a number of limitations with these past studies:

1. Researchers have examined relatively small ground truth topic sets: $n = 50$ (Guiver et al. 2009; Robertson 2011; Berto et al. 2013) and $n = 87$ (Carterette et al. 2009a). However, little is known about the generality of these results for larger n. Because the existing studies sampled from topic sets that are relatively small, as the cardinality of the subset becomes a substantial fraction of the ground truth set, the properties of the sample and the full set are guaranteed to become similar and the correlations between the rankings of runs will tend to 1. The observation in Fig. 1 that a topic subset of cardinality 22 has similar properties to the full set of 50 topics may not hold with a larger ground truth. This limitation is striking in the light of recent work by Sakai (2016b), who showed that for test collections to have reasonable statistical power, ground truth topic sets size should be at least around 200, if not higher. Therefore the results obtained on the basis of a ground truth of far fewer than 100 topics calls for further confirmation on higher cardinalities.
2. A limitation of past work (Guiver et al. 2009; Robertson 2011; Berto et al. 2013) is that the statistical significance of the differences between runs was not taken into account: τ values do not explain if a different run ranking is due to minor fluctuations or to statistically significant differences in measurement values. This is a notable omission, in the light of recent work from Sakai (2014) that emphasizes the link between topic set size and statistical power.
3. Almost no characterization of the best topic sets has been attempted [apart some results on stability of such sets, see e.g. Figures 5 and 6 in Guiver et al. (2009)]. However, it seems intuitive that smaller topic sets should be obtained by removing redundancy, for example by clustering topics and selecting representatives from each cluster.

In this paper we address three research questions:

RQ1. What effect does a larger ground truth topic set have on correlation curves? Are the results obtained in past work (Guiver et al. 2009; Robertson 2011; Berto et al. 2013) confirmed when using a larger ground truth? How does the minimum cardinality of a topic subset, needed in order to achieve a high correlation, depend on the cardinality of the ground truth, when using data from test collections?

RQ2. Are the results on topic subset size, obtained in past work (Guiver et al. 2009; Robertson 2011; Berto et al. 2013), still valid when statistical significance is considered?

[2] It is important to remark that this line of research focuses on an *a posteriori*, i.e., after-evaluation setting: it is not aimed at predicting in advance a good topic subset, but only at determining if such a subset exists.

RQ3. Is clustering an effective strategy to potentially[3] find and characterize the best topic sets? Does the choice of a specific clustering setting (features, algorithms, distance functions, etc.) make important differences? If so, what clustering settings are most effective in finding topic sets featuring high correlations?

In the rest of the paper, Sect. 2 frames the context of this research by analyzing related work. Section 3 describes the experimental setting. Section 4 discusses the results related to the first research question RQ1, highlighting the existence of even more extreme results when the number of topics increases. Section 5 focuses on RQ2 and addresses statistical significance, specifically discussing what kind of errors are more likely when using fewer topics. Section 6 examines RQ3, about clustering, and highlights how a rather natural approach turns out to be only slightly more effective than randomly chosen topics. Section 7 summarizes the contribution of this paper and sketches future developments.

2 Related work

In addition to the previously mentioned work examining topics (Guiver et al. 2009; Robertson 2011; Berto et al. 2013), a wide variety of studies analyze the components of test collections. Here, we focus on those that consider the number of topics needed and topic subsetting.

2.1 Number of topics

Buckley and Voorhees (2000) examined the accuracy of common evaluation measures relative to the number of topics used. They suggested using at least 25 topics, though stated having more was better. The authors concluded that 50 topics produce reliable evaluations. The conclusion on the number of topics was broadly confirmed by Carterette et al. (2006) who considered a larger number of topics (200).

While the methods used in earlier work to determine the appropriate number of topics for a test collection involved a range of empirical approaches, Webber et al. (2008) proposed the use of statistical power analysis when comparing the effectiveness of runs. The authors argued that a set of nearly 150 topics was necessary to distinguish runs. Building on suggestions by Sanderson and Zobel (2005), they also argued that using more topics with a shallow assessment pool was more reliable than using few topics with a deep assessment pool. Carterette and Smucker (2007) used power analysis statistics to study both topic set size and judgment set size.

Using the approach of Test Theory, introduced by Bodoff and Li (2007) and Urbano et al. (2013) examined test collection reliability considering all aspects of the collection. The authors tabulated their measures of reliability across a large number of TREC collections, and suggested that the number of topics used in most current test collections is insufficient.

[3] Consistently with this line of research (see Footnote 2), we investigate clustering of topics using an *a posteriori* setting; thus, we study an after-evaluation characterization of Best topic subsets, but do not aim at providing a methodology to find such subsets in practice.

More recently, Sakai (2014, 2016b) used power analysis to argue that more topics than are currently found in most test collections are required. He showed that many significant results may be missed due to the relatively small number of topics in current test collections. He concludes that potentially, hundreds of topics are required to achieve reasonable power in current test collections.

While the works here seem to draw contradictory conclusions of different minimum numbers, a common theme to the work is that the minimum number needed to separate the effectiveness of two runs depends on how similar the runs are. The earlier work examined runs more widely separated than more recent work.

2.2 Topic subsets

Separate to the question of how many topics are required, researchers have asked if some form of targeted topic sample could achieve the same measurement effect.

Subsequent to the work of Mizzaro and Robertson (2007) on topic subsets, Hauff et al. (2009) presented three approaches to measure effectiveness estimation using topic subsets: greedy, median Average Precision (AP), and estimation accuracy. Hauff et al. (2010) then presented evidence showing that the accuracy of ranking the best runs depended on the degree of human intervention in any manual runs submitted, and went on to show that this problem can be somewhat alleviated by using a subset of "best" topics. Cattelan and Mizzaro (2009) also studied whether it is possible to evaluate different runs with different topics. Roitero et al. (2017) generalized the approach to other collections and metrics, further investigating the correlations between topic ease and its capability of predicting system effectiveness.

In contrast to the work conducted by Mizzaro and Robertson—which looked for best and worst subsets in a "bottom up" approach, finding any topics that would fit into each subset—Carterette et al. (2009a) took a "top down" approach. They manually split the topics of the MQ collections into subsets based on groups of categories from Rose and Levinson (2004). They found little difference examining the groups. They also looked at different combinations of hard, medium, and easy topics (determined by the average score that runs obtained on the topics) and found similar conclusions to earlier topic subset work.

In related work, Hosseini et al. (2011b) presented an approach to expand relevance judgments when new runs are evaluated. The cost of gathering additional judgments was offset by selecting a subset of topics that discriminated the runs best, determined using Least Angle Regression (LARS) and convex optimization, up to a maximum topic set cardinality of 70. Later, Hosseini et al. (2011a) used convex optimization to select topics that needed further relevance judgments when evaluating new runs. The algorithm estimates the number of unjudged documents for a topic and identifies a set of query-document pairs that should be judged given a fixed budget.

Hosseini et al. (2012) proposed a mathematical framework to select topic subsets based on modeling the evaluation metric's uncertainty obtained when dealing with incomplete or missing relevance judgments for a set of topics. This work is particularly relevant as we will be able to compare some of our results with theirs.

Kutlu et al. (2018) developed a method for topic selection based on learning-to-rank; they took into account the effect of pool depth and focused on deep versus shallow judging.

Table 1 Test collections used for all experiments

Acronym	TREC collection	Year	Topics	Total runs	Used runs
AH99	Ad hoc	1999	50	129	96
R04	Robust	2004	249	110	82
TB06	TeraByte	2006	149	61	49
MQ07	Million query	2007	1153	29	26

3 Experimental setting and data

We describe the test collections, methods, and means of evaluation used in our experiments.

3.1 Data and collections

Our experiments require test collections with more than 50 topics, and for which a sufficient number of runs are available to be analyzed. The three instantiations of the Million Query track collections feature more than 1000 topics each year that are sampled from a query log. We use the data from the 2007 track. However, the Million Query datasets are not free from disadvantages: runs are evaluated using the statMAP and E[MAP] metrics, which are slightly different from classical Mean AP (MAP).[4] In addition, not as many runs are available (25–35). We also employ the TREC 2004 Robust and 2006 TeraByte track collections, using automatic runs only. To enable a comparison with the results obtained in previous studies (Guiver et al. 2009; Robertson 2011; Berto et al. 2013), we also use the TREC 8 ad hoc (AH) track (1999). Table 1 summarizes the four test collections. For the analyses in this paper, when not otherwise noted, we work on a subset of the runs. As is usual for the analysis of TREC run data (see e.g. Voorhees and Buckley 2002), we remove the least effective runs, obtaining the number of runs in the last column. For AH99 we removed the 25% least effective runs to have the same situation as in prior work (Guiver et al. 2009; Robertson 2011; Berto et al. 2013); for R04 we did the same; for TB06 and especially MQ07, which feature a smaller number of runs, we removed fewer (20% and 10%, respectively). The number removed was determined by manually examining the distribution of run effectiveness values, and pruning runs with a clear drop in effectiveness compared to others that are ranked higher.

3.2 Software

For our analysis, we employed the BestSub software that was used in previous studies (Guiver et al. 2009; Robertson 2011; Berto et al. 2013). The number of all subsets of a topic set of cardinality n is 2^n. The number of all possible topic subsets of cardinality c drawn from the larger set is $\binom{n}{c} = \frac{n!}{c!(n-c)!}$. Therefore, the software uses a heuristic to cope with the combinatorial explosion. The heuristic builds the best set of cardinality $c + 1$ on the basis of the best set of cardinality c by looking at those subsets of cardinality $c + 1$ that differ from the best set of cardinality c by at most k topics. In the previous studies, $k = 3$.

[4] The effect of statMAP, on which we focus in this paper, is discussed in more detail in Sect. 3.3.

Since in our case $n > 50$ (i.e. 149, 249, and 1153), the complexity is higher. This would mean that using BestSub was impractical, with months if not years of computation time required, even by resorting to lower k values. We therefore re-implemented BestSub to incorporate an evolutionary algorithm (Roitero et al. 2018a). This change has no effect when tested on small cardinalities: both versions of BestSub produce almost completely overlapping and graphically indistinguishable correlation curves. For higher cardinalities, the curves obtained are not distant from interpolating the curves from BestSub. We also needed stable results to work on the percentiles (as we discuss below). For this reason, the average correlation curves are obtained by averaging one million samples in place of 50,000 that was used in past work. Again, this larger sample did not substantially affect the average curves.

Using such heuristic searches means that the best and worst curves are not optimal: there could be topic sets that are even better or worse. However, correlation values should not change significantly, as shown by Guiver et al. (2009, Section 5.1).

3.3 Effectiveness metrics

The MQ07 collection differs from the other collections in that it uses statAP and statMAP (together with E[MAP], that we do not use in this paper), rather than AP and MAP, as its primary evaluation measure. The measure (Allan et al. 2007; Carterette et al. 2009a; Pavlu and Aslam 2007) is a version of MAP that is used to create a pool with a sampling strategy: each document is associated with an *inclusion probability*, used both to decide whether a document is in the pool, and to weight the importance of the document when computing the metric. Since the differences between statMAP and MAP may have implications for our analysis, we consider two approaches for comparing them.

The first is to produce scatter plots showing how the run ranks change when using the two metrics. This has been explored several times, and on different datasets, in previous work, e.g. over AH99 data by Pavlu and Aslam (2007, Figure 7), and over TB06 data by Allan et al. (2007, Figure 5); both analyses showed that while variations exist, they are limited.

A second approach is to compare the correlation curves produced by BestSub when using statMAP and MAP. To do so, we re-evaluated AH99 using statMAP. We selected the 57 runs in AH99 for which the statMAP sampling algorithm does not select any unjudged documents, and used statMAP software from MQ07,[5] thereby implementing "stratified sampling" (Pavlu and Aslam 2007, Section 2.4), where each document has a probability of being sampled that is proportional to its rank in the run outputs. We ran BestSub using both statMAP and MAP. The (best, average, and worst) correlation curves that we obtained for statMAP and for MAP are shown in Fig. 2. The lines are similar, and often overlap or cross each other. In fact the differences are much larger when comparing them with the full AH99 dataset, such as in Fig. 1; this is likely due to the different (smaller) number of runs, and the range of metric values, which have a larger impact than using statMAP in place of MAP. We therefore conclude overall that, although statMAP does create some differences, these appear to be smaller than the differences introduced by other variables, and that using statMAP in place of MAP should not introduce any strong bias into our

[5] Note, several versions of statMAP exist, we used statAP_MQ_eval_v3.pl: http://trec.nist.gov/data/million.query07.html.

