Effectiveness Evaluation with a Subset of Topics: A Practical Approach

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

Michael Soprano
University of Udine
Udine, Italy
soprano.michael@spes.uniud.it

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

ABSTRACT

Several researchers have proposed to reduce the number of topics used in TREC-like initiatives. One research direction that has been pursued is what is the optimal topic subset of a given cardinality that evaluates the systems/runs in the most accurate way. Such a research direction has been so far mainly theoretical, with almost no indication on how to select the few good topics in practice. We propose such a practical criterion for topic selection: we rely on the methods for automatic system evaluation without relevance judgments, and by running some experiments on several TREC collections we show that the topics selected on the basis of those evaluations are indeed more informative than random topics.

CCS CONCEPTS

• **Information systems** → **Test collections**;

KEYWORDS

Few topics, test collections, TREC, topic selection

ACM Reference Format:
Kevin Roitero, Michael Soprano, and Stefano Mizzaro. 2018. Effectiveness Evaluation with a Subset of Topics: A Practical Approach. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210108

1 INTRODUCTION

In Information retrieval (IR) test collection based evaluation, the number of topics used is a critical issue. Since it is one of the main parameters to determine the overall cost, it is not surprising that several researchers have studied how to reduce such a number in TREC-like initiatives. One research direction that has been pursued is to identify the optimal topic subset of a given cardinality that evaluates the systems/runs in the most accurate way [2, 6, 11]. The main limitation of such an approach is that it requires the full evaluation to be run, since it needs the effectiveness evaluation for each system/topic pair (usually represented as a system/topic matrix of, e.g., average precision values). So the results have been so far mainly theoretical, with almost no indication on how to select the few good topics in practice. In this paper we propose such a practical criterion for topic selection. We rely on the methods for automatic effectiveness evaluation without relevance judgments [1, 4, 9, 10, 12–14, 17], and by running some extensive experiments on several TREC collections we show that the topics selected on the basis of those evaluations are indeed more informative than a random selection of topics.

	t_1	$\cdots$	t_n	MAP
s_1	$\mathrm{AP}_{1,1}$	$\cdots$	$\mathrm{AP}_{1,n}$	MAP_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
s_m	$\mathrm{AP}_{m,1}$	$\cdots$	$\mathrm{AP}_{m,n}$	MAP_m

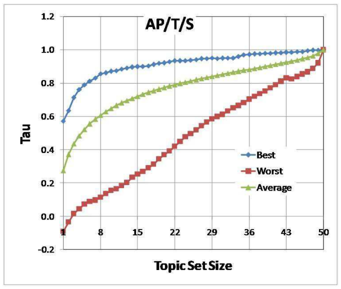

Figure 1: AP and MAP representation and correlation curves (both adapted from [6]).

2 RELATED WORK

2.1 A Few Good Topics

We briefly summarize the research on topic subsets and on effectiveness evaluation without human relevance judgments.

The table on the left of Figure 1 is a representation of the results of a TREC-like evaluation. Each row is a system / run and each column is a topic; each cell of the matrix $\mathrm{AP}_{i,j}$ is the effectiveness value (in this paper we focus on Average Precision, AP) of system i on topic j. When averaging across the n columns (topics) one obtains Mean AP (MAP), a measure of the effectiveness of a system.

The number n of topics in an evaluation initiative has received attention since the first TREC editions (and even before) [3]. The classical and main approach has been to understand what happens when selecting a *random* subset of topics of a given cardinality (i.e., a by computing MAP on a random subset of the columns in Figure 1). By doing so, the evaluation of systems / runs is in general different than when using the full set of topics. The effect is generally measured using the (linear, rank) correlation between the ground truth of the m original MAP values and the m predicted MAP values, obtained on the basis of the topic subset only.

A different approach, more related to our work, has been to find the *best* subset of topics of a given cardinality: one does not select a random subset of columns in Figure 1 but the optimal subset, i.e., the one leading to the highest correlation value between the real and the predicted MAP. The first proposal is by Guiver et al. [6], which performed a heuristic search on all possible topic subsets. More in detail, their work focused on the correlation of three series: (i) Best, the subset of topics which has the highest correlation (i.e.,

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210108

the predicted MAP of systems is the most similar) with the ground truth; (ii) Worst, the subset of topics which has the which has the smallest correlation (i.e., the predicted MAP of systems is the least similar) with the ground truth; and (iii) Average, the correlation that one might expect when choosing topics randomly. Guiver et al. used both Pearson's linear and Kendall's rank correlations; in this paper we focus on the latter. Their analysis shows that subsets of good and bad topic sets exist with a high / low correlation even at low cardinalities; for example (see the chart on the right of Figure 1), on a ground truth of 50 topics, at cardinality 8 one can identify a Best topic subset with $\tau \simeq 0.85$ and a Worst set with $\tau \simeq 0.1$).

