Comparing ANOVA Approaches to Detect Significantly Different IR Systems*

Discussion Paper

Guglielmo Faggioli[1], Nicola Ferro[1]

[1]University of Padova, Padova, Italy

Abstract

The ultimate goal of the evaluation is to understand when two IR systems are (significantly) different. To this end, many comparison procedures have been developed over time. However, to date, most reproducibility efforts focused just on reproducing systems and algorithms, almost fully neglecting to investigate the reproducibility of the methods we use to compare our systems. In this paper, we focus on methods based on ANalysis Of VAriance (ANOVA), which explicitly model the data in terms of different contributing effects, allowing us to obtain a more accurate estimate of significant differences. In this context, we compare statistical analysis methods based on "traditional" ANOVA (`tANOVA`) to those based on a bootstrapped version of ANOVA (b`ANOVA`) and those performing multiple comparisons relying on a more conservative Family-wise Error Rate (FWER) controlling approach to those relying on a more lenient False Discovery Rate (FDR) controlling approach. Our findings highlight that, compared to the `tANOVA` approaches, b`ANOVA` presents greater statistical power, at the cost of lower stability.

1. Introduction

Comparing IR systems and identifying when they are significantly different is a critical task for both industry and academia [2, 3, 4, 5]. The literature still lacks reproducibility studies on the statistical tools used to compare the performance of such systems and algorithms. Using reproducible statistical tools is crucial to drawing robust inferences and conclusions. In this context, ANalysis Of VAriance (ANOVA) [6] is a widely used technique, where we model performance as a linear combination of factors, such as topic and system effects, and, by developing more and more sophisticated models, we accrue higher sensitivity in determining significant differences among systems. We focus on two recently developed ANOVA models, `bANOVA`, developed by Ferro and Sanderson [7] and `tANOVA`, developed by Voorhees et al. [8]. Voorhees et al. [8] used sharding of the document corpus to obtain the replicates of the performance score for every (topic, system) pairs needed to develop a model accounting not only for the main effects, but also for the interaction between topics and systems; Voorhees et al. also used an ANOVA version based on residuals bootstrapping [9], which we call b`ANOVA`. Similarly, Ferro and Sanderson [7] used document sharding as well but they developed a more

* The full paper has been originally presented at ECIR 2021 [1]

IIR 2021 – 12th Italian Information Retrieval Workshop, June 29th – July 1st, 2022, Milano, Italy

comprehensive model, based on traditional ANOVA, which also accounts for the shard factor, the shard*system interaction, and the topic*shard interaction; we call this approach tANOVA. Another fundamental aspect to consider when comparing several IR systems is the need to adjust for *multiple comparisons* [10, 11]. Indeed, when comparing just two systems, significance tests control the *Type-I error* at the significance level α. However, when c simultaneous tests are carried out, the probability of committing at least one Type-I error increases up to $1 - (1 - \alpha)^c$. To correct for the multiple comparisons problem, Voorhees et al. adopted a lenient False Discovery Rate (FDR) correction by Benjamini and Hochberg [12]; Ferro and Sanderson used a conservative Family-wise Error Rate (FWER) correction, using the Honestly Significant Difference (HSD) method by Tukey [13]. In conclusion, we identified three aspects that can impact the reproducibility of the above-mentioned ANOVA approaches: *i)* the strategy used to obtain replicates, *ii)* the kind of ANOVA used, and *iii)* the control procedure for the pairwise comparisons problem. Our work investigates behaviour of tANOVA and bANOVA (Voorhees et al. [8]) under different experimental settings – with respect to the above-mentioned focal points – and the generalizability of their results.

2. Experimental Analysis

2.1. Experimental Setup & ANOVA Models

Akin to Voorhees et al., we used two collections: the TREC-3 Adhoc track [14] and TREC-8 Adhoc track [15]. In this work we report results only on TREC-8, we refer to [1] for all the experiments. We use Average Precision (AP) as performance measure. We consider three ANOVA models: (MD1) : a traditional two-way ANOVA that accounts only for the topic and the system factors; (MD2) : A second model, similar to the previous one, that considers also the interaction between topics and systems; (MD3) : A third model that includes also the shard factor and all the interactions between different factors.

2.2. Impact of the multiple comparison strategies and bootstrapping

To investigate the differences between ANOVA approaches, our first analysis compares the number of statistically significantly different (s.s.d.) system pairs found by them. We consider the following multiple comparison procedures: HSD for tANOVA, as originally proposed by Ferro and Sanderson, indicated with tANOVA(HSD); Benjamini-Hochberg (BH) for bANOVA, as originally proposed by Voorhees et al., indicated with bANOVA(BH); and, BH for tANOVA, indicated with tANOVA(BH). tANOVA with Benjamini-Hochberg correction is here employed and analyzed for the first time. It takes the p-values on the difference between levels of the factors produced by the traditional ANOVA, but corrects them using the BH correction. The rationale behind it is that it enjoys the statistical properties provided by the ANOVA while granting a higher discriminative power, due to the BH correction procedure. `zero` has been used as interpolation strategy; in Section 2.3 we empirically show that the interpolation strategy has a negligible effect on the results. Finally, we experiment all the models from (MD1) to (MD3) with all the ANOVA approaches; note that (MD3) has not been studied before for bANOVA and this represents another generalizability aspect. Table 1 reports the results averaged over the five samples of

Table 1
s.s.d. pairs of systems for different ANOVA approaches, using AP.