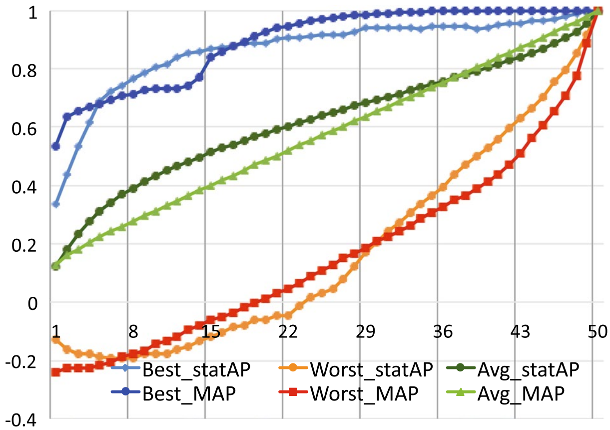

Fig. 2 Kendall's τ correlation curves for AH99, on a subset of runs, for both MAP and statMAP

analysis. This confirms the results obtained by previous studies (Guiver et al. 2009; Robertson 2011; Berto et al. 2013), where the evaluation metric usually did not make any noticeable difference.

4 RQ1: larger ground truth

To address RQ1, we first present a simulation experiment on synthetic data in Sect. 4.1. We then focus on real data starting with an overview of the results in Sect. 4.2, followed by descriptions of best, average, and worst curves in Sect. 4.3. In Sect. 4.4 we compare our results with those by Hosseini et al. (2012) and, finally, the worst sets are analyzed in more detail in Sect. 4.5.

4.1 A simulation experiment

Intuitively, given a larger initial topic set, it will be easier to find good (and bad) subsets, as the degrees of freedom increase. Analogously, when the number of runs in a test collection decreases, it should be easier to find good (and bad) topic subsets, as it is simpler to reorder fewer items in a given way since the size of the gaps between the runs becomes larger and the number of constraints is smaller. To have a first, less qualitative and more concrete, insight on what might happen when varying the number of topics and runs, we perform the following experiments. We randomly generate synthetic AP values for datasets having different sizes of topics (20, 50, 100, 1000) and runs (25, 50, and 100), using two strategies: (1) we generate random AP values normally distributed ($\mathcal{N}(\mu, \sigma^2)$), setting the μ and σ^2 parameters equal to the real μ and σ^2 values of AH99; and (2) we randomly sample with replacement real AP values from AH99 thus obtaining the same distribution of AH99. We then run BestSub on the synthetic datasets to obtain the best, worst, and average correlation values at each topic subset cardinality.

Figure 3 shows the results as correlation charts having the fraction of the full set of topics cardinality on the x-axis and τ on the y-axis. The four charts of the first two rows are obtained by using 50 runs and varying the number of topics. They clearly show that correlation curves become more extreme as the number of topics in the ground truth increases.

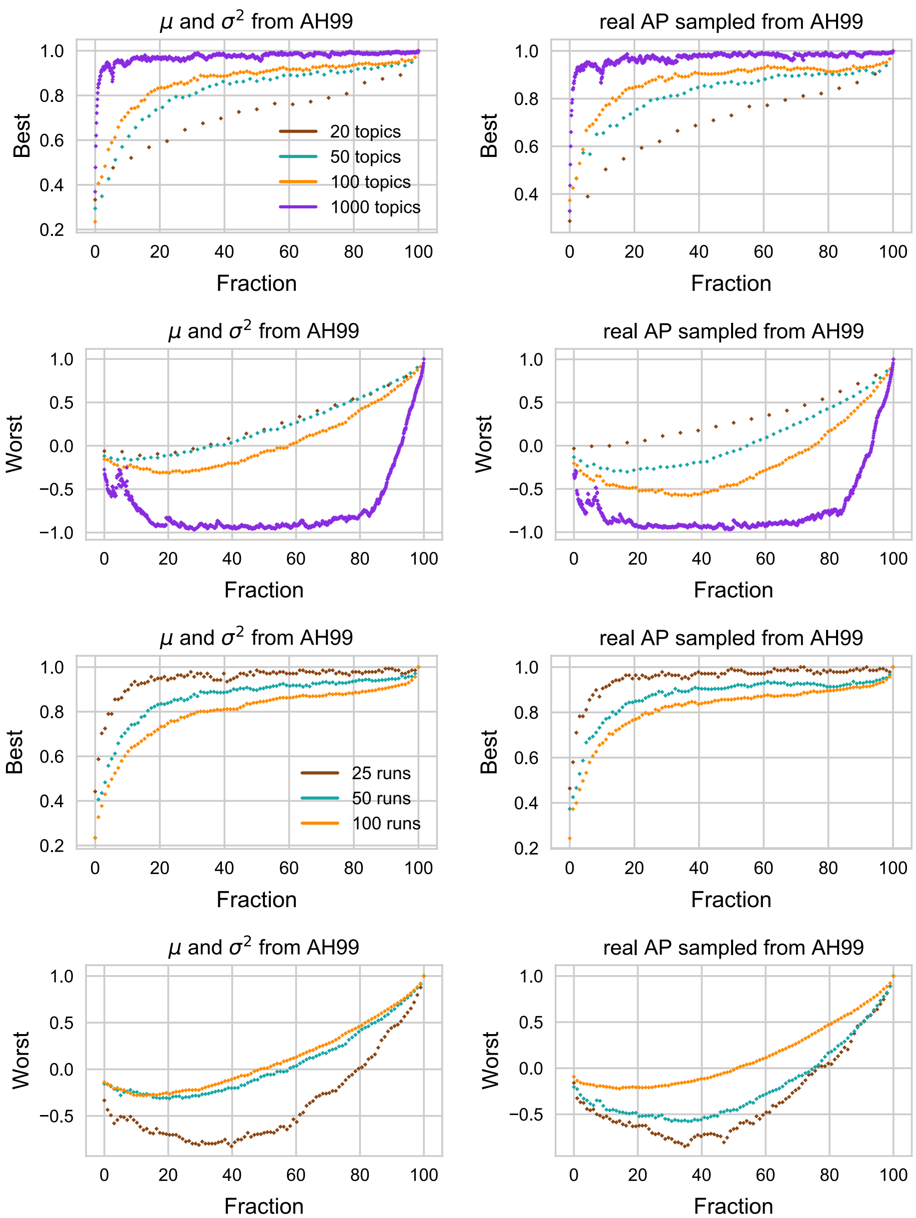

Fig. 3 Kendall's τ correlation curves for fractions of the full topic set, on synthetic data: best and worst curves on random data generated using μ and σ values of AH99 (left column), and on real AP scores sampled from AH99 (right column). In the first four plots, 50 runs are used; in the last four, 100 topics are used

The effect on the average curves (not shown) is less clear but are much smaller as they are quite similar to each other.

When using a fixed ground truth of 100 topics and varying the number of runs, the results are similar. The four charts on the last two rows of Fig. 3 show the correlation curves when varying the number of runs and using 100 topics; the best and worst correlation curves become more extreme as the number of runs decreases, as expected (and the correlation for the average series does not vary much). This is perhaps a less interesting result than the previous one, since the number of topics is related to test collection design and can be decided when building a test collection, whereas the number of runs depends on factors that are more difficult to control. Therefore in this paper we focus on the number of topics. Regardless, this confirms that the number of runs in a test collection can have an effect. Overall, comparing the two sampling strategies (i.e., the left and right columns in Fig. 3) we see that their behaviour is similar, although not identical, when considering a fixed number of both runs and topics.

The results of this simulation experiment hint that the extreme nature of the curves found in previous studies (Guiver et al. 2009; Robertson 2011; Berto et al. 2013) not only is confirmed on datasets with a larger topic set ground truth, but it can even become more striking in some cases. For example, in the worst curve for 1000 topics, even 75% of the topics (i.e., 750) would rank the 50 runs in almost the opposite way to the full topic set. Note that this is a setting similar to MQ07 (see Table 1): if these results were confirmed in the real datasets, they might have important practical implications.

However, the simulation experiment has some limitations: it relies on assumptions that might be not true in a real-world scenario, as different collections have different distributions and parameters, and complex systems topics interactions exist, as shown for example by Urbano (2016) and Urbano and Nagler (2018). For example, the charts on the left in Fig. 3 need to be interpreted with care, as real AP values are usually not normally distributed in practice. When running the Anderson–Darling normality test on each of the four test collections that we use in this paper (see Table 1), the set of all AP (or statAP) values for all topic/run pairs is not normally distributed (neither with $p < 0.05$ nor with $p < 0.01$). When considering the AP values for each single run, the distribution of values is not normal, accordingly to the same test, for 186 ($p < 0.01$) and 219 ($p < 0.05$) cases out of the total 253. By using random AP values we are assuming that the AP values of one run across different topics and the AP values of one topic across different runs are independent, both of which are false as usually the performance of a system across topics is relatively stable, and each research group submits usually many runs, which are somehow related.

Summarising, real test collections include many more variables and interactions than what our simulations can capture: the number of runs, dependencies between runs, the similarity and documents overlap of run variants, the topic system interactions (Urbano 2016; Urbano and Nagler 2018), etc. Moreover, it is of course interesting to see what happens in a real dataset, and in particular if there are particular "pathological" cases that might have occurred. For these reasons we turn to experiments on the real datasets.

4.2 General results

Figures 4 and 5 show correlation charts for the three new datasets TB06, R04, MQ07, as well as AH99 (correlation values are obtained using statMAP for MQ07 and MAP for the other datasets). Correlation is measured using τ. We plot best, average, and worst in separate charts. We also plot the best and worst 1% topic subsets found. In the graphs on the left side of the figures, the x-axis shows subset size measured by cardinality; the graphs on the right, subset size is measured as the cardinal fraction of the ground truth set. The graphs on the left have 250 as maximum cardinality so that we can fully represent the curves for AH99, TB06, and R04. As a consequence, the MQ07 curves are truncated, but their complete trend can be seen in the graphs on the right. To avoid clutter, we do not plot

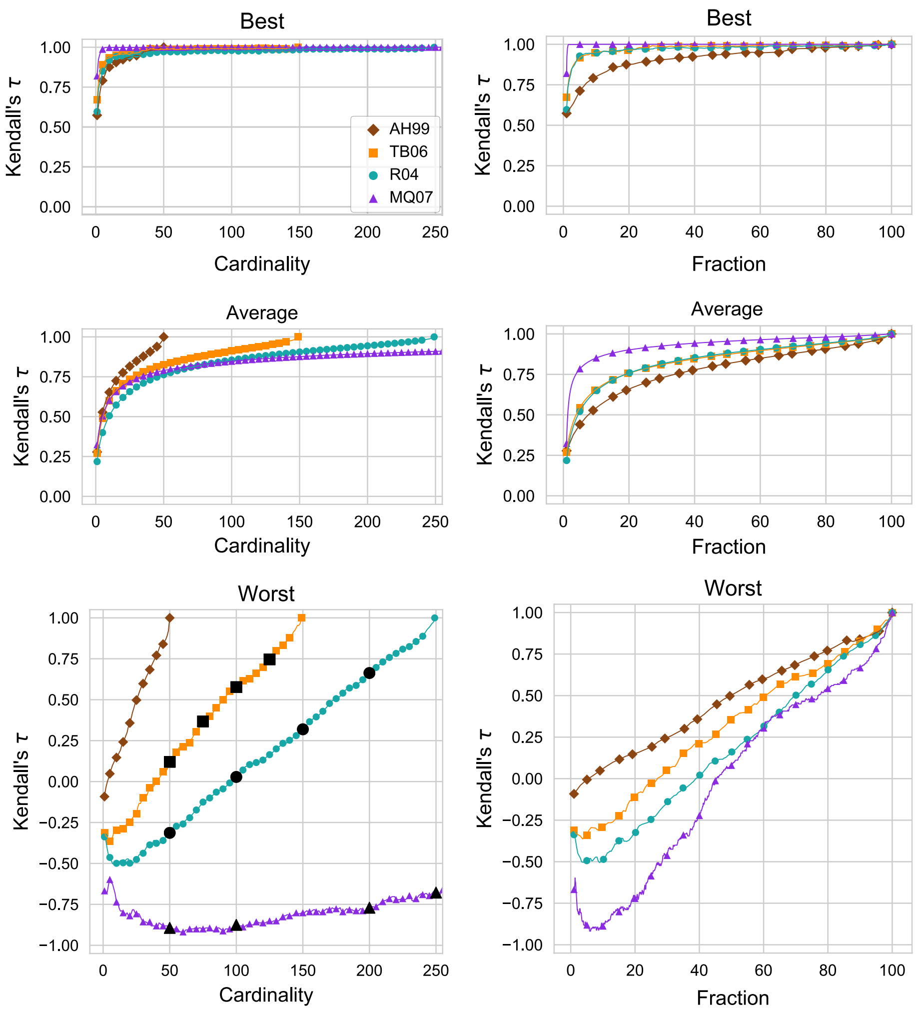

Fig. 4 Kendall's τ correlation curves for absolute cardinalities (left-side, cardinalities up to 250) and fraction of full set (right-side). Black markers in the Worst curves are further analyzed in Fig. 6

the markers for all cardinalities: on the left hand side markers are shown at multiples of 5, plus cardinality 1 and full set. On the right hand side, a marker is plotted at each multiple of 5% (or, when not available because of rounding, the closest value), plus the 1% marker. The lines in the charts are not interpolations, they follow the real values at each cardinality.

