Component-Based Evaluation using GLMM*

Nicola Ferro and Gianmaria Silvello

University of Padua, Italy
{ferro,silvello}@dei.unipd.it

Abstract. Topic variance has a greater effect on performances than system variance but it cannot be controlled by system developers who can only try to cope with it. On the other hand, system variance is important on its own, since it is what system developers may affect directly by changing system components and it determines the differences among systems. In this paper, we face the problem of studying system variance in order to better understand how much system components contribute to overall performances. To this end, we propose a methodology based on *General Linear Mixed Model (GLMM)* to develop statistical models able to isolate system variance, component effects as well as their interaction.

1 Introduction

The experimental results analysis is a core activity in *Information Retrieval (IR)* aimed at, firstly, understanding and improving system performances and, secondly, assessing our own experimental methods, such as robustness of experimental collection or properties of the evaluation measures. When it comes to explaining system performances and differences between algorithms, it is commonly understood [2] that system performances can be broken down to a reasonable approximation as

$$\text{system performances} = \text{topic effect} + \text{sys effect} + \text{topic/sys interaction effect}$$

even though it is not always possible to estimate these effects separately, especially the interaction one.

It is well-known that topic variability is greater than system variability and a lot of effort has been put in better understanding this source of variance [2] as well as in making IR systems more robust to it. Nevertheless, with respect to an IR system, topic variance is a kind of "external source" of variation, which cannot be controlled, but can only be taken into account to better deal with it. On the other hand, system variance is a kind of "internal source" of variation, since it is originated by the choice of system components, may be directly affected by developers by working on them, and represents the intrinsic differences between algorithms.

* This is an extended abstract of [1]. Please refer to the original paper for the full model and experimental results.

Currently, in experimental evaluation we consider system variance as a single monolithic contribution and we cannot break it down into the smaller pieces (the components) constituting an IR system.

We propose a methodology, based on *General Linear Mixed Model (GLMM)* and *ANalysis Of VAriance (ANOVA)* [3], to address this issue and to estimate the effects of the different components of an IR system, thus giving us better insights on what system variance and system effects are. In particular, the proposed methodology allows us to break down the system effect into the contributions of stops lists, stemmers or n-grams and IR models, as well as to study their interaction.

In this extended abstract we report the main ideas behind the adopted methodology and the main results we obtained from the experimental evaluation conducted on standard *Text REtrieval Conference (TREC)* Ad-hoc collections.

2 Methodology and Experimentation

The goal of the proposed methodology is to decompose the effects of different components on the overall system performances. In particular, we are interested in investigating the effects of the following components: stop lists; *Lexical Unit Generator (LUG)*, namely stemmers or n-grams; IR models, such as the vector space or the probabilistic model.

We considered three main components of an IR system: stop list, LUG and IR model. We selected a set of alternative implementations of each component and by using the Terrier open source system we created a run for each system defined by combining the available components in all possible ways. The components we selected are:

stop list: nostop, indri, lucene, smart, terrier;
stemmer: nolug, weak Porter, Porter, Krovetz, Lovins;
model: BB2, BM25, DFRBM25, DFRee, DLH, DLH13, DPH, HiemstraLM, IFB2, InL2, InexpB2, InexpC2, LGD, LemurTFIDF, PL2, TFIDF.

We conducted single factor and three-factors ANOVA tests for both the groups on TREC 05, 06, 07, 08, 09 and 10 collections, and by employing the following five measures: AP, P@10, nDCG@20, RBP and ERR@20.

The full GLMM model for the described factorial ANOVA for repeated measures with three fixed factors (stoplist α, stemmers β, models γ) and a random factor (topics τ) is:

$$Y_{ijkl} = \underbrace{\mu_{....} + \tau_i + \alpha_j + \beta_k + \gamma_l}_{\text{Main Effects}} + \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma_{kl} + \alpha\beta\gamma_{jkl}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}}$$

In Figure 1 we can see a graphical representation of the main analyses we conducted by running the ANOVA tests on the grids of points described above. We report only the plots for the TREC 09 and 10 collections. Here, we show three main plots: Tukey HSD plot, main effect plot and interaction effect plots.

