On Theoretically Valid Score Distributions in Information Retrieval

Ronan Cummins and Colm O'Riordan

Dept. of Information Technology,
National University of Ireland, Galway, Ireland
ronan.cummins@nuigalway.ie

Abstract. In this paper, we aim to investigate the practical usefulness of the Recall-Fallout Convexity Hypothesis (RFCH) for a number of document score distribution (SD) models. We compare SD models that do not automatically adhere to the RFCH to modified versions of the same SD models that do adhere to the RFCH. We compare these models using the inference of average precision as a measure of utility. For the three models studied in this paper, we conclude that adhering to the RFCH is practically useful for the two-normal model, makes no difference for the two-gamma model, and degrades the performance of the two-lognormal model.

1 Background

Document score distribution (SD) models offer a mathematically sound approach to modelling a document ranking in information retrieval (IR). This is because the entire ranking (and relevance information) is conflated to a fixed number of parameters in the SD model. The Recall-Fallout Convexity Hypothesis (RFCH) [3] has been proposed as a possibly useful constraint for valid SD models. The RFCH states that as we traverse a ranked-list, the recall should always be greater than fallout. This seems theoretically valid, as IR systems should at least provide a better than random ranking. More interestingly, this hypothesis constrains the parameters of certain models of score distributions. Therefore, in this paper we aim to investigate the practical usefulness of the RFCH. We do this by comparing a number of five-parameter SD models that do not automatically adhere to the RFCH, to modified four-parameter versions of the same SD models that do adhere to the RFCH.

2 Score Distributions and Parameter Estimation

Similar to previous approaches [3, 2, 1], we model a single ranking of document scores as a binary mixture of relevant and non-relevant documents, where a mixture parameter is the proportion of relevant documents R in the entire returned set N (i.e. $\lambda = \frac{R}{N}$). In this work, we use method-of-moments estimates (MME) to estimate the parameters of the model from an actual ranking (using labelled

data). In order to create SD models that adhere to the RFCH, certain parameters of both distributions must be constrained. As the set of non-relevant document scores (NR) is such a large sample of documents, it is justifiable to rely on the moments calculated from this sample. However, the sample of relevant document scores (R) is often very small, and therefore, we deem it justifiable to modify the moments of this sample to force the model, that will be estimated from the moments, to adhere to the RFCH. For all of the approaches in this paper, we modify the sample variance of the relevant scores (v_1) to enable the model to adhere to the RFCH, while ensuring that the remaining sample means and variances (m_1, m_0, and v_0) are calculated directly from the respective samples (i.e. relevant and non-relevant)[1].

The **Two-Normal (N_1N_0) Model** has been shown to adhere to the RFCH only when the variances of both relevant and non-relevant distributions are equal (i.e. $\sigma_1 = \sigma_0$) [3]. These variance parameters are very rarely equal when the variances are estimated from the sample variances. Therefore, to enforce this mixture to conform to the RFCH, the sample mean for both relevant and non-relevant documents (m_1 and m_0) are used as the mean of the both distributions respectively (μ_1 and μ_0), and the sample variance of the non-relevant documents (v_0) is used as both variance parameters ($\sigma_1^2 = \sigma_0^2$).

The **Two-Gamma (G_1G_0) Model** has been shown to adhere to the RFCH when either the shape parameter for both distributions are equal ($k_0 = k_1$), or when both scale parameters are equal ($\theta_0 = \theta_1$)[2] [3]. The MME estimates for the gamma distribution are $\theta = v/m$ and $k = m^2/v$. Therefore, to force this model to adhere to the RFCH, the sample variance for the relevant scores should be modified to $v_1 = v_0 \cdot m_1/m_0$ before the method-of-moments estimates are calculated to ensure that $\theta_1 = \theta_0$.

The **Two-Lognormal (L_1L_0) Model** adheres to the RFCH when the variance parameter for both distributions are equal ($\sigma_0 = \sigma_1$) [2]. The MME estimates for the gamma distribution are $\sigma^2 = ln(1 + v/m^2)$ and $\mu = ln(m) - 0.5 \cdot ln(1+v/m^2)$. Therefore, to ensure that $\sigma_1 = \sigma_0$, the sample variance for the relevant scores should be modified to $v_1 = v_0 \cdot m_1^2/m_0^2$ before the method-of-moments estimates are calculated.

Therefore, for each initial five parameter SD model that does not adhere to the RFCH, we can create a corresponding four parameter SD model that does adhere to the RFCH. It is obvious that these four parameter models are less flexible than their five parameter counterparts in terms of their *goodness-of-fit*. However, we do not know if these modified four parameter models have any

[1] When using maximum-likelihood estimates similar assumptions must be made to effectively link the parameters of both distributions.

[2] We also conducted experiments that ensured that $k_1 = k_0$ and determined that setting both scale parameters (θ) to be equal was more beneficial for the inference of average precision.

advantages in terms of their ability to correctly model relevance information (as measured by the ability of a model to infer average precision).