While there are similarities between the current charts and those previously published, the best and average curves seem higher when the ground truth cardinality increases (as predicted by the simulation experiment in Sect. 4.1). The worst curves are lower, particularly for the MQ07 dataset. For example, for the MQ07_W[6] curve, a τ of 0 is reached at

[6] We use the suffix B/A/W to indicate the correlation curve for the best/average/worst topic set.

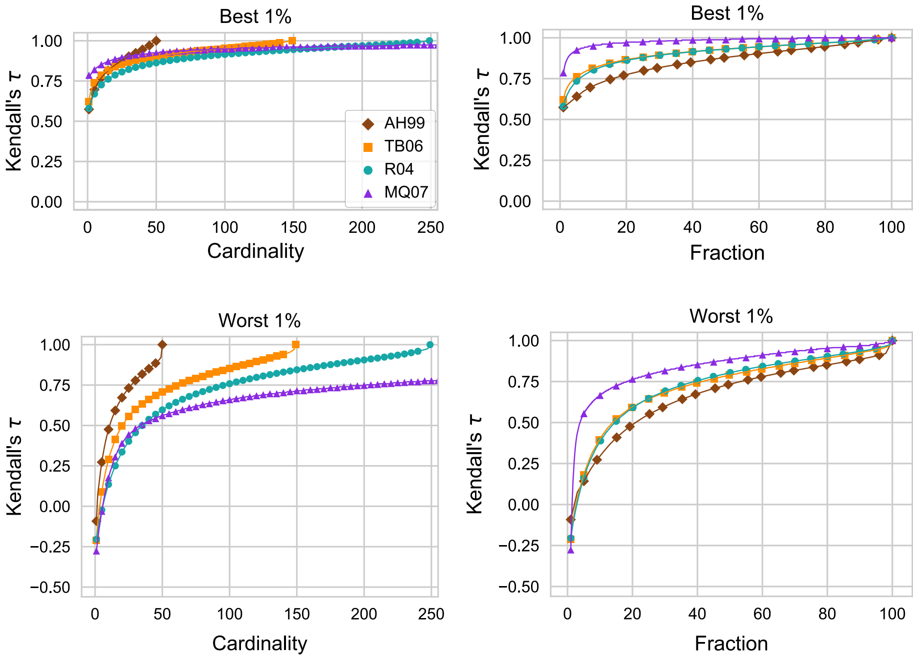

Fig. 5 Kendall's τ correlation curves for absolute cardinalities (left-side, cardinalities up to 250) and fraction of full set (right-side)

around 0.45 of the full cardinality set (around 500 topics) and a τ of 0.5 is reached at 0.8 (around 900 topics). In other words, it would appear that one can build a subset of around 500 MQ topics that ranks the runs randomly, compared to the ground truth. A subset of 900 topics can be found that ranks the runs in a still different way to the ground truth set. We analyze these curves in more detail in the following.

4.3 Best, average, and worst curves

4.3.1 Best correlation curves

From the best correlation curves we see that fewer topics can potentially be used on ground truth cardinalities of $n \gg 50$: the MQ07_B curve is highest, followed by R04_B and TB06_B, which are in turn both consistently higher than AH99_B. This answers the research question RQ1 by supporting, together with the experiment on synthetic data described in Sect. 4.1, the hypothesis that having a larger topic set as ground truth increases the possibility of finding a subset of good topics, thereby leading to higher correlation curves.

A further confirmation of that hypothesis comes from the fraction curves (right-side). Here, the two best curves R04_B and TB06_B are almost exactly overlapping, and they both stay well above the best curve AH99_B. The MQ07_B is even higher. Compared with the previous three studies (Guiver et al. 2009; Robertson 2011; Berto et al. 2013) we see that when using a higher cardinality ground truth (149, 249, or 1, 153 topics in place of only 50), run effectiveness can be predicted by using even fewer topics.

When comparing across the four test collections, it is prudent to examine other properties of the collections that might impact on the trend observed. One can see from Table 1 that as well as a change in ground truth topic cardinality, there is also a change in the number of runs associated with each of the test collections and that this might impact on the τ values.

Sanderson and Soboroff (2007) illustrated that the range of scores that a set of runs have has the greatest impact on τ and other correlation measures. As will be seen in Fig. 6, the range of scores of the runs is similar across the four test collections. However, as discussed in Sect. 4.1, a decreasing number of runs is another factor leading to more extreme curves. In fact, if the goal is to find extreme topic sets, as the number of run increases, there are more runs that need to be re ordered, and the chances of finding extreme topic subsets is lower. Although we leave to future work a complete study of the interplay between the number of topics and of runs, we observe that the effect of the number of topics seems to dominate that of the number of runs, as it can be seen by comparing the worst curves of R04 and TB06 and observing that R04 is clearly the most extreme. The reason is that R04 has more topics than TB06; even if R04 also has more runs (which leads to less extreme curves), this is less important.

4.3.2 Average correlation curves

When examining the average τ across topic subsets, we see that τ for AH99_A is higher than R04_A, TB06_A, and MQ07_A: on average, by selecting a random subset of topics of a given cardinality, this appears to be a better predictor of run rankings in the AH99 dataset than in R04, TB06, and MQ07. Returning to the example in Fig. 1 an average topic subset of cardinality 22 drawn from the collections with larger ground truth has a lower τ than on AH99.

The corresponding fraction curves tell a different story however: on average, by selecting a given fraction of the ground truth, the topic subset of AH99 turns out to be a worse predictor of run rankings than that of R04, TB06, and MQ07. Collections with larger ground truths appear to need a smaller fraction of topics to achieve high values of τ.

A particular feature of the MQ07_A curve is that its trend seems more similar to the best than to the average curves of the other datasets. For this dataset, on average, a good prediction of run ranks can be obtained with a small fraction of topics (around 5–10%) and a very good prediction of run ranks can be obtained with 20%. This result needs to be examined on other test collections with similarly large topic sets.

The curves for R04 and TB06 on the fraction charts are almost exactly overlapping. This might be an indication that a ground truth of cardinality 50 is somewhat different from a larger ground truth. There might be some numerical/statistical effect that does not appear when using only 50 topics.

4.3.3 Worst correlation curves

The most noticeable difference between AH99 and the larger datasets is in the worst curves: whereas best and average are broadly similar to past work, the worst curves are quite different.

The correlation values for the worst curves are strongly negative. This is a novel situation, not observed in the previous three studies (Guiver et al. 2009; Robertson 2011; Berto et al. 2013) where τ values were at worst negative with a low absolute value (around -0.2). Negative correlations show topic subsets that evaluate runs in broadly opposite ways. Also, the negative correlation values in R04_W persist for cardinalities much larger than 50, the usual number of topics used in evaluation exercises. The MQ07_W curve is even lower and stays below -0.5 up to 250 (and, as can be seen from the fraction curve on the right, even up to 300).

Although this is something expected after the simulation experiment in Sect. 4.1, it is somehow striking that on MQ07, a subset of more than 250 topics can be found that negatively correlates with the ground truth topic set. As mentioned above, a set of around 45% of the MQ07 topics (around 500 topics) results in a τ of zero.

Note, the reason three of the curves drop as the cardinality of the topic sets increase from 1 is due to the degrees of freedom there are when searching for topic subsets that are the worst: the value of $\binom{n}{c}$ initially increases as c gets larger. Therefore, the range of possible topic sets that are searched to find the worst also gets larger.

4.3.4 Best and worst first percentile curves

Given the extreme nature of the best and worst curves, we also computed the average τ of the best and worst 1st percentile of topic subsets. Figure 5 shows the resulting charts.

The Best 1% curves emphasize that although the quest of finding the best topic subsets is rather difficult since they are extremely rare, reasonably good results that can more easily be obtained in practical cases do exist. The Worst 1% curves are less worrying than the Worst ones, since they do not feature the same extremely low, if not negative correlations. Although these curves look more like those from the Average, it is worth noting that when trying to find subsets of topics for an effective test collection, a low positive correlation is not satisfying either. For example, the R04 1st percentile curve has low τ (< 0.6) even for cardinality 45, and the MQ07 1st percentile curve has a τ of about 0.75 at cardinality 250. These are not extremely unlikely topic sets, and it is possible that some test collections have been created with topics that rank runs quite differently from what might be expected.

4.4 Comparison with Hosseini et al. (2012)'s results

Hosseini et al. (2012) report in their paper some numeric correlation values for the AH99 and R04 collections to which we can compare. Since Hosseini et al. use all the runs in a collection, for this comparison we performed again our experiments using all the runs instead of the top 75% (thus, 129 for AH99 and 29 for MQ07, see Table 1), and all the values reported in this section concern such a setting.

Table 2 shows the results of the comparison. The first five rows in the table report Kendall's τ correlation values obtained for best, best 1st percentile, average, worst first percentile, and worst subsets at the specified fractions (20%, 40%, and 60%) of the full set cardinality, for the two collections AH99 and R04. Since we are using all the runs in this computation, the results do not exactly match with those presented in previous Figs. 4 and 5 . The next three rows in the table are the values reported in Hosseini et al. (2012, Table 1): "Oracle" is their attempt to find the highest possible correlation, and so it somehow corresponds to our Best topic subsets; "Random" is a random selection of

Table 2 Kendall's τ values for comparison with Hosseini et al. (2012)'s results. All runs used

	AH99			R04		
	20%	40%	60%	20%	40%	60%
Best	0.92	0.96	0.97	0.97	0.98	0.99
Best 1%	0.86	0.91	0.94	0.91	0.95	0.97
Average	0.80	0.87	0.91	0.85	0.91	0.94
Worst 1%	0.70	0.81	0.87	0.76	0.86	0.90
Worst	0.48	0.64	0.77	0.17	0.43	0.63
Oracle	0.88	0.93	0.95	0.90	0.92	0.94
Adaptive	0.83	0.90	0.93	0.77	0.87	0.91
Random	0.72	0.77	0.87	0.68	0.80	0.85
Clustering	0.79	0.88	0.93	0.84	0.92	0.95

topics, so it should correspond to our Average; and "Adaptive" values are those obtained by their topic selection algorithm. The values in the last row of the table will be discussed when focusing on RQ3 on clustering in the following.

We can draw several remarks.

- When comparing the correlation values in the first five rows of the table with those obtained on the top 75% runs (Figs. 4 and 5), it is clear that the correlation values obtained using all systems are higher. This is expected, as the bottom runs are usually consistently ineffective on all topics. In other terms, focusing on the top systems only as we are doing in this paper is a more difficult setting for our task than using all systems.
- When comparing our Best with Hosseini et al.'s Oracle, we note that Best values are always higher than Oracle. Indeed, Oracle is always closer to Best 1% than to Best, and for R04 it is even closer to Average than to Best.
- When comparing Average with Random, we expected no differences, but it turns out that some clear differences exist. Our Average values are clearly higher than their Random. Indeed, Random is always closer to, and often lower than, Worst 1%. We have not been able to obtain the original code used by Hosseini et al. to replicate their experiment and although we tried, we could not obtain their values. We double checked and we are quite confident that Average values are correct, for several reasons: they correspond to the values in previous work (Guiver et al. 2009; Robertson 2011; Berto et al. 2013) obtained with the different implementation of BestSub (see Sect. 3.2); the new implementation of BestSub is publicly available (https://github.com/Miccighel/NewBestSub) and it has been flawlessly used in other experiments (Roitero et al. 2018b, a), including some specifically aimed at reproducing previous results (Roitero et al. 2018a, Section 4.3). As a further check, we also reproduced the random series of the plots in (Kutlu et al. 2018, Figure 1) for two datasets (Robust 2003 and 2004 reduced to 149 topics): also in this case our average values correspond to Kutlu et al. random ones.
- When looking at the Adaptive values (that will be further analyzed in the last part of this paper), one can notice that Adaptive is clearly higher than Random (the baseline used by Hosseini et al.) but, as a consequence of the previous remark, it is very simi-

lar to Average for AH99 and even always lower than Average for R04. Therefore, it turns out that Adaptive is not effective when compared to our, higher, baseline.

4.5 Worst subset analysis

Although exceptionally rare, the very worst topic subsets feature extremely low correlations. In this section we try to better understand how the subsets produce such low τ correlations.