Model	Approach	bANOVA(BH)	tANOVA(BH)	tANOVA(HSD)
(MD1)	bANOVA(BH)	6866.60 ± 36.965	329.20 ± 22.027	2275.80 ± 39.844
	tANOVA(BH)	-	6537.40 ± 57.107	1946.60 ± 23.190
	tANOVA(HSD)	-	-	4590.80 ± 75.850
(MD2)	bANOVA(BH)	7231.80 ± 51.085	375.20 ± 17.436	2133.40 ± 70.456
	tANOVA(BH)	-	6856.60 ± 65.859	1758.20 ± 54.580
	tANOVA(HSD)	-	-	5098.40 ± 113.429
(MD3)	bANOVA(BH)	7563.40 ± 15.273	262.00 ± 11.681	1655.80 ± 25.377
	tANOVA(BH)	-	7301.40 ± 11.734	1393.80 ± 32.585
	tANOVA(HSD)	-	-	5907.60 ± 37.359

Table 2
Average number of Passive Disagreements (PD) for ANOVA model (MD2) and (MD3).

Model	Approach	Interp.	zero	lq	mean	one
((MD2))	tANOVA(HSD)	zero	230.60± 21.55	23.00± 15.21	100.20± 74.45	89.80± 82.47
		lq	—	239.20± 22.56	77.20± 62.86	85.60± 96.98
		mean	—	—	253.20± 32.18	124.40± 92.81
		one	—	—	—	265.80± 53.21
	bANOVA(BH)	zero	282.60± 13.70	5.80 ± 3.45	41.60 ± 24.44	33.20 ± 28.83
		lq	—	280.80± 12.99	35.80 ± 21.12	32.60 ± 30.75
		mean	—	—	285.00 ± 13.24	49.20 ± 40.73
		one	—	—	—	288.40 ± 18.59
((MD3))	tANOVA(HSD)	zero	222.60± 15.392	0.00± 0.000	0.00± 0.000	0.00± 0.000
		lq	—	222.60± 15.392	0.00± 0.000	0.00± 0.000
		mean	—	—	222.60± 15.392	0.00± 0.000
		one	—	—	—	222.60± 15.392
	bANOVA(BH)	zero	279.20± 16.60	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
		lq	-	279.20± 16.60	0.00 ± 0.00	0.00 ± 0.00
		mean	-	-	279.20± 16.60	0.00 ± 0.00
		one	-	-	-	279.20± 16.60

shards together with their confidence interval. Numbers on the diagonal of Table 1 describe how many pairs of systems are considered s.s.d. by a given approach; numbers above the diagonal are the additional s.s.d. pairs found by one method with respect to the other. Table 1 shows that, as the complexity of the model increases from (MD1) to (MD3), the pairs of systems deemed significantly different increase as well, confirming previous findings in the literature. tANOVA(HSD) controls tANOVA(BH) since all the s.s.d. pairs for tANOVA(HSD) are significant also for tANOVA(BH); this was expected since FWER controls FDR [16]. It is possible see this by considering the differences between approaches (above diagonal): by summing the difference between tANOVA(HSD) and tANOVA(BH) to the tANOVA(HSD) you obtain back the number of s.s.d. pairs identified by tANOVA(BH). However, this pattern holds also for bANOVA(BH) and tANOVA(BH), i.e. all the s.s.d. pairs of tANOVA(BH) are s.s.d. pairs for bANOVA(BH) too. While the relation between BH and HSD was expected, this finding sheds some light on the difference between using a traditional or a bootstrapped version of ANOVA. In summary, most of the increase in the s.s.d. pairs is due to the correction procedure rather than the use of bootstrap.

