Do Hard Topics Exist? A Statistical Analysis

(Discussion Paper)

J. Shane Culpepper[1], Guglielmo Faggioli[2], Nicola Ferro[2] and Oren Kurland[3]

[1]*RMIT University, Melbourne, Australia*
[2]*University of Padova, Padova, Italy*
[3]*Technion, Israel Institute of Technology, Haifa, Israel*

Abstract

Several recent studies have explored the interaction effects between topics, systems, corpora, and components when measuring retrieval effectiveness. However, all of these previous studies assume that a topic or information need is represented by a single query. In reality, users routinely reformulate queries to satisfy an information need. Recently there has been renewed interest in the notion of "query variations" which are essentially multiple user formulations for an information need. Like many retrieval models, some queries are highly effective while others are not. In this work, we explore the fundamental problem of studying the interaction components of an IR experimental collection. Our findings show that query formulations have a comparable effect size to the topic factor itself, which is known to be the factor with the greatest effect size in prior ANOVA studies. This suggests that topic difficulty is an artifact of the collection considered and highlights the importance of further research in understanding link between the complexity of a topic and the query rewriting in IR related tasks.

1. Introduction

The interplay between simple keyword queries and large document collections has challenged researchers in *Information Retrieval (IR)* for more than half a century. Some queries are highly effective, while others perform poorly, and changing the ranking models to compensate for *difficult* queries can have negative effects on the performance of queries that were performing well previously. This notion of *query difficulty* has received a great deal of attention over the years. For example, NIST ran the Robust Track in 2004 and 2005 to reexamine sets of queries which had performed poorly across all systems evaluated in the Ad hoc track [2]. It is clear that certain queries challenge even the best performing systems. The distinction between a topic (information need) and a query can have a profound impact on the effectiveness of retrieval, as well as how IR researchers typically categorize and compare system performance [3, 4]. In this paper, we reexamine the idea of query difficulty from the topic perspective, where a topic can have many different query formulations and the retrieval system and the underlying document collection can change. We explore this issue by addressing the following research questions:

RQ1: How does the formulation of a topic impact system performance *within corpora*?

RQ2: How does the formulation of a topic impact system performance *across corpora*?

RQ3: How does topic difficulty vary *across corpora* based on the formulation of a topic?

RQ1 allows us to investigate the effect size of topics and query formulations with respect to

IIR 2021: The 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy

This is an extended abstract of [1]

Table 1
A summary of the effect sizes in within-corpora ANOVA.

	Robust 2004	CORE 2017	CORE 2018
Topic	0.7639	0.8215	0.7834
Formulations (Topic)	0.6941	0.6833	0.6038
System	0.1080	0.2193	0.1445
Topic*System	0.3385	0.3510	0.4386

systems and their components in order to better understand what contributes to topic difficulty and the magnitude of the effects. RQ2 extends RQ1 by looking at what happens across corpora and allows us to also explore corpora-specific topic / query formulations. Finally, RQ3 examines the topic difficulty across multiple corpora. To address our research questions we develop a set of *ANalysis Of VAriance (ANOVA)* models which allow us to break down the overall system performance into topic, query formulation, system, and corpora effects.Additionally, to investigate RQ3 more deeply, we measure variance in arbitrarily ranked topics across corpora. The key idea is that the likelihood of observing arbitrary rank orderings of topics by effectiveness is analogous to topic difficulty being an *intrinsic* property. That is, high volatility in topic ordering suggests that topic hardness is not absolute. Rather it is an artifact of system / corpora interaction. Our experiments highlights that the idea of a single topic being difficult is an artifact of collection design, and the topic difficulty can reliably be circumvented through careful query reformulation. This is a promising step in a fundamentally important problem in IR — that of robust system effectiveness. The paper is organized as follows: Section 2 discusses the experimental setup and the experimental findings; finally, Section 3 draws some conclusions and outlooks for future work.

2. Experiments

Data and Methods. We used the following collections: TREC Robust 2004 Ad Hoc, TREC Common CORE 2017, and TREC Common CORE 2018 for our experiments. The Robust Ad Hoc track contains approximately 528K documents; the TREC 2017 Common CORE contains over 1.8 million articles; finally, the TREC Common CORE 2018 track roughly containing 600K news articles. A large seed set of 3,402 human curated query formulations originally developed using the TREC Robust 2004 Ad Hoc search collection were used in our experiments [5].

RQ1: Effect Size of Query Formulation within Corpora.

To determine the impact of the topics, their formulations and the systems on the overall performance, we run an ANOVA [6, 7] on a *Grid of Points (GoP)* of 144 different systems (9 ranking functions, 2 stemmers, 4 query expansion approaches and 2 stoplists). We also include in the ANOVA model the interaction between the systems and the topics. Table 1 provides a summary of the ANOVA effect size for each factor using each corpus separately. We observe similar performance trends across all of them. All factors are statistically significant. The topic factor has a large-size effect size, and it is indeed the largest effect for this configuration. We can also clearly see that query formulations also have a large effect size in our experiments

Table 2
Across-corpora ANOVA summary table. The letter in the column "effect size" indicates whether the effect is large (L), medium (M) or small (S).