4.5.1 Overlap

Examining intersections between the best and worst topic subsets, we find that there is a quite large overlap between them: at cardinality 100, R04 and MQ07 have a topic overlap of around 40%. This means that it is possible to select a set of 40 topics, then to add to it either a first or a second set of 60 (different) topics, and obtain completely different, even almost opposite, rankings of runs.[7]

A possible explanation for this overlap could be that there are two small subsets of topics, one good and one bad, that are used to build the low cardinality best and worst sets; then a set of common "neutral" topics are added to both to obtain the higher cardinality sets. However, this needs further study, as this possibility is not consistent with the data, since the 40% overlap can be found from cardinality 50 up to 200.

4.5.2 Comparing worst with best

It is also possible that some conceptual features of the topic subsets exist that could explain the low correlations. Therefore, some of the worst topic subsets are characterized here for analysis. We manually selected illustrative topic subsets that have low τ correlations and high cardinalities:

- TB06: cardinalities 50, 75, 100, and 125.
- R04: cardinalities 50, 100, 150, and 200.
- MQ07: cardinalities 50, 100, 200, and 250.

These are the subsets represented by black markers in Fig. 4. Figure 6 shows scatter plots for these subsets. We see that the effectiveness measure computed on the worst subset (y-axis) usually has both a smaller range and lower values when compared to the measure computed on the ground truth set (x-axis). This is especially true for MQ07, but the same effect can also be found on the other datasets. To better understand this observation, the mean effectiveness over all topic subset cardinalities was computed for the best and worst topic subsets. The results are shown in the left part of Table 3. It can be seen that the best subsets contain topics that lead to higher effectiveness values than the worst subsets. The right part of the table shows the Δ between the average subset effectiveness and the ground truth effectiveness. As might be expected, in all cases the best subsets contain topics that lead to values more similar to ground truth effectiveness.

[7] Note, the overlap that we find might be an effect of the heuristic used; we can say no more than it is possible to build a best and a worst set of topics with a high overlap.

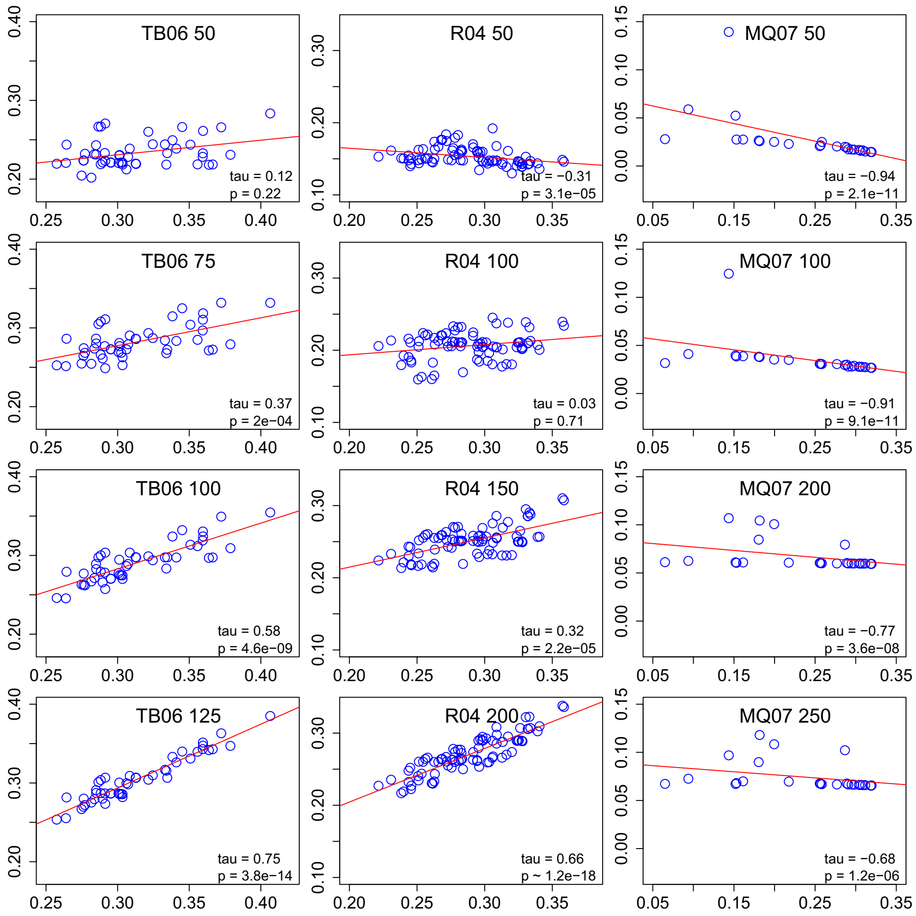

Fig. 6 Scatter plots for some selected notable worst topic subsets. Each dot is a run, the x-axis shows MAP (statMAP for MQ07) computed on the ground truth full topic set, the y-axis shows effectiveness computed on the worst subset indicated. The τ of the correlation along with the significance of the correlation (indicated by a p value) is detailed on each plot

Table 3 Effectiveness measures (MAP, except statMAP for MQ07) over the best and worst subsets and between ground truth

	Av. effectiveness of subset				Subset Δ from ground truth			
	AH99	TB06	R04	MQ07	AH99	TB06	R04	MQ07
Best	0.298	0.277	0.264	0.148	0.017	0.036	0.025	0.092
Worst	0.201	0.263	0.224	0.049	0.080	0.050	0.066	0.191

5 RQ2: statistical significance

We now turn to RQ2. While the previous results demonstrate that it is possible to find topic subsets that lead to run rankings that are highly correlated with the rankings obtained when using a full (ground truth) set of topics, in order for one run to be considered more effective than another, a statistical significance test is usually carried out. The number of topics that are used to evaluate effectiveness has a direct impact on significance calculations. For example, for a paired t-test, the test statistic includes the sample size (Sheskin 2007), and the larger the sample, the lower the p value. In IR experiments the sample size is the number of topics. Some analysis of statistical significance is therefore due in the fewer topics scenario. We present two different and somehow dual approaches to do such an analysis in the next two subsections: the first approach is based on the work by Sakai (2016b) that determines the number of topics needed when aiming at a given statistical power; the second is aimed at determining the amount of error that is introduced when using topic subsets, as well as at understanding what kind of errors are made.

5.1 Power analysis

Sakai (2016b) recently proposed three methods to compute the cardinality of a topic set size to ensure that a test collection has enough statistical power to distinguish effectiveness of the systems/runs. The methods compute the estimated topic set size on the basis of three different tests:

- Method 1, based on t-test, and used when one wants to compare two system scores, or the score of one system against all the other systems.
- Method 2, based on one way ANOVA, and used when one wants to compare the scores of more than two systems, or to compare all systems against each other.
- Method 3, similar to Method 1, but it allows one to specify a confidence interval δ to ensure that the outcome of this test is bounded with precision δ. As Sakai points out: "a wide confidence interval that includes zero implies that we are very unsure as to whether systems X and Y actually differ".

We computed and/or estimated the parameters required by Sakai's methods and ran them on our four collections, using the software (Excel spreadsheets) provided by Sakai. Tables 4, 5, and 6 show the results.

The topic set size cardinalities in Table 4 are those required to find statistical significance when comparing two systems, or a system against a set of other systems (e.g.

Table 4 Number of estimated topics using the first method, based on t-test. The required σ_t^2 parameter has values 0.096 (for AH99), 0.071 (TB06), 0.100 (R04), and 0.118 (MQ07). The values in bold represent the maximum and minimum estimated number of topics for the given parameters, for each collection

α	β	AH99			TB06			R04			MQ07		
		$minD_t$			$minD_t$			$minD_t$			$minD_t$		
		0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
0.01	0.1	**575**	147	40	**426**	109	30	**599**	153	41	**706**	179	48
	0.2	452	116	32	336	87	25	471	121	33	554	142	38
0.05	0.1	406	103	28	301	77	21	422	106	29	498	126	33
	0.2	304	78	**21**	225	58	**16**	315	81	**22**	373	95	**26**

Table 5 Number of estimated topics using the second method, based on ANOVA.The required σ^2 parameter has values 0.048 (for AH99), 0.036 (TB06), 0.050 (R04), and 0.59 (MQ07). The values in bold represent the maximum and minimum estimated number of topics for the given parameters, for each collection

α	β	AH99			TB06			R04			MQ07		
		minD			*minD*			*minD*			*minD*		
		0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
0.01	0.1	**2352**	588	147	**1341**	336	84	**2295**	574	144	**3446**	862	216
	0.2	2001	501	126	1131	283	71	1948	487	122	2879	720	180
0.05	0.1	1860	465	117	1050	263	66	1810	453	114	2669	668	167
	0.2	1529	383	**96**	855	124	**54**	1485	372	**93**	2151	538	**135**

Table 6 Number of estimated topics using the third method, based on confidence intervals. The required σ_t^2 parameter has values 0.096 (for AH99), 0.071 (TB06), 0.100 (R04), and 0.118 (MQ07)

α	δ	AH99	TB06	R04	MQ07
0.01	0.05	1019	754	1061	1253
	0.1	255	189	266	314
	0.2	64	48	67	79
0.05	0.05	591	437	615	726
	0.1	148	110	154	182
	0.2	37	28	39	46

when trying to understand if system s_1 is better than both systems s_2 and s_3). The three parameters are: α, which is the probability of Type I error (to find a difference that does not exist; that is, one concludes that s_1 is more/less effective than s_2 but this is not true); β, which is the probability of Type II error (not to find a difference that does exist; that is, one does not conclude that s_1 is more effective than s_2 when it is in fact better); and $minD_t$, which is the minimum detectable difference in MAP values. We use the same values for these three parameters as adopted by Sakai in the examples in his paper. $minD_t$ is computed considering the estimated within-system variance from past collections, σ^2. To compute σ^2 we used, as Sakai suggests, Formula (36) of Sakai (2016b), that is the residual variance from one-way ANOVA computed considering all the AP values for a given collection (i.e., all the systems and all the topics): we applied Formula (36) to our collections when using the AP (statMAP for MQ07) metric. As discussed by Sakai (2016b, Section 3.2), σ^2 represents the common system variance computed under the so called homoscedasticity assumption, which means that σ^2 is considered to be common for all the systems. Carterette (2012) shows that this assumption does not hold for IR evaluation, and discusses how this fact is not important; indeed, as remarked by Sakai (2016b, Footnote 16), ANOVA is widely used in the IR field.

The values in the table (the estimated required number of topics) range from 16 to 706. Besides the considerations that could be made on the values of the three parameters α, β, and $minD_t$ (probabilities of Type I and II errors, and the minimum detectable difference), what is important to note for our purposes is that quite often the required number of topics is even higher than the cardinality of the full topic set size for the corresponding collection.

This is even more true when using the second method (based on ANOVA), see Table 5: in this case values range from 54 to 3446. The parameters for this method have a similar meaning to the previous method based on the t-test, with some technical differences. It is important to notice that the estimates obtained with the second method are probably more related to the approach in this paper, since we generally compare all the systems together, rather than a single system to the other ones. The third method returns intermediate results (see Table 6).

This analysis led to reappraising the results on the best correlation curves: whereas it is true that small good topic sets exist, using them would, unsurprisingly, lead to less statistical power (which is defined according to Sakai as $1 - \beta$, and represents the capability of finding a difference between system scores which is statistically significant), or in other words it is a move away from the number of topics required to have such statistical power.

We note that this approach (Sakai 2016b) does not directly quantify how much statistical power we are losing when using the smaller good topic sets. In future work we intend to further explore the relationship between the factors of (sub-)topic set size and quality, and statistical power. Moreover, this method does not consider what kind of errors are made: when using fewer topics, there are different possible specific outcomes besides the result of a statistical test: one might find significance for a sub-set while according to the full set of topics there is not, or vice versa one might not find significance for a sub-set while for the full there is; one might even find statistically significant disagreement; and so on. For these reasons, we conducted another, more general, experiment, described in the following subsection.

5.2 Statistically significant agreement and disagreement

We conduct an empirical investigation into the relationship between the number of topics considered in an IR experiment and the observed outcomes of statistically significant differences between runs. We first discuss some methodological issues and then describe our experimental results.

5.2.1 Methodology

Consider a typical IR effectiveness experiment, where a researcher is seeking to demonstrate that one retrieval approach is superior to another. The researcher chooses a test collection consisting of (say) 50 topics, and generates two sets of 50 effectiveness scores (two runs). If the mean score for one run is higher than that for the other, it is standard to carry out a significance test such as a paired t-test. This will indicate whether the two scores are indeed likely to come from populations with different means, at some specified level of confidence.