This work has been continued by Robertson [11], who questioned the generality of Guiver et al. results, as well as attempted a first example of a practical topic selection strategy, that however turned out to be ineffective. Berto et al. [2] confirmed the findings of Guiver et al. and extended their work by looking at the number, the distribution, and the stability of the Best and Worst topic sets. Such a research direction has been so far purely theoretical, with no indication on how to select the few good topics in practice. In this paper we propose such a practical approach. In this respect, our work is an attempt to make the methodology proposed in the above described publications [2, 6, 11] more similar to the more recent work by Kutlu et al. [7] who proposed a learning-to-rank based approach for topic selection, and analyzed in detail the role of deep and shallow pooling in the topic selection process, and its impact on the evaluation.

2.2 Evaluation without Relevance Judgments

The first proposal of evaluating IR systems without relevance judgments is by Soboroff et al. [13]; their proposal is simply to random sample documents from the pool and treat such documents as relevant. The intuition is that the random sample is performed on a biased set of documents: if many different systems retrieve the same document (maybe even in the first rank positions) that document is probably relevant. Results show that the estimated final rank of systems correlates decently (Kendall's $\tau \simeq 0.5$) with the official TREC ranking; the method fails in predicting the rank of best systems, which are somehow "peculiar".

Wu and Crestani [17] proposed a method based on data fusion techniques, which are used to merge the ranked lists of documents retrieved by the systems, assigning a popularity score to each document, and using such score to provide an estimated final rank of systems. Wu and Crestani propose five variants to assign the popularity score to each document.

Another approach, proposed by Aslam and Savell [1], measures the similarity of the ranked lists of each pair of systems, and ranks such systems by the computed similarity index. This methodology is strongly correlated with Soboroff et al.'s one. Aslam and Savell also raise the issue that the predicted ranking of systems is based on document popularity rather than relevance.

Nuray and Can [9, 10] proposed a method based on three strategies used in democratic elections to measure popularity of candidates: the "RankPosition", the "Borda", and the "Condorcet" method. Furthermore, such indexes are computed considering either all the system/runs which participated in a given TREC edition (the "normal" method), or selecting just the most "peculiar" systems (the "bias" method), that are the systems that have a ranked list which deviates more from the norm.

Spoerri [14] proposed an evaluation method which relies on a set of trials between runs; each run is assigned five times to a set with other four runs, thus in this way each system participate in exactly five trials. Then, for each trial, for each system the method computes the percentage of documents retrieved by the system alone ("Single" index), the percentage of documents retrieved by all the systems in the trial ("AllFive"), and the algebraic difference between the Single and the AllFive indexes ("SingleMinusAllFive"); those three indexes are averaged across the trials and then retrieval systems are ranked according to their SingleMinusAllFive index, the lower the index the better.

Sakai and Lin [12] proposed a variation of the Condorcet method proposed in [10], which is more feasible to compute even for deep pools, and it is strongly correlated although statistically different from Nuray and Can's proposal.

All the above methods provide an approximate evaluation matrix; that is a system by topic matrix similar to the one in Figure 1, but with predicted $\text{AP}_{i,j}$ and MAP_i values.

3 OUR APPROACH

The main limitation of the above "few good topics" approaches [2, 6, 11] is that they are only theoretical, not practical. For example, they provide an optimum to aim at (the Best correlation curves) but to actually select the Best subset of topics, the whole TREC evaluation exercise has to be performed, since the matrix in Figure 1 is needed. However, each of the methods for effectiveness evaluation without relevance judgments [1, 4, 9, 10, 12–14, 17] indeed provides an approximation of that matrix, which can possibly be used as input to the approaches that find the optimal subset of a few good topics. This is the main idea of this paper.

Thus, we run the few topics approach used in [2, 6, 11] on the approximate matrices. Then, for each cardinality, we consider the Best subset of topics found on the approximate matrices and we use such topics on the real matrix to produce a predicted evaluation of systems using the selected few topics. We then measure the correlation obtained by such a reduced topic subset: if the approximate matrices are representative of the real matrix, then the obtained correlation should be higher than the correlation found with a random selection of topics (i.e., the Average correlation curve in Figure 1).

