How Topic and System Size Affect the Correlation among Evaluation Measures*

Nicola Ferro

University of Padua, Italy
ferro@dei.unipd.it

Abstract. In this paper, we investigate the effect of topic and system sizes on the correlation among evaluation measures for both τ and τ_{AP}. We found that topic size matters more than system size and that τ and τ_{AP} does not lead to noticeably different rankings among measures.

1 Introduction

Correlation analysis plays a central role in *Information Retrieval (IR)* evaluation where it is one of the tools we use to study properties and relationships among evaluation measures. When a new evaluation measure is proposed, correlation analysis is used to assess how the new measure ranks IR systems with respect to the other existing measures and, thus, to understand whether it actually grasps different aspects of the systems and its introduction is somehow motivated. In this context, the most used correlation coefficients are the Kendall's tau correlation τ [4] and the AP correlation τ_{AP} [6].

In this paper, we investigate what is the effect of the number of systems and topics on the correlation among evaluation measures and what are the differences in using τ or τ_{AP}.

In order to answer these research questions, we rely on 3 different *Text REtrieval Conference (TREC)* collections and, for each collection, we create a *Grid of Points (GoP)* [2, 3], i.e. a set of system runs originating from all the possible combinations of the following components: 6 different stop lists, 6 types of stemmers, 7 flavors of n-grams, and 17 distinct IR models, leading to 1,326 distinct system run. These GoPs basically represent nearly all the state-of-the-art components which constitute the common denominator almost always present in any IR system for English retrieval.

We consider 8 different evaluation measures – namely, AP, P@10, Rprec, RBP, nDCG, nDCG@20, ERR, and Twist – and we compute the correlation among them over the created GoPs. Finally, we use *General Linear Mixed Model (GLMM)* and *ANalysis Of VAriance (ANOVA)* [5] to conduct the analyses needed to answer the above research questions.

The paper is organized as follows: Section 2 introduces the GLMM used for the analyses; Section 3 discusses the experimental findings; finally, Section 4 draws some conclusions and provides an outlook for future work.

* Extended abstract of [1].

IIR 2018, May 28-30, 2018, Rome, Italy.

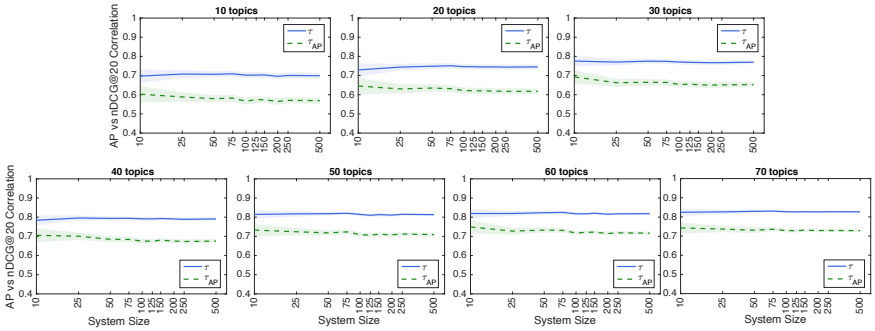

Fig. 1. `AP vs nDCG@20`: each plot shows the correlation, both τ and τ_{AP}, for a given number of topics as the number of systems increases.

2 Model

We create a GoP using the TREC 13, 14, and 15 Terabyte track, thus containing 149 topics and 1,326 runs. For each topic size $t \in T = \{10, 20, 30, 40, 50, 60, 70\}$ and system size $s \in S = \{10, 20, 50, 75, 100, 125, 150, 200, 250, 500\}$, we independently draw $H = 100$ random samples of t topics and $H = 100$ random samples of s systems from the the GoP. Overall, for each combination $(t, s) \in T \times S$ of topic and system sizes and for each measure pair, this procedure originates $H = 100$ samples of correlation values for both τ and τ_{AP}.

We use the following model

$$Y_{ijkl} = \underbrace{\mu.... + \kappa_i + \alpha_j + \beta_k + \gamma_l}_{\text{Main Effects}} + \underbrace{(\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}} \quad (1)$$

where: κ_i is the effect of the i-th subject, i.e. one of the $h = 1, \ldots, H$ samples; α_j is the effect of the j-th factor, i.e. measure pairs; β_k is the effect of the k-th factor, i.e. number of topics; γ_l is the effect of the l-th factor, i.e. number of systems; $(\alpha\beta)_{jk}$, $(\alpha\gamma)_{jl}$, and $(\beta\gamma)_{kl}$ are, respectively, the interactions between measures pairs and number of topics, measure pairs and number of systems, and number of topics and number of systems; and, ε_{ijkl} is the error.

3 Experimental Results

General Trends As Figure 1 highlights, the number of topics affects both τ and τ_{AP}, since their average value increases as the number of topics increases. On the other hand, the number of systems exhibits less impact on the two correlation coefficients: indeed, apart from a small transient up to around 75-100 systems, the trend for both coefficients is somehow constant, especially when the number of topics increases. We can note how, in the transient phase, τ and τ_{AP} behave differently: τ tends to slightly increase before reaching stability while τ_{AP} manifests an initial decrease, sometimes followed by an increase, before getting more or less constant.

Table 1. Kendall's τ correlation: ANOVA table for the GLMM model of equation (1).