We are interested in investigating the question: if the researcher had carried out the same experiment but with a subset of topics, would the same results have been observed? This is somehow related to a similar question that has been investigated by Sakai (2007), who studied the effect of collection incompleteness on the discriminative power using Sakai's bootstrap sensitivity method; however, we focus on subsets of topics rather than subset of documents. More concretely, let us consider a test collection with a ground truth set, G, of topics of cardinality b. Let there also be a subset of topics, S, with cardinality a, where $a < b$. For a pair of runs X and Y, calculate their MAP using topic set S, and carry out a paired 2-tailed t-test to determine whether the difference is statistically significant. Repeat the process for the same pair of runs, but using the topic set of full cardinality, G. There are five possible outcomes (Moffat et al. 2012):

- SSA: run X is significantly better than run Y on both topic sets, S and G.
- SSD: run X is significantly better than run Y on one topic set, but Y is significantly better than X on the other topic set.
- SN: one run is significantly better than the other on topic set S, but there is no significant difference on topic set G.
- NS: one run is significantly better than the other on topic set G, but there is no significant difference on topic set S.
- NN: there is no significant difference between the runs on either topic set.

The first two letters of each label indicate the outcome of the experiment (Significant or Not significant) on topic set S and G, respectively, while A and D stand for Agreement and Disagreement, respectively. Note that only two of the five outcomes, SSA and NN, are cases where consistent conclusions would be drawn from the experiments regardless of which topic set is used. For the other three, a researcher who happened to use a topic subset S would reach a different conclusion about relative run effectiveness, than if they had used the ground truth G.

When considering topic subsets, it is desirable to maximize the number of SSA and NN cases (SSA if the researcher is looking for a publishable result), and to avoid SN and NS cases (where significant differences are found with one topic set but not with the other) and in particular SSD (where significant differences are found with both topic sets, but with different runs being indicated as being better).

5.2.2 Results

The results of the simulated experiments are shown in Fig. 7 for the four collections (columns), and for the best, random,[8] and worst subsets (rows). For each sub-figure, the x-axis shows the cardinality of topic set S, which is being compared to the full cardinality ground truth, topic set G. The y-axis shows the proportion of occurrences for each of the five experimental outcomes: SSA (blue), NS (green), SN (yellow), SSD (red) and NN (orange). It can be observed that when the subsets reach their maximum cardinality (on the right of the plots), only two outcomes are possible, SSA and NN. This must be the case, since at full (ground truth) cardinality S and G are the same, and so the outcomes of the two experiments are identical. (Recall that in the MQ07 collection, subsets do not reach full cardinality by 250.) When the full topic sets are used, SSA is dominant, accounting for around 55–70% of cases. This is reassuring, since it shows that using the full collections, it is possible to statistically distinguish between the runs more often than not.

The figures also clearly confirm that the larger the cardinality, the higher the likelihood that the same significant evaluation results will be observed as when using the ground truth topic set. The NS class shows the cases where a significant difference would be found between two runs using G, but no corresponding difference is found when using S. For these cases, the reduction in topic set cardinality has compromised the ability of the

[8] Here, for speed of calculation reasons only a single random topic subset is drawn from the set of all topic subsets of a given cardinality. The histograms of random are consequently more "spiky" than if we averaged several random subsets. However, the broad signal of the result is still visible in the plots.

significance test to identify significant effects, a false negative. Moreover, when comparing the charts "horizontally", the NN orange areas decrease with the full cardinality of the dataset, in both best and random, while the SSA blue areas increase. As expected, results tend to be statistically significant more often when the full cardinality of the ground truth is higher.

Considering the best, random, and worst topic subsets, over all four collections, as cardinality increases, the best subsets lead to the most rapid maximizing of the SSA class (and quickest reduction of the NS class), though on the MQ07 collection, the best subset is only somewhat better than random. The best subsets, besides allowing to use fewer topics in evaluation, also lead to finding SSA results (both in agreement with the ground truth and statistically significant) more often than random topic subsets.

The worst subsets lead to experimental significance results that have the least correspondence with the ground truth topic set. Perhaps unsurprisingly, the heuristic that selected the best and worst subsets, which was optimized to maximize and minimize τ respectively, also maximized and minimized on significance.

The size of the SN category (false positive) is generally very small—there are few cases where significant differences are detected on S while no significance is found on G.

The most problematic case, SSD, where one run is significantly better than another on topic set S, while the other run is found to be significantly better on topic set G, is fortunately rare, although it should be noted that it is possible, for all collections except AH99, to find a (worst) subset of topics of rather high cardinality that would lead to such contradicting results. In particular, for the MQ07 collection, even by cardinality 250, for the worst subset, the proportion of SSD cases is substantially higher than SSA cases. In other words, the worst chart for MQ07 shows that we can not only generate a topic subset of cardinality 250 with strong, and statistically significant, negative correlation with the full set (as already shown by the MQ07 series in the worst plots in Fig. 4 and, more in detail, by the last plot of Fig. 6), but moreoever that the "aberrant" subset of MQ07 topics of cardinality 250 would also feature a very small amount of SSA and NN, and many NS and even SSD. Experiments using that subset would lead a researcher to derive a statistically significant result that is very different from the full set. Whether this is a temporary manifestation, or is maintained into higher cardinalities, needs to be investigated in future work. However, it must be noted that best, random, and worst charts for MQ07 are consistent with the other datasets (once the fact that we do not reach the full cardinality for MQ07 is taken into account).

By and large the NN cases stay constant over the best and random topic subsets and there is only some variation of these on the worst subset.

5.2.3 Conclusions

Overall, this analysis demonstrates that while it is possible to find a subset of topics that lead to run effectiveness rankings that are highly correlated with rankings from a ground truth set, a side effect of doing so is that a researcher is sacrificing the ability to identify statistically significant differences between runs.

An experimenter using a topic subset in general does not risk having to deal with false positive significant results, however, they do risk having a number of false negatives in their experiments. As seen in the ratio of SSA to NS in the plots, the magnitude of the problem reduces as the subset cardinality increases. Indeed, many experimenters might

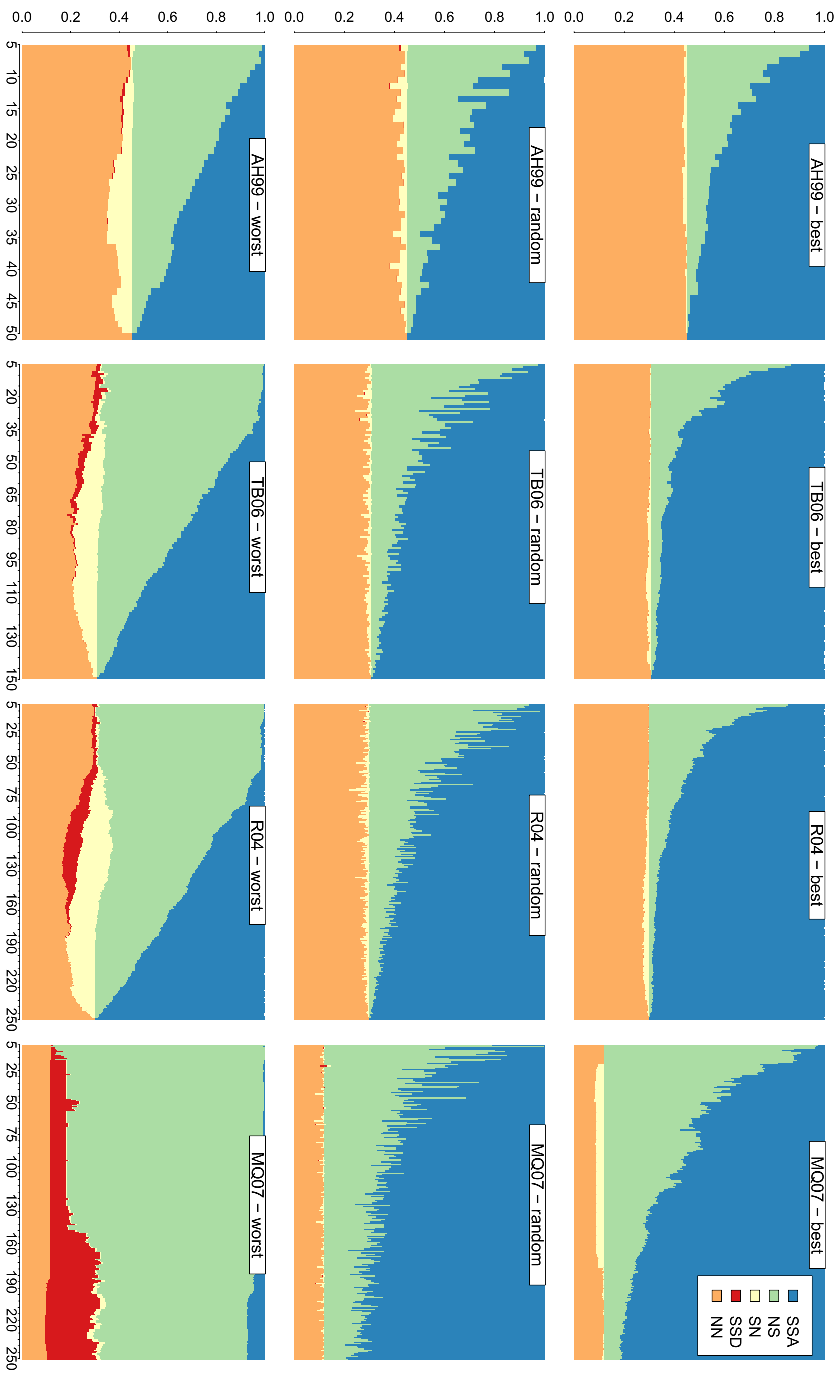

Fig. 7 The results of typical IR effectiveness experiments, showing the proportion of cases where statistically significant differences (two-tailed paired t-test, $p < 0.05$) are observed between two runs when comparing them using a reduced cardinality (shown on the x-axis) compared to the full topic set for a collection

view a small amount of NS acceptable if it means they can build their test collection more quickly using fewer resources.

Perhaps more worryingly, it is difficult to state that the topics used in IR test collections are sampled randomly and independently from the population of all topics: TREC topics are created by analyzing the document collection and by selecting those topics that, for example, guarantee a minimal number of relevant documents. The bias introduced by such a process is far from being clearly measured. Therefore, one might question the general applicability, in IR evaluation, of statistical tests which usually require specific conditions, and how much confidence one should attach to such results in terms of estimating the generalizability of experiments to larger topic sets.

5.2.4 Caveat

It should be noted that this simulation of typical IR experiments includes a large number of pairwise significance tests. One might therefore argue that corrections for multiple testing, such as the Bonferroni correction (Feise 2002), should be applied. However, while individual researchers might use such corrections from time to time, the fact is that IR test collections are used again and again, often to compare against standard baselines, and there is no way of knowing what corrections should be made to account for all (reported as well as unreported) tests that are ever carried out by the population of IR researchers as a whole. Not applying multiple comparison corrections is therefore a more accurate simulation of the typical IR experimental environment. This choice is also supported by Carterette (2012), who argues that it is not clear how to properly correct values in a TREC-like setting, or whether it should be done at all. In this respect, it has to be remarked that in the second method of Sect. 5.1 the variance estimates are indeed computed applying the correction method for multiple comparisons.

Finally, we note that the *t*-test is the most widely used statistical test in IR experiments (Sakai 2016a); however, we also repeated the experiment using the Wilcoxon signed rank test instead of the *t*-test, and the trends were consistent.

6 RQ3: topic clustering

We now turn to RQ3. As already stated in Sect. 1, it seems intuitive that by (1) clustering the topics, and (2) selecting representatives from each cluster, the topic set obtained should be more representative of the full ground truth than an average or random topic subset of the same cardinality. Furthermore, the selection of a subset of representative queries has been proven to be effective in a Learning to Rank scenario (Mehrotra and Yilmaz 2015), and indeed the clustering of topics approach follows the same principles as it is clearly based on the representativeness notion. Therefore, a topic clustering process should be an effective strategy to find good topic subsets. However, such a process could involve many different settings. We present several approaches, their results, and a discussion on clustering effectiveness.

We start by presenting in Sect. 6.1 the overall experimental setting. We then discuss two possible approaches: the first in Sect. 6.2, that will is not effective despite attempting many variants, and the second, in Sect. 6.3, which is slightly more effective than the first. Section 6.4 discusses the clustering approach.

6.1 Experimental setting

We start by defining the experimental settings and notation that are common to the experiments described below.

6.1.1 The clustering process

We denote with n the number of topics and with $c \in \{1, \dots, n\}$ the cardinality of the topic subset; also, $m \in \{2, \dots, n\}$ is the number of clusters obtained when performing a clustering process. Our method is composed of the following three steps.