Moreover, instead of using the individual methods alone, we combine such methods. The intuition on which we rely is that while a single method can produce a poorly representative matrix on a particular collection and a highly representative matrix on another collection, a combination of all the approximate matrices should provide a more stable and general AP prediction. To combine the methods, we follow the same approach of Mizzaro et al. [8]: we train a machine learning system that, on the basis of the TREC data of the previous years, learns a model that is then applied on a subsequent year TREC test collection. So, previous years test collections are the training set and the new test collection is then the test set; the features are the approximate matrices produced by the individual methods, and the combination function is the learned best combination of them to fit the real AP values. In other terms, the combination function, or the machine learning model, is

Table 1: The datasets used in this paper.

Acronym	Name	Year	Topics	Runs	Used Topics
TREC5	Ad Hoc	1996	50	61	251-300
TREC6	Ad Hoc	1997	50	74	301-350
TREC7	Ad Hoc	1998	50	103	351-400
TREC8	Ad Hoc	1999	50	129	401-450
TREC01	Ad Hoc	2001	50	97	501-550
TB04	TeraByte	2004	49	69	701-750 (no 703)
R05	Robust	2005	50	74	See [15, Figure 1]
W11[1]	Web Track	2011	50	61	101-150
W12[1]	Web Track	2012	50	48	151-200
W13[1]	Web Track	2013	50	55	201-250
W14[1]	Web Track	2014	50	30	251-300

[1] Binarized: collapsed relevance levels $\{-2, 0\}$ into 0, and $\{1, 2, 3\}$ into 1.

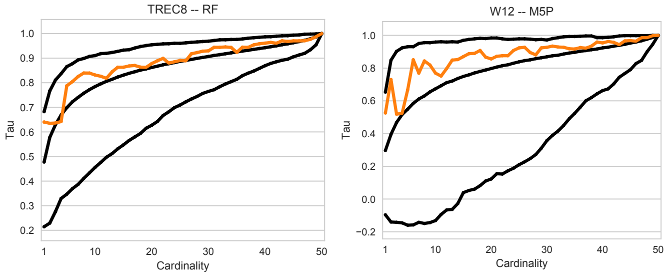

Figure 2: Two examples of the obtained correlation curves.

the one that, on the basis of historical data, provides the best prediction of real AP values. We try six machine learning algorithms [16]: Linear Regression (LR), M5P model tree (M5P), Random Forest (RF), Neural Networks (NN), Support Vector Machine with Polykernel (SVM_Poly), and Support Vector Machine with Radial Basis Function Kernel (SVM_RBF).

4 EXPERIMENTS AND RESULTS

To run our experiments we use the ten datasets shown in Table 1. When compared to the experiments by Kutlu et al. [7], we attempt a more general and systematic evaluation, as we consider 10 different test collections instead of the 4 used by them. For each of the test collections, we compute the 22 approximate matrices (the 16 individual methods of Section 2.2 plus their 6 machine learning combinations above described).

Figure 2 shows two examples of the obtained correlation curves, compared to the Best, Average and Worst. The one on the left is obtained by RF on TREC8 (the same data as in Figure 1); the one on the right by M5P on W12. To better understand, as well as objectively quantify, the effectiveness of the methods, we define the following simple index that measures if and how much the obtained correlation curve is above the Average curve. We denote with $M_i(k)$ the Kendall's τ correlation of the i-th method at cardinality k, and with $A(k)$ the τ correlation of the Average series at cardinality k. If we have C collections, each of them having T topics, then we can define the Predictive Power Index of method i (PPI(i)) as

$$\text{PPI(i)} = \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{T}\left(M_i(k) - A(k)\right).$$

PPI describes the behavior of a topic selection strategy with respect to the Average series (i.e., a random topic selection). If PPI > 0 the topic selection strategy is effective, if PPI < 0 it is not, and if PPI $= 0$ the topic selection strategy is equivalent to the expected value of a random topic selection.

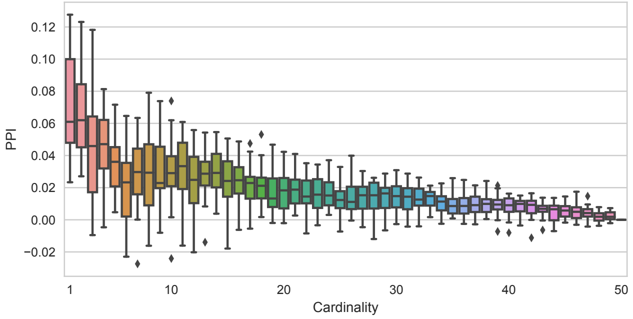

Figure 3: The distributions of PPI values for each cardinality.