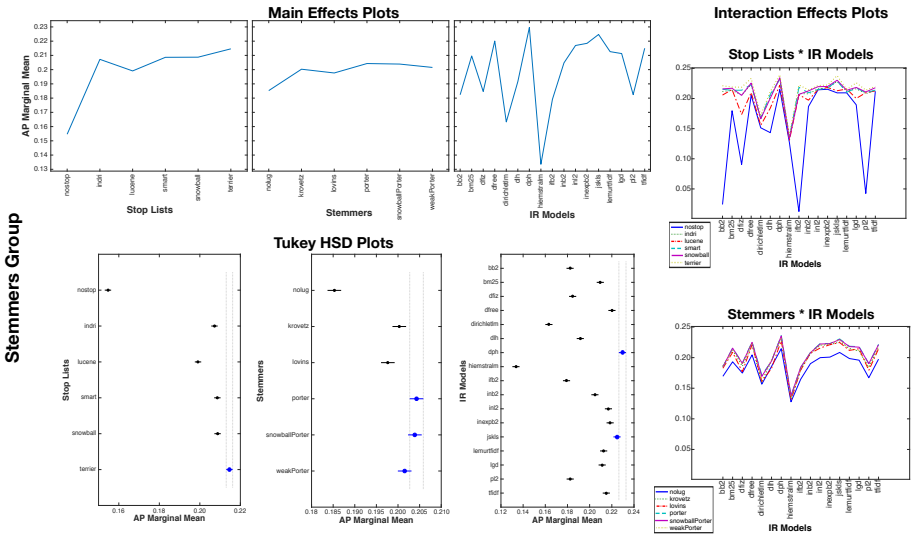

Fig. 1. TREC 09-10 Web search: main effects, interaction effects, and Tukey HSD plots for average precision on stemmers group.

In the Tukey HSD plots, each point represents the mean performances of an approach, where best approaches are at the top of the figure. We also show both the main effects and the interaction effects plots in order to get a better appreciation of the behaviour of the different levels of each factor. By means of this plot we can easily determine the impact of the different levels of a factor. An interaction effects plot displays the levels of one factor on the X axis and has a separate line for the means of each level of the other factor on the Y axis; it allows us to understand whether the effect of one factor depends on the level of the other factor.

From the main effects and Tukey *Honestly Significant Difference (HSD)* plots of the stop lists in Figure 1, we can see that, there is a substantial difference between systems employing or not a stop list, while the top performing stop list is `terrier`, i.e. the longest one, followed by a second group constituted by `snowball`, `smart`, and `indri`; `lucene` is still the lowest performing stop lists. The fact that there is a clearer distinction between stop lists than in the news search case, and that the longest stop list is the top performing one, lead us to hypothesize that the noisy Web context benefits more from the aggressive filtering of a longer stop list.

As far as stemmers are concerned, the Porter-based stemmers constitute the top group in the case of Web search, while `krovetz` and `lovins` stay together in the second group, well above the group employing no stemmer at all. With respect to the news search case, the less aggressive stemmers perform better for Web search and this may be motivated again by the hypothesis that the noisy Web context benefits more from avoiding further noise due to over-stemming.

3 Discussion and main results

In general, from the experimental analysis we have seen that linguistic preprocessing and linguistic resourcesare very important and contributed pretty much to the effectiveness of an IR system. So, the role of the stop list is significant as well as choosing between stemmers or n-grams.

In particular, we have seen that the choice of the stop list does not make a big difference with respect to use or not use a stop list; indeed, we have seen that there are no significant differences between the "indri", "smart" and "terrier" stop lists, whereas the "lucene" stop list (which is composed by 15 words) is significantly different from the other three.

The main effect of the stemmer is always significant even though its size is quite small; nevertheless, there is a tangible difference between systems using or not using a stemmer. In particular, we observe that there is no significant difference between the Porter and the Krovetz stemmer which are the stemmers with the highest impact on variance followed by the weak Porter and the Lovins ones.

For all the collections, consistently across the measures and both for the stemmer and the n-grams group, the higher effect size is reported by the *stop list*model* interaction effect which is always of medium or large size. This effect shows us that the variance of the systems is explained for the bigger part by the stop list and the model components. The *stop list*stemmer* interaction effects are always not significant and a very similar trend can be observed for the *stemmer*model* interaction effect.

It is interesting to note that the second order interactions for the n-grams group are all statistically significant and that, in particular, we can see that n-grams, differently than the stemmers, have a bigger effect on the stop list than on the IR model.

We observe that different measures see the stop lists in a comparable way in terms of effect size. This is valid also for the stemmer, with the exception of ERR@20 for which it has an almost negligible effect size even though it is statistically significant. For the n-grams group all the measures are comparable and ERR@20 is not as low as it happens for the stemmers.

References

1. N. Ferro and G. Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM Press, New York, USA, 2016.
2. S. E. Robertson and E. Kanoulas. On Per-topic Variance in IR Evaluation. In W. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 891–900. ACM Press, New York, USA, 2012.
3. A. Rutherford. *ANOVA and ANCOVA. A GLM Approach.* John Wiley & Sons, New York, USA, 2nd edition, 2011.