1. For each cardinality c, we build a set of m clusters.
2. Then a topic subset of cardinality c is formed by selecting random representatives from each cluster. In the following we refer to this selection method as *one-for-cluster* (note that one might devise different methods, e.g., selecting a number of topics proportional to cluster size, selecting from some clusters only, etc.)
3. Finally, we build the usual correlation curves, and we compare the one-for-cluster series with random topic selection, which is the average series (such as the ones represented in Fig. 4, left-side, second row).

We use a standard, effective clustering algorithm, *hierarchical* clustering with a *complete linkage* method, and the *cosine similarity* as the distance function. We also try variants, as specified below. We conduct 10,000 repetitions to compute the one-for-cluster series, to avoid noise.[9] Note that we are only considering clustering as the main analysis technique. We leave as future work more complex machine learning approaches, that could make use of multiple features such as for example the μ and σ^2 parameters from Sect. 4.1.

6.1.2 Feature space

We take as topic features the AP (or statAP) values over the run population, by clustering topics in a multi-dimensional space, where each dimension is the effectiveness on a specific run, and each topic is a vector of AP (statAP) values. The idea is that topics that have similar AP values for all runs are redundant: one topic should be as effective as all of the "similar" ones. Clustering should group together those topics that have similar scores, and by picking representatives from each cluster we should select a good topic subset. For each dataset, the number of dimensions is therefore the number of runs (the last column in Table 1). We also experiment with a variant of this approach, as detailed below.

6.1.3 Number of clusters and topic cardinality

We can think of two possible overall settings, that affect Steps 1 and 2 above. For each cardinality c and number of clusters m:

– We can perform clustering with the constraint $c = m$; we refer to this setting with the term *cardinality-driven clustering*;

[9] We tried with up to 1 million repetitions, but the series are already stable with 1000 repetitions.

– We can determine the number of clusters a priori, independently from c, and subsequently select the topic subset; we refer to this setting with the term *cardinality-independent clustering*.

Both settings have pros and cons. The first approach forces the clustering algorithm to produce a clustering of exactly $m = c$ clusters, which might be unnatural for certain c values: for example, if the topics are naturally form two clusters, forcing them into three will produce clusters that are less complete and more heterogeneous, thus potentially of lower quality. However, once the clusters are formed, the selection of topics is straightforward, since there is the guarantee that when c topics are to be represented, there are exactly c clusters. Furthermore, even if the $c = m$ constraint might lead to unnatural clusters for certain c values, in general just decent clusters, even if not perfect, might be of a sufficient quality to guarantee higher correlation values for the one-for-cluster series than for random topic selection.

Conversely, with the second setting, the topics can be clustered in a more natural way, but then the selection process is slightly more complicated: there is no equivalence between the number of clusters and the number of topics to select, thus there is not a unique selection method, and the selection process has to take into account the empty clusters that might occur during the process. Finally, whereas with the first setting the choice of the number of clusters m is straightforward and determined, with the second setting m is a parameter to be chosen, and it is not clear which criteria should be used. In the following Sects. 6.2 and 6.3 we analyze both settings, starting with the first one.

6.2 Cardinality-driven clustering

6.2.1 A first attempt

We compute the clustering as described above, with the constraint $c = m$; then, we compare the one-for-cluster with the average series. It is found, however, that this clustering of topics approach does not result in any topic subset having a τ correlation higher than the average; indeed usually τ is lower. There are multiple possible explanations for this behavior. First is the choice of clustering algorithm. Therefore, we tried different variations of the clustering, for example, using a non hierarchical algorithm such as K-means (with the algorithm variations Hartigan-Wong, Lloyd, and MacQueen[10]), and/or using different distance functions (including as different kinds of proximity measures[11] both linear metrics, e.g., Euclidean, Manhattan, Divergence, etc., and similarity-angular distances, e.g., Cosine, Correlation, Jaccard, Phi, etc.), or using different methods to join clusters (thus different linkage techniques including single, average, mean, median, Ward). However, τ was never found to be higher than average for any of these clustering methods, and we can be confident that these negative results are not affected by a particular clustering setting.

A second possible explanation is related to the feature vector: our feature vectors are in a high-dimensional space, and therefore most of the distances tend to be similar, and vectors

[10] See the R function "kmeans" in the "stats" package (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html), and "k-means" of "scikit-learn" for Python 3 (http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html).

[11] For an exhaustive list see the R package "proxy" (https://cran.r-project.org/web/packages/proxy/proxy.pdf), and the "Distance computations" section of Python 3 (https://docs.scipy.org/doc/scipy/reference/spatial.distance.html).

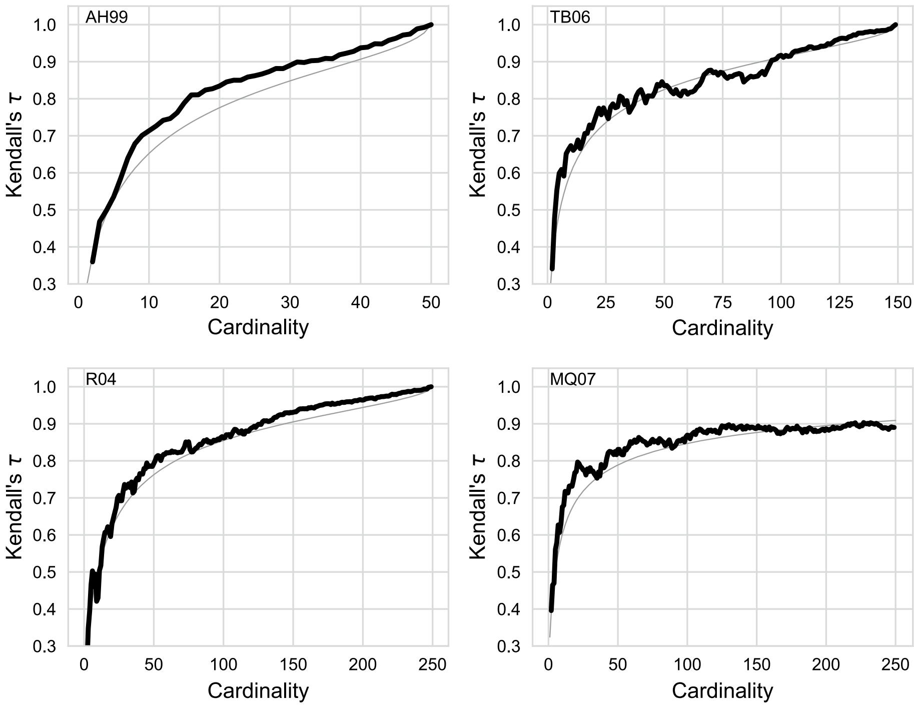

Fig. 8 Kendall's τ correlation curves for the four datasets: averaged (thinner gray lines), and obtained using clustering (thicker darker lines)

tend to be orthogonal. To be more precise, as soon as the number of dimensions grows, the number of possible distance values drops. This is a well known phenomenon, referred to as "the curse of dimensionality" (Rajaraman and Ullman 2011, Chapter 7), and it occurs for both linear and angular distance values (Cosine, Euclidean, and Manhattan). This could of course harm the clustering process. To address this limitation we tried to combine clustering with dimensionality reduction, as described in Sect. 6.2.2. Finally, in this setting we have the constraint that the number of clusters m must be equal to the topic subset cardinality c, and that could lead to forming unnatural clusters, as already mentioned; we discuss this third possible explanation in Sect. 6.2.3.

6.2.2 Dimensionality reduction

To deal with the curse of the dimensionality effect, a second attempt makes use of Principal Components Analysis (PCA). To express around 85–90% of the total variance of the data, three components/dimensions are needed for AH99, R04, and TB06, and five for MQ07. Each topic vector is then heavily reduced, to very few components: from the values in the last column of Table 1 to 3, 3, 3, and 5, respectively. We then repeat the clustering process with the same primary settings as above ($c = m$, hierarchical algorithm, cosine distance, and complete linkage).

With PCA, the results are different to clustering. Figure 8 compares the correlation curves for average subsets, which are gray and thin in the figure, with the correlation curves obtained with the one-for-cluster method: the latter are usually above the former. Moreover, the differences between one-for-cluster and average correlation values are statistically significant for most of the cardinalities: in around 90% of the $50 + 249 + 149 + 250 = 698$ total cases for the four collections, the difference is statistically significant according to the Wilcoxon signed rank test, $p < 0.01$, and there are no noticeable differences across datasets (the number of statistically significant cases varies between 86% and 92%).

In summary, topic subsets found by clustering combined with dimensionality reduction show correlations with the ground truth that are statistically significantly higher than average/random subsets. However, the difference is rather small: although clustering helps, it helps just a little. Indeed, considering the results of Fig. 8, one might be tempted to conclude that clustering of topics is not an effective technique, at least with the constraint $c = m$. Also, the oscillations of the one-for-cluster correlation curves that can be seen in Fig. 8 call for an explanation. To address these issues, and to present a detailed analysis of clustering of topics with the constrain $c = m$, we perform another experiment, described in the next section.

6.2.3 A simulation experiment

To further understand what is happening during the clustering process, and to further investigate the capabilities of the clustering process with the constraint $c = m$, as well as the limitations, we design the following simulation experiment. The aim of the simulation is to show what happens with clustering of topics in an ideal situation, where the data is distributed with a minimum and controlled amount of noise, and the topics are artificially clustered in a neat way. This represents the most favorable scenario for the topic clustering process. We will discuss the same experiment for cardinality-independent clustering in Sect. 6.3.

The experiment is as follows. We select s topics, called *seeds*. We experiment with choosing as seeds the topics from a collection in two ways: either randomly, or choosing a set of well separated topics after projecting the multidimensional topic space onto two dimensions. In the following, we report the results of the random selection only, as the other one provides a comparable result.

Given the seed topics, we form a set of new topics, placing in the neighborhood of each seed r fictitious topics in a hyper-sphere of radius ϵ; we call these topics the *surrounding* topics of the seed topics. Thus, we simulate an ideal scenario for clustering of topics where we have s ideal clusters of r topics each; 2ϵ is the maximum distance, in terms of AP (statAP), between two topics in the same ideal cluster. Note that, the higher ϵ, the higher the probability that the ideal clusters overlap, and therefore that a topic, during an automatic clustering process, is placed in a cluster different from that of its seed, and of the other topics in the same ideal cluster.

We now perform clustering as we did in Sect. 6.2.2; we use the constraint $c = m$, PCA, hierarchical clustering with a complete linkage method, and the cosine similarity as the distance function. We vary the three parameters as follows: $s \cdot r = 150$, with $s \in \{15, 30, 50\}$, and thus $r \in \{10, 5, 3\}$, and $\epsilon \in \{0.01, 0.02, 0.05\}$.

Results of the experiment are shown in Fig. 9. In panel (a), the one-for-cluster series for the three topic seeds (15, 30, and 50) are represented with different colors, and the different line types (continuous, dashed, and dotted) identify the different ϵ values (0.01, 0.02, and 0.05). The figure also shows the average series as gray thin lines. Figure 9b shows the same data with a different representation. Each series is obtained subtracting the corresponding

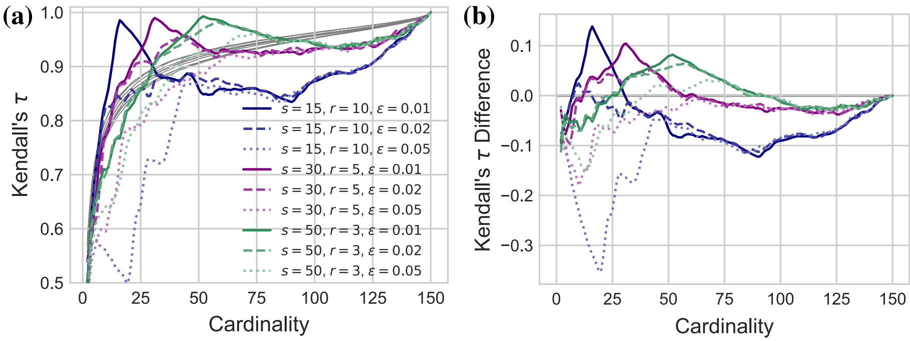

Fig. 9 On the left, Kendall's τ correlation values for the average and the one-for-cluster series. On the right, the series obtained subtracting the average series to the one-for-cluster one. The three series represent the three number of topic seeds: 15, 30, and 50, represented with different colors. The different colors and line types of the series represents different epsilon values. The series are smoothed using a mobile mean with a window of three elements. The gray lines represent the average series

average series from the one-for-cluster one. The horizontal gray line highlights the value of zero: if the series in the plot is above zero it means the one-for-cluster series has higher correlation values with the ground truth than the average one, if the series is below zero vice-versa the average series has higher correlation values.