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	effect size
Topic	1544.69	24	64.36	22370.77	0	0.7682	L
Formulation (Topic)	804.82	350	2.30	799.25	0	0.6330	L
System	87.80	143	0.61	213.41	0	0.1579	L
Corpus	33.04	2	16.52	5742.72	0	0.0662	M
System*Topic	216.04	3432	0.06	21.88	0	0.3067	L
System*Form.	240.36	50050	0.00	1.67	0	0.1713	L
System*Corpus	28.57	286	0.10	34.72	0	0.0562	S
Topic*Corpus	960.75	48	20.01	6956.96	0	0.6733	L
Form.*Corpus	527.09	700	0.75	261.72	0	0.5298	L
Topic*System*Corpus	214.42	6864	0.03	10.86	0	0.2946	L
Error	287.99	100100	0.00				
Total	4945.58	161999					

– approaching the topic effect size – suggesting that query formulations strongly influence topic difficulty. Overall, query formulation has the second largest effect, with nearly 1.5 times the size of the topic*system interaction which has historically been a point of emphasis in similar performance comparisons. This provides important evidence that query formulation is crucial in retrieval effectiveness, and has deeper implications in rethinking the way many IR experiments currently formalize query / topic difficulty.

RQ2 and RQ3: Effect Size of Query Formulation across Corpora. To study the effect of the topics and query formulations across corpora, we slightly change the ANOVA model, by adding the additional corpus factor. Table 2 shows the results for the ANOVA across-corpora. The introduction of the new factor allows to compute the effect of some interactions between factors that could not be computed previously. All the factors are statistically significant. We can observe that the topic factor has a large-size effect even across corpora and that the system factor becomes a moderately large-size effect, being bigger than in the single corpus case (see Table 1); the corpus factor has a medium-size effect. We also note that the query formulation factor has a remarkably large-size effect, even across corpora, observed here for the first time, suggesting it is a key contributor to topic difficulty. Both the query formulation*corpus interaction and the topic*system*corpus interaction, observed here for the first time, are clearly important large-size effects. Overall, these findings provide further evidence supporting the possibility that difficult topics do not actually exist in any absolute sense.

Topic Difficulty. To determine whether the difficulty is an intrinsic property of the topic we execute the following experiment. We randomly sample 20,000 permutations of the topics. For each of such permutations and for each pair system-corpus, using a greedy approach we select formulation to represent each of the topics to maximize the correlation between the ranking of the topics based on the *Average Precision (AP)* and the random permutation. We then select, among all the rankings of topics, the one that maximizes the correlation with the random permutation of topics. We finally compute the Kendall's τ correlation between the sampled and the constructed rankings of topics. We have a mean Kendall's τ of 0.85, indicating that the

queries selected to induce the desired topic rankings were consistently close to the arbitrary target ordering. This is a strong evidence that topics can be "arbitrarily" easy or difficult across many different formulations, corpora and system combinations. This is empirical evidence that topic difficulty is not an intrinsic property of an information need – meaning that query formulation based on a corpus and retrieval system, can be combined to sort topics arbitrarily based on a performance goal. How can such a finding help researchers develop more effective retrieval systems? Firstly, it is worth noting that current evaluation paradigms usually consider a single formulation for each topic. Such an arrangement prevents us to observe system behavior with small changes to each query. We believe that multiple formulations are a key omission in our current evaluation campaigns, and we are hopeful future campaigns will incorporate them into their methodology. If our goal is to model real performance of systems, collections creators should explore how to best include multiple formulations of each topic. Our isolation of multiple formulations of topics has allowed us to study in detail the concept of "topic difficulty", which is construct of a specific retrieval configuration – the collection, the system and the query which represents the topic – and not a property intrinsic to the topic alone.

3. Conclusion

In this work, we have presented a comprehensive ANOVA analysis that compares the effect sizes across multiple corpora and retrieval system configurations. We have also generalized previous model configurations in order to incorporate a new nesting factor which maps an information need (topic) to multiple query formulations. The removal of the constraint of a 1:1 mapping between a query and a topic has led to several interesting observations which have important implications on the notion of topic difficulty. We also propose an analysis methodology, based on a permutation algorithm, to further explore topic difficulty. Based on this new knowledge, we were able to show conclusive evidence that topic difficulty is not an intrinsic property and therefore query formulations should be included in future evaluation campaigns.

References

[1] J. S. Culpepper, G. Faggioli, N. Ferro, K. Oren, Topic difficulty: Collection and query formulation effects, Transactions on Information Systems (2021).

[2] E. M. Voorhees, Overview of the trec 2004 robust retrieval track., in: Proc. TREC, 2004.

[3] P. Bailey, A. Moffat, F. Scholer, P. Thomas, User Variability and IR System Evaluation, in: Proc. of SIGIR, 2015, pp. 625–634.

[4] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An enhanced evaluation framework for query performance prediction, in: Proc. ECIR, 2021, pp. 115–129.

[5] R. Benham, J. S. Culpepper, Risk-reward trade-offs in rank fusion, in: Proc. ADCS, 2017, pp. 1–8.

[6] D. Banks, P. Over, N.-F. Zhang, Blind Men and Elephants: Six Approaches to TREC data, Information Retrieval 1 (1999) 7–34.

[7] G. Faggioli, N. Ferro, System effect estimation by sharding: A comparison between anova approaches to detect significant differences, in: Proc. of ECIR, 2021.