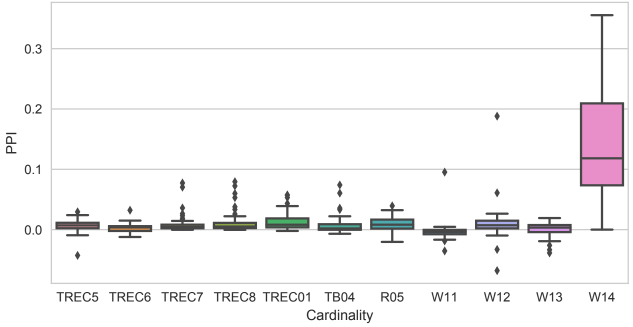

Figure 4: The distributions of PPI values for each collection.

The box-plots in Figure 3 show the distributions of the PPI values (y-axis) for each cardinality (x-axis). Each box-plot is a representation of the 22 PPI values obtained, one for each method. The PPI index is almost always positive, for both the median values and the lowest quantiles. Also the number of negative outliers is very low. However, this positive result depends in part on a skewed distribution of the PPI values for different collections, as demonstrated in Figure 4, that shows a breakdown of the PPI values for each collection separately: PPI values are still on the y-axis, the different collections are on the x-axis, and each box-plot is still obtained combining the single PPI values for the 22 methods. The W14 collection is a clear outlier, and the positive results obtained might depend just on it. We therefore repeat the same analysis without W14. Figure 5 shows that, although to a smaller extent, the positive result still hold: PPI are positive. This is also confirmed by running a statistical significance test: all series have p-value < 0.01 according to the paired Wilcoxon signed rank test.

Thus, when selecting a few good topics using the approximate matrices obtained with the methods for effectiveness evaluation without relevance judgments, one indeed finds topic subsets that are significantly better than a random selection. This is true for most of such methods; however, some methods could be more effective in this respect. In the last part of the paper we address this issue. We start by remarking that in our scenario not all the cardinalities have the same importance, for two main reasons: (i) above a certain cardinality threshold, the Average τ correlation is close to 1 (i.e., the higher the cardinality the harder is to beat the Average baseline),

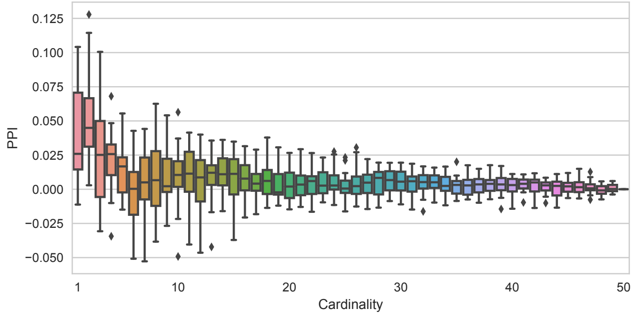

Figure 5: The distribution of PPI values for each cardinality, having excluded the outlier collection W14.

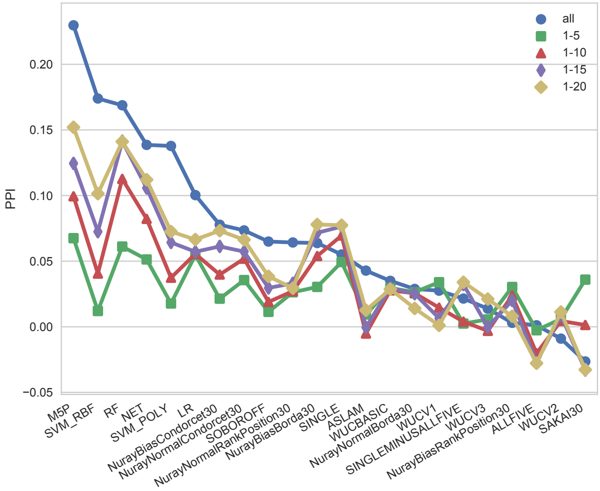

Figure 6: PPI of the methods over cardinality ranges.

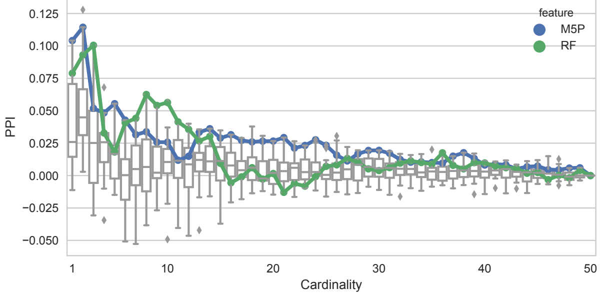

Figure 7: The PPI values obtained by M5P and RF, compared to the distribution of PPI values of Figure 5.

and (ii) higher cardinalities are less interesting, since the higher the cardinality the less useful the topic set reduction is; in other words, since we are interested in reducing as much as possible the topic set size, the lower the cardinality the better.