2.3. Stability of ANOVA Models with respect to Different Interpolation Values

Both tANOVA and bANOVA are based on the concept of "corpus sharding": divide the corpus in non-overlapping subcopora, and use those to compute the systems performance. Problems arise when a shard do not contain any relevant documents, since several Information Retrieval (IR)

measures are not defined. Thus, we study the impact of the interpolation strategy, i.e. how to substitute missing values for topics without any relevant document on a given shard, for the different approaches. If a shard does not contain any relevant document for a topic, we interpolate the missing value using 4 possible strategies: `zero`; `lq`, the value of the lower quartile of the measure scores; `mean`, the average value of the measure scores; and, `one`. To assess the stability with respect to interpolation strategies, we resample shards 5 times ans we consider the number of Passive Disagreements (PD), i.e. the number of pairs of systems A and B for which an approach considers A to be significantly better than B on a sample but A is not significantly better than B on the other sample. Here, for space reasons, we report only the results for tANOVA(HSD) and bANOVA(BH), being the tANOVA(BH) midway between these two. Table 2 reports the average PD counts together with their confidence interval for models (MD2) and (MD3). Values on the diagonal are the average PD observed using the same interpolation strategy, but over the pairs of shards samples. The upper triangle of the Table contains the average PD when using two different interpolation values. Table 2 shows what happens if, using model (MD2) by Voorhees et al., instead of re-sampling shards we use an interpolation value. We can note that, as the interpolation value increases, the PD count on the diagonal tends to increase too. When it comes to the upper triangles, we interestingly find that bANOVA(BH) is much less sensitive to the interpolation values than tANOVA(HSD), being the PD counts substantially lower. The bootstrapped version of ANOVA (bANOVA) appears to be less stable with respect to the resharding (higher diagonal values). This phenomenon is likely due to its greater discriminative power: since a small evidence for bANOVA is enough to assess when two systems are different, the random resharding might produce spurious evidence and thus large variation among different samples. In the part of Table 2 concerning (MD3), both tANOVA(HSD) and bANOVA(BH) have upper triangle equal to zero, and thus are independent from the interpolation values. Indeed, the bANOVA approach samples the residuals and Ferro and Sanderson proved that they are independent of the interpolation value for (MD3). Therefore, using (MD3) also the bootstrap approach by Voorhees et al. does not need to re-sample shards.

3. Conclusions and Future Work

In this work, we compared bANOVA [8] and tANOVA approaches under different conditions. We found out that tANOVA tends to be more robust than bANOVA with respect to the actual random shards used, suggesting more reliability in drawing the same conclusions. On the other hand, when using partial ANOVA models like (MD2) which are not able to deal with shards without relevant documents, bANOVA is more robust than tANOVA to the chosen interpolation value. Regarding the multiple comparison strategy, we have found that tANOVA with HSD is more restrictive than bANOVA but tANOVA with BH correction behaves similarly to bANOVA. Overall, we can conclude that, the decision of the model and the correction technique depends on the final aim of the researcher. If stability is more important, tANOVA(HSD) is preferable, since it is more stable with respect to random shards and less computationally expensive. Conversely, if the focus is on the number of pairs, bANOVA(BH) gives the maximum boost, at the price of lower stability for random shards. Future work will investigate the use of uneven-size random shards, instead of the even-size ones used in the literature so far.

References

[1] G. Faggioli, N. Ferro, System effect estimation by sharding: A comparison between anova approaches to detect significant differences, in: European Conference on Information Retrieval, Springer, 2021, pp. 33–46.

[2] D. A. Hull, Using Statistical Testing in the Evaluation of Retrieval Experiments, in: Proc. SIGIR, 1993, pp. 329–338.

[3] J. Savoy, Statistical Inference in Retrieval Effectiveness Evaluation, Information Processing & Management 33 (1997) 495–512.

[4] B. A. Carterette, Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments, ACM Trans. Inf. Syst 30 (2012) 4:1–4:34.

[5] J. S. Culpepper, G. Faggioli, N. Ferro, O. Kurland, Topic difficulty: Collection and query formulation effects, ACM Transactions on Information Systems 40 (2021).

[6] A. Rutherford, ANOVA and ANCOVA. A GLM Approach, 2nd ed., John Wiley & Sons, New York, USA, 2011.

[7] N. Ferro, M. Sanderson, Improving the Accuracy of System Performance Estimation by Using Shards, in: Proc. SIGIR, 2019, pp. 805–814.

[8] E. M. Voorhees, D. Samarov, I. Soboroff, Using Replicates in Information Retrieval Evaluation, ACM Trans. Inf. Syst 36 (2017) 12:1–12:21.

[9] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall/CRC, USA, 1994.

[10] N. Fuhr, Some Common Mistakes In IR Evaluation, And How They Can Be Avoided, SIGIR Forum 51 (2017) 32–41.

[11] T. Sakai, On Fuhr's Guideline for IR Evaluation, SIGIR Forum 54 (2020) p14:1–p14:8.

[12] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, J. Royal Stat. Soc. 57 (1995) 289–300.

[13] J. W. Tukey, Comparing Individual Means in the Analysis of Variance, Biometrics 5 (1949) 99–114.

[14] D. K. Harman, Overview of the Third Text REtrieval Conference (TREC-3), in: Proc. TREC, 1994, pp. 1–19.

[15] E. M. Voorhees, D. K. Harman, Overview of the Eigth Text REtrieval Conference (TREC-8), in: Proc. TREC, 1999, pp. 1–24.

[16] J. C. Hsu, Multiple Comparisons. Theory and methods, Chapman and Hall/CRC, USA, 1996.