We can draw several conclusions from these results. Looking at the highest peaks, one can see that they occur at cardinalities corresponding with the number of ideal clusters (equal to the number of seeds s): the clustering approach works well if the topics can be "naturally clustered" in a number of clusters corresponding to the cardinality of the subset of a few good topics; this is true when all the surrounding topics are placed in the same cluster as the corresponding seed topic. However, the further the cardinality is from this ideal number of clusters (the number of seeds), the more the correlation of the one-for-cluster series decreases, and becomes comparable with the random selection of topics (the average series), or even worse.

Focusing on the "negative" peaks (e.g. for the series with 15 seeds, for $\epsilon = 0.05$ at the cardinalities around 20, and for all three ϵ values at cardinalities around 90) we note that the negative peaks achieve lower values of correlation as ϵ increases, as expected. These negative peaks confirm that, if a natural clustering of topics is not possible, clustering of topics worsens the selection of a few good topics with respect to random selection. This effect can be explained by looking at the composition of the clusters produced during the cluster process, where we notice that surrounding topics of different seeds indeed tend to be clustered together even when ϵ is small. This is likely caused by the constraint $c = m$, that forces the number of clusters. Furthermore, the desired behavior would be that when increasing cardinalities, the clusters split into balanced sub-clusters; for example, with $s = 15$, at cardinality 30 each cluster containing the seed should split into 2 balanced clusters, at cardinality 45 into 3 balanced clusters, and so on. However, in practice this is not the case: on the contrary, there are always few clusters split into smaller clusters, while other larger clusters remain intact. This results in a "bad" clustering of topics: in the one-for-cluster series the majority of topics come from more fragmented clusters. We can say that the more fragmented weight more than the other in the evaluation; on the contrary, the average series chooses topics uniformly.

Finally, we note that there are some lower positive peaks in the series. For example, see in the chart on the right the series with 30 seeds with $\epsilon = 0.02$, for the cardinalities around the values of 18, 22, 39, and 41. These lower peaks suggest that it is not always the case that the data can be explained with only one number of clusters, but multiple numbers of clusters are possible to obtain a natural clustering of topics.

Summarizing, it seems reasonable to conclude that the $c = m$ constraint makes clustering ineffective for most of the cardinalities, even in the most favorable scenario. Moreover, considering real data, ϵ will be quite high, since in general it is unlikely that our vectors (topics) have similar values, with just a small ϵ difference. Thus cardinality-driven clustering does not seem to be a feasible technique to be applied on real data. For this reason, in the following we study cardinality-independent clustering, starting by repeating the simulation experiment of this section.

6.3 Cardinality-independent clustering

In our previous experiments, the number of clusters is equal to the number of selected topics. Now, we perform clustering of topics with a number of clusters m independent from the topic subset cardinality c and hopefully matching the number of clusters in a natural clustering.

6.3.1 The clustering process

In the case of cardinality-independent clustering, differently from cardinality-driven clustering, m is a parameter to be chosen. There are several ways of selecting such a parameter. The first alternative is to try all possible values from 2 to the number of topics. A second approach could be to rely on some index of goodness of the obtained clusters. Another possibility is to look at the results of cardinality-driven clustering: in cardinality-driven clustering, due to the constraint $c = m$, the positive peaks in the one-for-cluster series (see Figs. 8 and 9) correspond to m values leading to an effective clustering of topics; this fact can be exploited to choose the value of m for the cardinality-independent clustering: we can focus on the cardinalities corresponding to the positive peaks of the one-for-cluster series in cardinality-driven clustering. In the following we investigate the latter approach; we also tried various indexes on clustering goodness (e.g Connectivity, Dunn, and Silhouette) with no positive result, and we leave for future work the study of other feasible a priori approaches to find m.

Once the m clusters are formed, the probably most natural algorithm for selecting the topics from the clusters is as follows. Considering the one-for-cluster series, there exist three possibilities for each cardinality $c \in \{1, \ldots, n\}$:

- Case $c < m$: we select randomly c clusters, and then we select c elements, one for each cluster.
- Case $c = m$: we select one topic per cluster, as we did in the case of cardinality-driven clustering (Sect. 6.2).
- Case $c > m$: we select m topics as in the previous $c = m$ case; we then repeat for the remaining $c - m$, until we fall in the first $c < m$ case. When a cluster becomes empty during the process, we skip it in the following iterations.

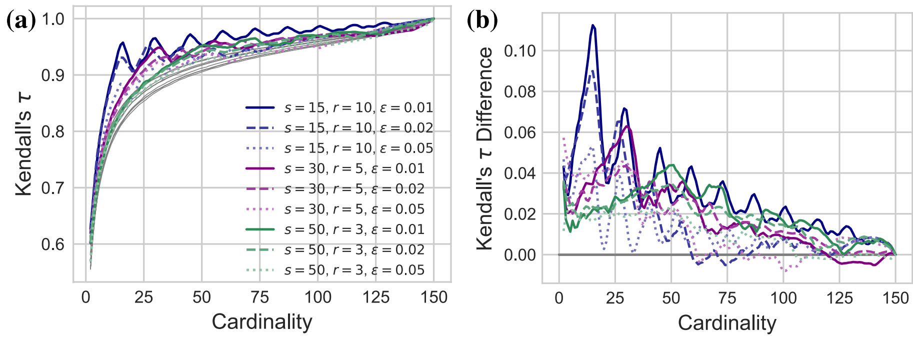

Fig. 10 Results of the cardinality-independent clustering for the artificial experiment. Compare with Fig. 9

Note that cardinality-driven and cardinality-independent clustering coincide only when $c = m$.

6.3.2 Cardinality-independent clustering on the simulated example

Figure 10 shows the results for cardinality-independent clustering for the same simulated experiment. The figure shows that in general, we obtain topic subsets that always have higher τ values than the average; this holds for almost all the s, r, and ϵ values.

Also, there are several positive peaks in the series. These occur at cardinalities corresponding to multiples of the number of topic seeds s; e.g. considering $s = 15$, the positive peaks are around cardinalities 15, 30, 45, and so on. This is an indication that multiple effective m values exist. Indeed, clustering is effective not only for m corresponding exactly to the cardinalities of the peaks, but also for near values, and this fact can be exploited for m selections.

Finally, the lower negative peaks of Fig. 9 almost disappear, even for the largest ϵ value of 0.05: even if the topics are difficult to cluster, the clustering process is still effective.

6.3.3 Cardinality-independent clustering on real data

Figure 11 shows the results of cardinality-independent clustering for the real-data experiment, for some selected m values, corresponding to the cardinalities of the positive peaks of the series of Fig. 8: 9 clusters for AH99, 5 clusters for TB06, 31 clusters for R04, and 24 clusters for MQ07. We choose to report the results corresponding to the highest peak at the lowest possible cardinality: for AH99 a similar behavior is found for cardinalities 16, and 31, for TB06 for cardinalities 22, 43, 45, and 75, for R04 for cardinalities 16, 25, 31, 45, 60, and 75, and finally for MQ07 for cardinality 64.

The figure shows that the one-for-cluster series always has higher τ values than the average series, for all the collections, with a single exception for R04. Cardinality-independent clustering is effective. As we have seen in Sect. 6.2, in the cardinality-driven clustering, even with PCA, the one-for-cluster series is often significantly lower than the average in the simulation experiment (see Fig. 9) and sometimes lower in the real datasets (see Fig. 8). In cardinality-independent clustering, this never happens: in the least favorable case, the one-for-cluster and average series are equivalent (the series overlap).

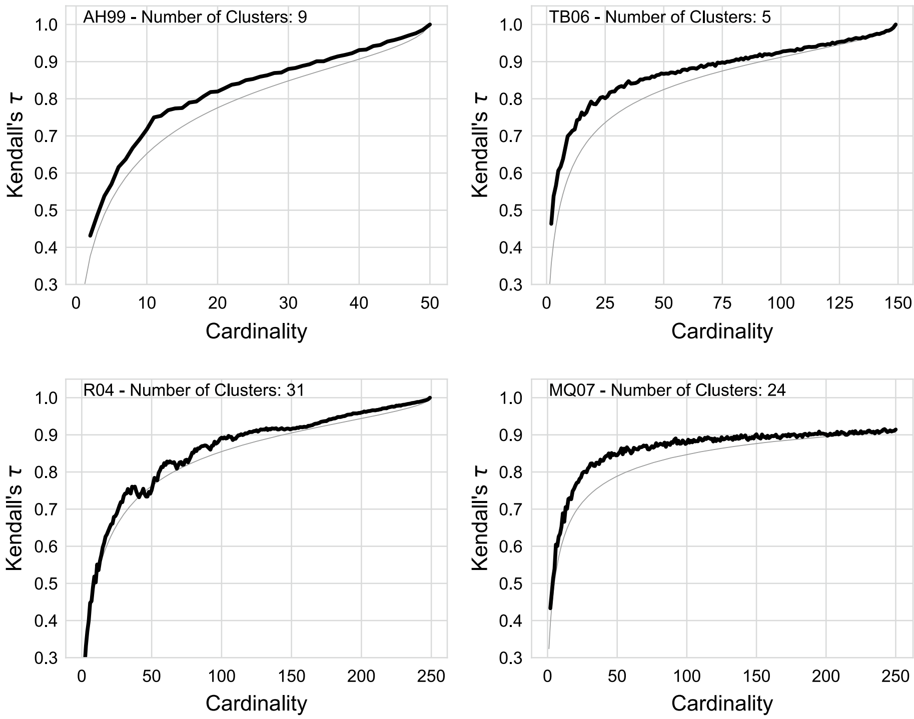

Fig. 11 Kendall's τ correlation curves for the four datasets: averaged (thinner gray lines), and obtained using clustering (thicker darker lines)

We also verified that the series oscillations do not depend on noise, as they still occur with 1M repetitions, as noted previously (see Footnote 9): the one-for-cluster series always fluctuates a little, but these oscillations are small and do not affect the results.

Another final result is that, in cardinality-independent clustering, the choice of the number of clusters m can be critical. A detailed analysis of our data shows that good m values can be found by looking at the positive peaks on the one-for-cluster series of the cardinality-driven clustering. For the m values corresponding to the cardinalities of such peaks, as well as the nearest cardinalities, the one-for-cluster series of the cardinality-independent clustering tend to have higher τ values than the average series, for almost any cardinality. It has to be noted that these are not all the good m values, as there exist other m values such that for the cardinality-independent clustering the one-for-cluster series is always above the average, but there is not a corresponding peak in the cardinality-driven clustering. However, this provides a general criterion for the choice of m. For example, considering our datasets, to obtains one-for-cluster series that are better than random topic selection: for AH99 any m value can be used (but cardinalities around 8, 15, and 30 are better), for TB06 the best values are around 10, for R04 the best values are 25, 75, and 110; and, finally, for MQ07 the best values are around 25, 45, and 60.

6.4 Discussion

The above results show that cardinality-independent clustering of topics is an a posteriori topic selection strategy that is more effective than the random selection of topics. The effectiveness increase is still not large but it is consistent across all cardinalities and collections. As all the other results of this line of research, this is an a posteriori strategy that is only potentially useful and cannot be applied in practice. However, it can give useful insights for a priori strategies, like suggesting the number of clusters to be used.

Note that although the setting is still a posteriori, clustering of topics shows only a limited effectiveness as a strategy to find good topic subsets. That is, even if we focused on a context where we expected clustering to be clearly effective, this was not the case. This is perhaps surprising and might even cast some doubts on the effectiveness of clustering also for an a priori approach; however, in that case the features used would be very different, and therefore this claim needs to be verified with further experiments, that we leave as future work.

Also, note that comparing the clustering curves with the average series, as we have done, might even be unfair, since the clustering approach needs the whole topic set to produce the topic subset at a given cardinality c, whereas the average series are produced using just c topics at cardinality c. In this respect, the clustering is even less effective. For instance, focusing on cardinality 50 for MQ07 (fourth chart in Fig. 11), we can indeed say that clustering has a higher correlation than average (0.85 vs. 0.79), but that a clustering-based topic subset is generated using all 1153 topics, whereas by using around 100 random topics, one would get the same correlation.

We can also compare clustering correlations with Hosseini et al. (2012)'s "Adaptive" ones. As in Sect. 4.4, we need to change again our setting to perform clustering on the dataset with all the runs (instead of the top 75% only); the obtained correlation values are shown in the last row of Table 2. Incidentally, by doing so, we are not able to obtain correlation values higher than the average as those on the top 75% runs; indeed, as it can be seen in Table 2, when using all the runs the correlation values obtained by our clustering are hardy distinguishable from Average values. This is consistent with the remark in Sect. 4.4: when using all the runs in a collection, the Average curves achieve higher values of τ, and therefore it is more difficult to do better than the Average baseline in such a case.

