

Improving the Efficiency of Retrieval Effectiveness Evaluation: Finding a Few Good Topics with Clustering?

Kevin Roitero¹ and Stefano Mizzaro²

¹ Dept. of Maths, Computer Science, and Physics. University of Udine, Italy
`kevin.roitero@outlook.com`

² Dept. of Maths, Computer Science, and Physics. University of Udine, Italy
`mizzaro@uniud.it`

Abstract. We consider the issue of using fewer topics in the effectiveness evaluation of information retrieval systems. Previous work has shown that using fewer topics is theoretically possible; one of the main issues that remains to be solved is how to find such a small set of a few good topics. To this aim, in this paper we try a novel approach based on clustering of topics. We consider various algorithms, metrics, and various features of topics that can be helpful in identifying such a set.

1 Introduction

Effectiveness evaluation is of paramount importance in the Information Retrieval (IR) field. The effectiveness evaluation method commonly used by Text REtrieval Conference (TREC) and other similar initiatives is based on a set of topics, which are textual descriptions of information needs, and on computing relevance assessments for these topics in a pooled set of documents. Participants run their own retrieval systems on the topics provided, and return a ranked list of the retrieved documents. Those are then pooled and their relevance is judged by human assessors. The process of creating these relevance assessments is very expensive in terms of money (circa \$30M from 1999 to 2009 according to [3]) and in terms of human effort; every technique developed for reducing the topic set size is a significant progress in the efficiency of the whole evaluation process.

We present a novel approach, based on clustering, to select such a good subset of topics. This paper is structured as follows. Sect. 2 describes the state of the art and the research question, Sect. 3 describes the experiments performed. We conclude in Sect. 4, where we also provide some directions for future work.

2 Background and Research Question

The starting point of our research is shown in Table 1. Each row represents a system (or, in TREC terms, run). Each column represents a topic. Each $AP(s_i, t_j)$ represents the effectiveness of the system s_i on the topic t_j . AP stands for the Average Precision measure, while MAP stands for Mean AP.

Table 1: AP and MAP, for n topics and m systems (from [2], Table 1)

APs	t_1	\dots	t_n	MAP
s_1	$AP(s_1, t_1)$	\dots	$AP(s_1, t_n)$	$MAP(s_1)$
\vdots		\ddots		\vdots
s_m	$AP(s_m, t_1)$	\dots	$AP(s_m, t_n)$	$MAP(s_m)$

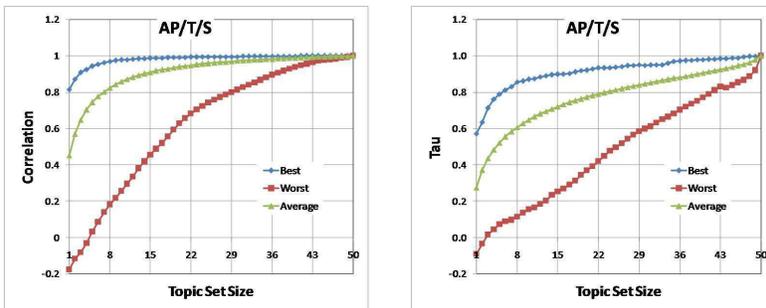


Fig. 1: Best, worst, and average-topic set for AH99 (from [2], Figures 2 and 3)

Guiver et al. [2] show that a topic set reduction is possible, and that some topic sets are in fact more informative than others. Starting with a larger set of topics, it is possible to discover a good-topic subset; that is, a subset which has a value of correlation of 0.96 for Pearson’s ρ (henceforth, ρ) and a value of Kendall’s τ (henceforth, τ) above 0.85³. This behavior is shown in Fig. 1; the plot shows the ρ and τ values for the three types of topic subsets for each topic set size: the best/worst, that is the topic subset with the highest/lowest value of correlation to the full topic set, and the average, that is the average value for ρ and τ of all the topic subsets considered. This result is found by exhaustive search (and heuristic search when exhaustive was not feasible): enumerating all subsets and measuring the correlation of the MAP obtained on those subsets only with the MAP on the full topic set. Berto et al. [1] find further confirmations and generalization of these results.

It is reasonable to conjecture that similar topics will evaluate systems in a similar way, and therefore they might be redundant: by removing some of the similar topics the final system evaluation should not change much. When looking for smaller topic subsets a natural strategy is then: (i) to cluster topics into a set (henceforth we call this set a *cluster-set* of topics), and (ii) to select a topic representative of each cluster and form a topic subset (henceforth we call this set a *one-for-cluster-topic* subset). Therefore, we inquire into the following

³ The values are justified on the basis of previous proposals: [4] states that evaluation schemes with values τ above 0.9 should be considered equivalent, while there is noticeable changes in the ranking with values below 0.8; [1] chose the value in the middle, 0.85; values of ρ are higher so they chose 0.96 as its value.