3 Experiments

To measure the performance of a particular SD model in terms of IR utility, we compared the average precision inferred from the SD model with the actual average precision of that ranking. We do this over a set of queries and use both Pearson's and Kendall's τ correlation measures to measure how well the output of a particular model (i.e. inferred average precision) agrees with the actual average precision. Average precision can be inferred from an SD model by calculating the area under the precision-recall curve [2] and is a natural candidate as a measure of performance for a number of theoretical reasons [4]. Furthermore, as different IR systems create different rankings, we conducted the same experiment over 14 different IR systems so that our results would be more general. The IR systems used were TFIDF, Pivoted document normalisation (with s=0.01, s=0.05, and s=0.2), BM25 (with b=0.25, b=0.5, and b=0.75), divergence-from-randomness model (PL2 with c=1, c=2, and c=5), two language models (Dirichlet priors and Jelinek-Mercer smoothing), F2EXP (axiomatic term-weighting), and ES (a learned term-weighting model).

3.1 Results and Discussion

Figure 1 shows box plots of Kendall's τ correlation on the 14 systems for three SD models that are not forced to adhere to the RFCH, and the modified version of the SD models that are forced to adhere to the RFCH on four collections. We can clearly see that for the two-normal model adhering to the RFCH is beneficial, as a marked increase in performance is indicated. For the two-gamma model, in general there is no loss in performance when the model is forced to adhere to the RFCH. However, for the two-lognormal model there is a decrease in performance. These results are consistent when using a linear correlation as the measure of performance (not shown due to space limitations). This is an interesting outcome as each modified SD model is less complex than its five-parameter counterpart. Overall, the five-parameter two-lognormal model is the best performing model. However in general, when looking for the best theoretically valid SD model, we can see that the two-gamma model tends to slightly outperform the valid two-lognormal model.

4 Conclusion and Future Work

This paper has shown that adhering to the RFCH is beneficial for some SD models. There is a degradation in performance for one of the SD models when it adheres to the RFCH. In general, this empirical validation of the RFCH is significant due to the fact that models that adhere to the RFCH have a reduced number of parameters, and therefore are inherently less complex.

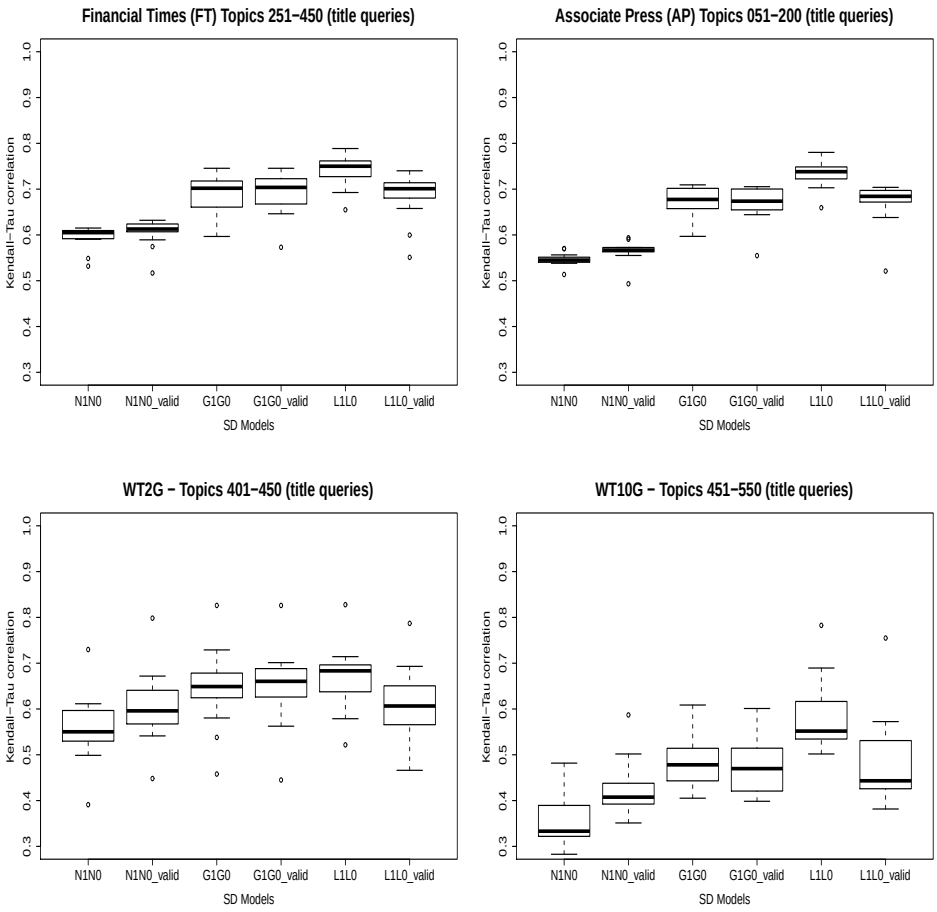

Fig. 1. Kendall's τ Correlations for mixtures that violate the RFCH and those that adhere to the RFCH for title queries on two Newswire and two Web collections

References

1. Avi Arampatzis and Stephen Robertson. Modeling score distributions in information retrieval. *Inf. Retr.*, 14(1):26–46, 2011.
2. Ronan Cummins. Measuring the ability of score distributions to model relevance. In *Proceedings of the 6th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, UAE, December 18-20*, 2011.
3. Stephen Robertson. On score distributions and relevance. In *ECIR 2007*, pages 40–51, Berlin, Heidelberg, 2007. Springer-Verlag.
4. Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. Extending average precision to graded relevance judgments. In *SIGIR '10*, SIGIR '10, pages 603–610, New York, NY, USA, 2010. ACM.