Focusing on the comparison between Adaptive and clustering, we see from Table 2 that Adaptive is more effective than clustering on the smaller (having a lower number of topics) AH99 dataset, and conversely clustering is more effective than Adaptive on the larger R04 dataset. This result will need to be confirmed by further experiments, but it suggests that the two approaches could be fruitfully combined.

As a final remark, we conjecture that one general reason for the less than satisfactory effectiveness obtained with a posteriori clustering could simply be a "tyranny of majority" effect. If there is a large subset of topics that can be "naturally clustered" together, and that cluster is indeed recognized by the clustering algorithm (as is quite likely), then the one-for-cluster selection method will be forced to pick up just one topic from that largest cluster. However, the topics in that large natural cluster are driving the evaluation in a specific direction – these topics "weigh more" than the other topics. This will result in penalizing the one-for-cluster selection method, that is forced to not recognize this majority. This conjecture is true at least to some extent in our datasets: in our experiments the largest cluster usually contained around 75–90% of the topics.

To analyse this conjecture, we performed a last experiment. Given a clustering of topics, and the topic subset obtained from it, we computed not only the MAP by averaging the AP values, but also a Weighted MAP (WMAP) in which the AP values are averaged with a weight corresponding to the size of the cluster the topic belongs to. Note that both MAP and WMAP make sense: the WMAP approach somehow assumes that the full topic set is a representative sample of the whole topic population, and therefore if some topics are clustered together, that happens because the whole topic population contains many topics like those; conversely, the MAP approach is based on the assumption that since some topics are very similar, picking just one of them avoids a biased sampling, in which the topics of larger clusters are over represented. Therefore, the two approaches differ on the weight given to each sampled topic; MAP assumes all topics to be of the same importance, conversely WMAP assumes topics that are sampled from a larger population are more representative, and thus more important. We also remark that by using WMAP we are not guaranteed that by using the full topic population we reach correlation 1 with MAP. Results are clearly negative: all correlation values obtained when using WMAP are not only lower than those obtained when using MAP, but also always lower than Average.

7 Conclusions and future work

Compared to previous work on using fewer topics in the evaluation of IR systems, our contributions are threefold. Addressing RQ1, we show that examining subsets of a larger ground truth topic set results in average and best subsets that are more highly correlated with the ground truth topic set than found in previous work (Guiver et al. 2009; Robertson 2011; Berto et al. 2013). It would appear that as the cardinality of the ground truth increases, the size of the subset (relative to ground truth) required to obtain a high correlation also decreases.

We also find that under large cardinalities, worst topic subsets are notably worse than shown in past work. Although finding a few bad topics was perhaps to be expected, when a larger pool of topics could be drawn from the large size of worst topic subsets that still had very low correlations was striking. Examination of the effectiveness of worst subsets shows that they were mainly composed of topics with poor effectiveness scores.

Addressing RQ2, we analyze the role of statistically significant differences between runs for different topics subsets. The ability to distinguish statistically between the effectiveness of two runs is impaired when topic cardinality is lowered. The main problem is an increase in false negatives (type II errors) when making comparisons. This issue has not been shown before in this area of topic subsetting research, although it has been addressed in conjunction with incomplete relevance judgments: see for example Carterette and Smucker (2007, Table 2), which agree with our findings. Some subsets were shown to be better than others at minimizing type I and II errors. The analysis showed that the level of error reduced relatively quickly as subset cardinality increased. Nevertheless, because all of our experiments still use relatively small populations of topics when compared to “the set of all topics in the world”, it is not clear if the level of type II error will reduce sufficiently. The collections still don’t give us a sense of what the “true” population of possible topics is like, and we have no way to be sure that the full cardinality is the truth. In a way, the results in this paper suggest that all test collections are suspect, since their very small subset of topics might be completely un-correlated with the “true” population of all possible topics.

Our findings on the overlap of best and worst topics sets confirm that being a good topic largely depends on the other topics in the subset. In general, the previously established terminology of best/worst topic sets is perhaps misleading since it can be argued that the worst topics are actually the most interesting ones (they rank runs in ways contrary to the majority of topics), whereas the best topics feature a high degree of redundancy that might lead to a waste of resources. Indeed, the high degree of redundancy is manifested in the best correlation curves, that have high correlation values also for low cardinalities.

Addressing RQ3, our analysis showed that clustering is effective in finding topic subsets that are more representative than simply taking average or random subsets, as long as the clustering is combined with dimensionality reduction. However, the topic subsets obtained by clustering are only slightly more effective than random topic subsets, and are far from featuring correlations that are as high as the best topic sets. A comparison with, and an analysis of, related work shows that we are in good company, though: good topics subsets exists but finding them seems a rather daunting task. While the work here is a first step in finding representative and effective topic subsets, there is still much work to be done to improve topic subset selection.

In future work, we plan to consider the correlation between topic subsets (rather than between a topic subset and the full topic set) as well as top-heavy measures of association such as Rank Biased Overlap or τ_{AP}, to give more importance to the most effective systems. We have only started to analyze how best and worst topic sets are formed. Considering the extreme nature of Best and Worst series, extreme value theory might be useful to better understand and model the stochastic behavior of Best / Worst series and topic subset distributions. We also plan to deepen the analysis by finding more semantic features that characterize a good/bad topic set. Indeed, as in previous research, we have not attempted to devise methods to find good topic subsets before the evaluation exercise is performed, or while it is ongoing; the focus of our research so far has been on working to understand how different topic sets interact. Future work studying more semantic topic features, combined with many runs, will hopefully help to provide a set of guidelines for sound topic set engineering.

Acknowledgements We thank several colleagues for useful discussions: Jamie Callan, Ben Carterette, Bruce Croft, David Lewis, Ian Soboroff, and Ellen Voorhees. Virgil Pavlu provided useful information on statMAP. The reviewers provided very insightful and detailed remarks that helped to improve the paper. This work was supported by the Australian Research Council's *Discovery Projects* Schemes (DP170102231 & DP170102726).

References

Allan, J., Carterette, B., Aslam, J. A., Pavlu, V., Dachev, B., & Kanoulas, E. (2007). Million query track 2007 overview. In *Proceedings of TREC*.

Bartlett, J. E., Kotrlik, J. W., & Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, *19*(1), 43–50.

Berto, A., Mizzaro, S., & Robertson, S. (2013). On using fewer topics in information retrieval evaluations. In *Proceedings of the ICTIR*, (p. 9).

Bodoff, D., & Li, P. (2007). Test theory for assessing ir test collections. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, (pp. 367–374). New York: ACM.

Buckley, C., & Voorhees, E. (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd SIGIR*, (pp. 33–40).

Carterette, B., Allan, J., & Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proceedings of the 29th SIGIR*, (pp 268–275).
Carterette, B., Pavlu, V., Fang, H., & Kanoulas, E. (2009a). Million query track 2009 overview. In *Proceedings of TREC*.
Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., & Allan, J. (2009b). If i had a million queries. In *Proceedings of the 31th ECIR, ECIR '09*, (pp. 288–300).
Carterette, B., & Smucker, M. D. (2007). Hypothesis testing with incomplete relevance judgments. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management*, (pp 643–652). New York: ACM. CIKM '07. https://doi.org/10.1145/1321440.1321530.
Carterette, B. A. (2012). Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 4.
Cattelan, M., & Mizzaro, S. (2009). IR evaluation without a common set of topics. In *Proceedings of the ICTIR*, (pp. 342–345).
Feise, R. (2002). Do multiple outcome measures require *p*-value adjustment? *BMC Medical Research Methodology*, *2*, 8.
Guiver, J., Mizzaro, S., & Robertson, S. (2009). A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems*, *21*(1–21), 26.
Hauff, C., Hiemstra, D., Azzopardi, L., & de Jong, F. (2010). A case for automatic system evaluation. In *Proceedings of the ECIR*, (pp. 153–165).
Hauff, C., Hiemstra, D., de Jong, F., & Azzopardi, L. (2009). Relying on topic subsets for system ranking estimation. In *Proceedings of the 18th CIKM*, (pp. 1859–1862).
Hosseini, M., Cox, I. J., Milic-Frayling, N., Shokouhi, M., & Yilmaz, E. (2012). An uncertainty-aware query selection model for evaluation of ir systems. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*, (pp. 901–910). New York, NY, USA: ACM. SIGIR '12. https://doi.org/10.1145/2348283.2348403
Hosseini, M., Cox, I. J., Milic-Frayling, N., Sweeting, T., & Vinay, V. (2011a). Prioritizing relevance judgments to improve the construction of IR test collections. In *Proceedings of the 20th CIKM 2011*, (pp. 641–646)
Hosseini, M., Cox, I. J., Milic-Frayling, N., Vinay, V., & Sweeting, T. (2011b). Selecting a subset of queries for acquisition of further relevance judgements. In *Proceedings of the ICTIR*, (pp. 113–124). lNCS 6931.
Kutlu, M., Elsayed, T., & Lease, M. (2018). Intelligent topic selection for low-cost information retrieval evaluation: A new perspective on deep vs. shallow judging. *Information Processing and Management*, *54*(1), 37–59. https://doi.org/10.1016/j.ipm.2017.09.002.
Mehrotra, R., & Yilmaz, E. (2015). Representative & informative query selection for learning to rank using submodular functions. In *Proceedings of the of the 38th international ACM SIGIR conference on research and development in information retrieval*, (pp. 545–554). New York, NY, USA: ACM, SIGIR '15. https://doi.org/10.1145/2766462.2767753
Mizzaro, S., & Robertson, S. (2007). HITS hits TREC—Exploring IR evaluation results with network analysis. In *Proceedings of the 30th SIGIR*, (pp. 479–486).
Moffat, A., Scholer, F., & Thomas, P. (2012). Models and metrics: IR evaluation as a user process. In *Proceedings of the Australasian document computing symposium, Dunedin, New Zealand*, (pp. 47–54).
Pavlu, V., & Aslam, J. (2007). A practical sampling strategy for efficient retrieval evaluation. Tech. rep., technical report, college of computer and information science, Northeastern University.
Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets* (1st ed.). Cambridge: Cambridge University Press.
Robertson, S. (2011). On the contributions of topics to system evaluation. In *Proceedings of the ECIR, lNCS 6611*, (pp. 129–140).
Roitero, K., Maddalena, E., & Mizzaro, S. (2017). Do easy topics predict effectiveness better than difficult topics? In J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, & J. Tait (Eds.), *Advances in information retrieval* (pp. 605–611). Cham: Springer International Publishing.
Roitero, K., Soprano, M., Brunello, A., & Mizzarom, S. (2018a). Reproduce and improve: An evolutionary approach to select a few good topics for information retrieval evaluation. *ACM Journal of Data and Information Quality*, *10*(3), 12:1–12:21. https://doi.org/10.1145/3239573.
Roitero, K., Soprano, M., & Mizzaro, S. (2018b). Effectiveness evaluation with a subset of topics: A practical approach. In *The 41st international ACM SIGIR conference on research and development in information retrieval*, (pp. 1145–1148). New York, NY, USA:ACM, SIGIR '18. https://doi.org/10.1145/3209978.3210108
Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, (pp. 13–19). New York, NY:ACM Press.

Sakai, T. (2007), Alternatives to bpref. In *Proceedings of the 30th annual international ACM SIGIR Conference on research and development in information retrieval*, (pp. 71–78). New York, NY:ACM, SIGIR '07. https://doi.org/10.1145/1277741.1277756
Sakai, T. (2014). Designing test collections for comparing many systems. In *Proceedings of the 23rd CIKM 2014*, (pp. 61–70).
Sakai, T. (2016a). Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th SIGIR*, (pp. 5–14). ACM.
Sakai, T. (2016b). Topic set size design. *Information Retrieval Journal*, *19*(3), 256–283.
Sanderson, M., & Soboroff, I. (2007). Problems with Kendall's Tau. In *Proceedings of the 30th SIGIR*, (pp. 839–840).
Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th SIGIR*, (pp. 162–169).
Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton: CRC Press.
Urbano, J. (2016). Test collection reliability: A study of bias and robustness to statistical assumptions via stochastic simulation. *Information Retrieval Journal*, *19*(3), 313–350. https://doi.org/10.1007/s10791-015-9274-y.
Urbano, J., Marrero, M., & Martín, D. (2013). On the measurement of test collection reliability. In *Proceedings of the 36th SIGIR*, (pp. 393–402).
Urbano, J., & Nagler, T. (2018). Stochastic simulation of test collections: Evaluation scores. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, (pp. 695–704). New York, NY, USA: ACM, SIGIR '18. https://doi.org/10.1145/3209978.3210043.
Voorhees, E., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In*Proceedings of the 25th SIGIR*, (pp. 316–323).
Webber, W., Moffat, A., & Zobel, J. (2008). Statistical power in retrieval experimentation. In *Proceedings of the 17th CIKM*, (pp. 571–580).