Table 2: Composition of each TREC dataset

Dataset	Description	Number of	Number of
		Topics	Systems
AH99	“Ad Hoc” track from TREC-8	50	96
TB06	“Terabyte” track from TREC 2006	149	49
R04	“Robust” track from TREC 2004	249	82
MQ07	“Million Query” track from TREC 2007	1153	26

research question (RQ): “Can clustering help to find a good-topic subset?”. Clustering may depend on several aspects; in this paper we show some preliminary exploration of this space. Table 2 shows the datasets used in these experiments.

3 Clustering of Topics

To answer our research question we build a cluster-set of topics. We use the columns of Table 1 as vectors, so that each topic is represented with an m -dimensional vector $t_k = \{\text{AP}(s_1, t_k), \dots, \text{AP}(s_m, t_k)\}$, where $\text{AP}(s_i, t_j)$ represents the effectiveness of the evaluation of the system s_i on the topic t_j . We use hierarchical clustering, with complete method to join clusters, and the cosine distance. Then, in order to compare the one-for-cluster set with the average set (similar to the comparisons done in [1, 2]), we compare the correlation values of the one-for-cluster topic subsets and the average-topic subsets, across each topic set size. Figure 2 shows the results for two datasets (the other two show similar behavior). In the x axis we find the topic set size (that is equal to the number of clusters) while in the y axis we find the values of ρ and τ correlations of MAP values with the full topic set. The one-for-cluster-topic subset does not feature a higher correlation than the average-topic subset; on the contrary, it usually has a lower correlation. This can be caused by the fact that experimentally we found that there is a cluster which has much more topics than the others: this cluster “attracts” the MAP of the average set, penalizing the one-for-cluster set, which takes only one item from that cluster. Furthermore, there is a not negligible overlap between the best and the worst set of topics, as discussed in [2].

We also tried other methods and features to cluster topics: we considered other methods to join clusters (i.e., single, average, Ward); we considered another clustering algorithm, k-means; we changed from cosine distance to other ones (i.e., Euclidean, Manhattan, ...). We used also other topics features: we considered the set formed by the identifiers (and the numbers) of the documents (and the relevant ones) in the pooled set, the textual description of topics (only the title field and the full description), plus a set of experiments in which we considered combinations of all the features and all the parameters. In all these experiments the one-for-cluster topic subset shows the same correlation values (or worse) than the average-topic subset.

These results seem to rule out the possibility of using clustering to select a few good topics, at least when the columns of Table 1 are used as vectors.

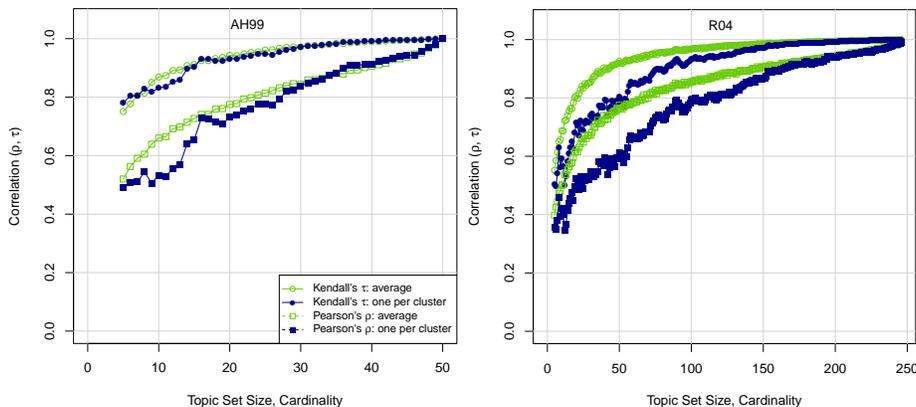


Fig. 2: ρ and τ for one-for-cluster and average set

4 Conclusions and Future Work

In summary, we did not find any evidence supporting our research question clustering does not seem usable to find a good-topic subset. Furthermore, we believe we provided convincing evidence that our results are independent from the parameters used to form the cluster-set or from the topic features used. We started from the hypothesis that clustering can be used to chose a subset of a few informative topics in the evaluation of information retrieval effectiveness. We believe that we have provided convincing evidence that performing clustering is not effective in the choice of a good-topic subset, at least when straightforward clustering strategies are used. It might be that more complex approaches, that we intend to try in the future, provide better results: for example, we did not yet try dimensionality reduction.

Bibliography

- [1] A. Berto, S. Mizzaro, and S. Robertson. On using fewer topics in information retrieval evaluations. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, pages 30–37. ACM, 2013.
- [2] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, 27(4):1–26, 2009.
- [3] G. Tassef, B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic impact assessment of NIST’s text retrieval conference (TREC) program. *Report prepared for National Institute of Technology (NIST)*, 2010.
- [4] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM, 2001.