We therefore perform a breakdown of the PPI measure over the first 5 cardinalities (1,5], the first 10, the first 15, the first 20, and all cardinalities. Figure 6 shows the result: whereas when considering all cardinalities M5P, SVM_RBF, and RF are the three most effective methods, when focusing on lower cardinalities clearly M5P and RF show consistently higher PPI. We thus focus on this two methods in Figure 7. The plot shows that both RF and M5P have a PPI score greater than zero, especially for lower cardinalities, i.e., the most interesting ones. Also note that the two curves shown in Figure 2 are two examples of those obtained by these two methods: in both cases, using fewer than 10 topic produces ranking of systems with a Kendall's correlation higher than 0.8 with the ranking obtained with all the topics. Overall, the topics selected applying the fewer topics algorithm to the results of the M5P and RF methods are those with the best results.

5 CONCLUSIONS AND FUTURE WORK

We have applied the few topics approach to the outcome of the methods for effectiveness evaluation without relevance judgments. Our experimental results on ten TREC test collections show that such methods (and especially M5P and RF) allow to select a subset of a few topics that evaluate a population of systems / runs in a more accurate way than a random selection of topics of the same cardinality. This practical result is the first successful attempt of showing the practical usefulness of the few topics approaches [2, 6, 11], thus addressing their main limitation. For example, it could be used to select which topics to evaluate when resources are limited. In future work we plan to include TREC collections with a higher number of topics, as well as report on a more detailed comparison with the work by Kutlu et al. [7]. Furthermore, we plan to apply models used in response theory [5] to study the relationship between topic sets.

REFERENCES

[1] Javed A. Aslam and Robert Savell. 2003. On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In *Proceedings of 26th ACM SIGIR*. 361–362.

[2] Andrea Berto, Stefano Mizzaro, and Stephen Robertson. 2013. On Using Fewer Topics in Information Retrieval Evaluations. In *Proc. of ACM ICTIR 2013*. 9:30–9:37.

[3] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd ACM SIGIR*. ACM, New York, NY, USA, 33–40.

[4] Fernando Diaz. 2007. Performance Prediction Using Spatial Autocorrelation. In *Proceedings of 30th ACM SIGIR*. 583–590.

[5] Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

[6] John Guiver, Stefano Mizzaro, and Stephen Robertson. 2009. A Few Good Topics: Experiments in Topic Set Reduction for Retrieval Evaluation. *ACM Trans. Inf. Syst.* 27, 4, Article 21 (Nov. 2009), 26 pages.

[7] Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Intelligent topic selection for low-cost information retrieval evaluation: A New perspective on deep vs. shallow judging. *Inform. Processing & Management* 54, 1 (2018), 37–59.

[8] Stefano Mizzaro, Josiane Mothe, Kevin Roitero, and Md Zia Ullah. 2018. Query Performance Prediction and Effectiveness Evaluation Without Relevance Judgments: Two Sides of the Same Coin. In *Proc. of the 41st ACM SIGIR*. In press.

[9] Rabia Nuray and Fazli Can. 2003. Automatic Ranking of Retrieval Systems in Imperfect Environments. In *Proceedings of 26th ACM SIGIR*. 379–380.

[10] Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management* 42, 3 (May 2006), 595–614.

[11] Stephen Robertson. 2011. On the Contributions of Topics to System Evaluation. In *Proceedings of the 33rd ECIR*. 129–140.

[12] Tetsuya Sakai and Chin-Yew Lin. 2010. Ranking Retrieval Systems without Relevance Assessments — Revisited. In *Proceeding of 3rd EVIA — A Satellite Workshop of NTCIR-8*. National Institute of Informatics, Tokyo, Japan, 25–33.

[13] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems Without Relevance Judgments. In *Proc. of 24th ACM SIGIR*. 66–73.

[14] Anselm Spoerri. 2007. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Processing & Management* 43, 4 (2007), 1059 – 1070.

[15] Ellen M Voorhees. 2003. Overview of the TREC 2003 Robust Retrieval Track.. In *Trec*. 69–77.

[16] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[17] Shengli Wu and Fabio Crestani. 2003. Methods for Ranking Information Retrieval Systems Without Relevance Judgments. In *Proceedings of the 2003 ACM Symposium on Applied Computing*. 811–816.