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	Power
Subject	55.8432	99	0.5641	99.6711	0.0000		
Measure Pair	1,666.5150	27	61.7228	10,906.3664	0.0000	0.6004	1.0000
Topic Size	418.7689	6	69.7948	12,332.6910	0.0000	0.2740	1.0000
System Size	2.3875	9	0.2653	46.8753	3.64e-85	0.0021	0.9999
Measure Pair*Topic Size	33.1886	162	0.2049	36.2001	0.0000	0.0283	1.0000
Measure Pair*System Size	0.9346	243	0.0038	0.6796	1.0000	0.0000	0.8043
Topic Size*System Size	0.6136	54	0.0114	2.0079	1.69e-05	0.0002	0.5137
Error	1,105.8283	195,399	0.0057				
Total	3,284.0798	195,999					

Table 2. AP correlation τ_{AP}: ANOVA table for the GLMM model of equation (1).

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$	Power
Subject	71.0670	99	0.7178	86.7603	0.0000		
Measure Pair	2,536.3318	27	93.9382	11,353.5200	0.0000	0.6100	1.0000
Topic Size	612.0528	6	102.0088	12,328.9432	0.0000	0.2740	1.0000
System Size	12.0979	9	1.3442	162.4638	4.20e-308	0.0074	1.0000
Measure Pair*Topic Size	26.9371	162	0.1662	20.0967	0.0000	0.0155	1.0000
Measure Pair*System Size	1.2495	243	0.0051	0.6214	1.0000	0.0000	0.7467
Topic Size*System Size	0.8735	54	0.0162	1.9550	3.44e-05	0.0002	0.4966
Error	1,616.7174	195,399	0,0083				
Total	4,877.3271	195,999					

When it comes to confidence intervals, lower number of topics and systems call for larger intervals, which is not surprising. However, τ generally exhibits smaller confidence intervals than τ_{AP}, especially for low number of topics. Moreover, τ seems to be a bit more effective than τ_{AP} in benefiting from the increasing number of topics and systems; indeed, correlation values get more stable and confidence intervals get smaller in a "faster" way for τ than for τ_{AP}.

ANOVA Analysis Tables 1 and 2 report the results of the ANOVA analyses on the GLMM model of equation (1) for τ and τ_{AP}, respectively. The most prominent effect is the measure pair one, which is a large size effect in terms of $\hat{\omega}^2$, and it has almost the same size for both τ and τ_{AP}. The second biggest effect is the topic size one, which again is a large size effect and it has the same size for both τ and τ_{AP}. This supports the previous observations about Figure 1 when we noted that the topic size is the most prominent factor influencing the correlation among evaluation measures. Finally, the system size effect, even if significant, is a very small size effect and we can consider it almost negligible; however, it should be noted that this effect is a little bit more than three times bigger for τ_{AP} than for τ. Overall, this sustains the observations made above about the smaller importance of the number of systems on the correlation among evaluation measures, with τ_{AP} being more sensitive to this factor than τ.

When it comes to the interaction between effects, for both τ and τ_{AP}, the measure pair and topic size $(\alpha\beta)_{jk}$ and the topic size and system size $(\beta\gamma)_{kl}$ interactions are statistically significant. On the other hand, the measure pair and system size $(\alpha\gamma)_{jl}$ interaction is not significant and this further stress the fact that the number of systems does not influence much the correlation among evaluation measures.

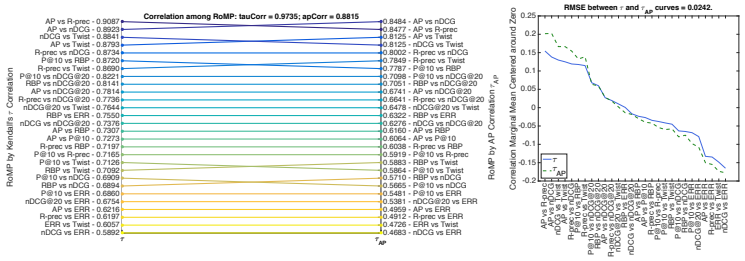

Fig. 2. Comparison between τ and τ_{AP} in terms of how they rank evaluation measures.

τ and τ_{AP} Comparison Figure 2 on the left shows the rankings of measures according to τ and τ_{AP}: we can note how there are very few swaps and always among adjacent rank positions. On the right, we show the actual correlation values but with means centered around zero: it is evident how close are τ and τ_{AP}, apart from a constant offset; indeed the *Root Mean Square Error (RMSE)* among the two curves is just 0.0242, indicating very small differences.

Overall, these findings suggest that, if you consider a set of evaluation measures and you compare them across a large set of topic and system sizes, removing those effects, τ and τ_{AP} have different absolute values but they provide a quite consistent assessment of what the differences among these measures are.

4 Conclusions and Future Work

We investigated how topic and system size affect the correlation among evaluation measures. We discovered that the number of topics impacts more than the number of systems and that the number of systems does not cause the correlation to steadily increase but it reaches a stable point quite quickly. We also observed that the behavior of τ and τ_{AP} is quite consistent when comparing a whole set of evaluation measures, yet producing different absolute correlation values.

As future work, we plan to investigate how the different system components, e.g. stop lists, stemmers, IR models, affect the correlation among evaluation measures.

References

1. Ferro, N.: What Does Affect the Correlation Among Evaluation Measures? TOIS 36(2), 19:1–19:40 (2017)
2. Ferro, N., Harman, D.: CLEF 2009: Grid@CLEF Pilot Track Overview. In CLEF 2009. pp. 552–565. LNCS 6241 (2010)
3. Ferro, N., Silvello, G.: Toward an Anatomy of IR System Component Performances. JASIST 69(2), 187–200 (2018)
4. Kendall, M.G.: Rank correlation methods. Griffin, Oxford, England (1948)
5. Rutherford, A.: ANOVA and ANCOVA. A GLM Approach. John Wiley & Sons, New York, USA (2011)
6. Yilmaz, E., Aslam, J.A., Robertson, S.E.: A New Rank Correlation Coefficient for Information Retrieval. In SIGIR 2008. pp. 587–594. (2008)