Investigating Performance Predictors Using Monte Carlo Simulation and Score Distribution Models

Ronan Cummins
Department of Information Technology
National University of Ireland, Galway
ronan.cummins@nuigalway.ie

ABSTRACT

The standard deviation of scores in the top k documents of a ranked list has been shown to be significantly correlated with average precision and has been the basis of a number of query performance predictors [8, 6, 3]. In this paper, we outline two hypotheses that aid in understanding this correlation. Using score distribution (SD) models with known parameters, we create a large number of document rankings using Monte Carlo simulation to test the validity of these hypotheses.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval: Retrieval models

Keywords: Monte Carlo Simulation, Score Distributions

1. DOCUMENT SCORE DISTRIBUTIONS

Many works [2, 1] have shown that document scores can be modelled using score distributions (SD), while others [5] have shown that unexpected observations in large collections can be explained using SD models. Therefore, and in particular, we assume document scores can be drawn from an SD model comprised of two-lognormal distributions, where $f(s|1)$ and $f(s|0)$ are the probability density function of the relevant (1) and non-relevant (0) document scores respectively, and where λ is the mixing parameter (i.e. $f(s) = \lambda \cdot f(s|1) + (1-\lambda) \cdot f(s|0)$). As a consequence of the recall-fallout convexity hypothesis (RFCH) [7], each lognormal SD model can be described using four parameters $\{\ \mu_1,\ \mu_0,\ \sigma_1 = \sigma_0,\ \lambda\}$ where $\mu_1 > \mu_0$ and $\sigma_1 = \sigma_0$. Recent research [4] has shown that the RFCH implies the following relationship[1] between the moments in a two-lognormal model:

$$\frac{E[s_1]}{E[s_0]} = \frac{\sqrt{Var(s_1)}}{\sqrt{Var(s_0)}} \tag{1}$$

where $E[s_1]$, $E[s_0]$, $Var(s_1)$, $Var(s_0)$ are the expected value (E) and variance (Var) of relevant (1) and non-relevant (0) document scores respectively. It can be seen that the RFCH implies that the standard deviation of relevant scores (monotonically related to variance) is proportional to the expected value of the relevant scores. All else being equal, it is intuitive to assume that the larger the expected value of the relevant scores is (i.e. $E[s_1]$), the higher average precision will be.

1.1 Hypotheses

We now outline two hypotheses that aim to explain the correlation between average precision and the standard deviation in the head of a ranked list:

H1: As a consequence of the RFCH, the standard deviation at the head of a ranked list is positively correlated with the mean score of relevant documents, which in turn is positively correlated with average precision.

H2: A lower standard deviation of document scores in the head of a ranked list indicates that the separation between the relevant and non-relevant distributions is low, and therefore there is a higher contamination of non-relevant documents in the head of the list. This leads to a lower average precision.

2. MONTE CARLO SIMULATION

We now investigate these hypotheses using Monte Carlo simulation. We simulate rankings by drawing samples of document scores from SD models with known parameters. In particular, to test the first hypothesis (**H1**), we simulate rankings returned for 50 queries. Each ranking is drawn from an SD model with a different μ_1 value **(variable $\boldsymbol{\mu_1}$)** ranging uniformly from 1.5 to 2.5, while $\mu_0 = 1.5$, $\sigma_0 = 0.25$, and λ remains fixed[2]. This experimental setting changes the distribution (i.e. the mean score) of relevant documents while keeping the other parameters constant. The Kendall-τ correlation of average precision with the standard deviation-at-k documents for the 50 rankings is recorded. We repeat this process 50 times to ensure that the resultant correlation coefficients are not spurious. The average Kendall-τ is reported.

To test the second hypothesis (**H2**), we simulate rankings for another 50 queries. Each of these rankings is drawn from an SD model with a different μ_0 value ranging uniformly from 1.5 to 2.5 **(variable $\boldsymbol{\mu_0}$)** while $\mu_1 = 2.5$, $\theta_0 = 0.25$, and λ remains fixed. This experimental setting changes the distribution (i.e. the mean score) of non-relevant documents

[1] A two-gamma model is another suitable SD model that yields similar results to those in this paper. The moment relationship is similar for a two-gamma SD model that adheres to the RFCH. In that model, the variance replaces the standard deviation (i.e. $\frac{E[s_1]}{E[s_0]} = \frac{Var(s_1)}{Var(s_0)}$).

[2] These values were chosen by fitting a two-lognormal SD model to scores returned by a Language Model run on actual TREC data (disks 4 and 5)

Copyright is held by the author/owner(s).
SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

while keeping the remaining parameters static. This essentially increases the contamination of non-relevant documents in the head of a ranking. Again we record the Kendall-τ correlation of average precision with the standard deviation-at-k documents and average the correlation for 50 simulations.

We perform both of these experiments, **(variable μ_1)** and **(variable μ_0)** for different values of λ (i.e. the parameter controlling the portion of relevant documents drawn). We wanted each sample ranking to contain approximately 50 relevant documents[3] (R) and so we drew different sample sizes (N) from each SD model so that we simulate returned sets of various sizes. In particular, we set $N = \{250, 500, 1000, 2000, 4000, 8000, 16000, 32000, 64000, 128000\}$ and so the mixing parameters was $\lambda = 50/N$ for each value of N.

3. RESULTS

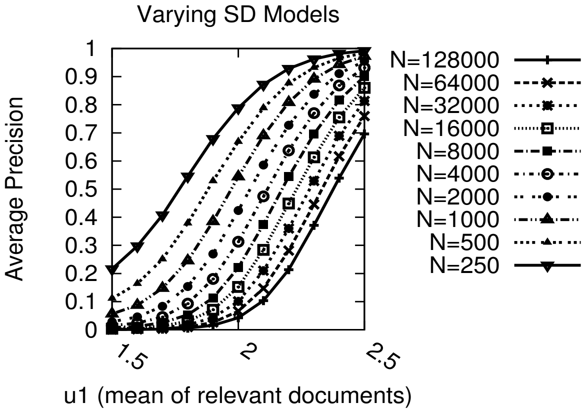

Figure 1: Average precision of rankings drawn from SD models with $\sigma_0 = 0.25$, $\mu_0 = 1.5$ and variable μ_1 for various mixing levels $\lambda = 50/N$.

Each point in Figure 1 shows the mean average precision of 50 simulated rankings drawn from an SD model. Plotted are SD models for various μ_1 and λ values. Intuitively, we can see that the average precision of the rankings increase as both λ and μ_1 increase. This shows that our simulated rankings cover a large range of average precision values.

Table 1 shows the average Kendall-τ correlation of average precision with the standard deviation-at-k for the simulated rankings for both experimental settings (**variable** μ_1 and **variable** μ_0). We can see that when only **varying** μ_1, the correlation coefficient is high and significant for all values of λ. This is a consequence of the RFCH (equation 1) and is similar for many values of k (not shown due to space limitations).

We can also see that there is a significant correlation between standard deviation-at-100 and average precision when **varying** μ_0 for most settings of λ. By examining our simulated rankings, we have determined that this is because the standard deviation is measured at a value $k = 100$, which is higher than the number of relevant documents (R) in the average ranking. Therefore, the standard deviation-at-k has the potential to measure the score of all relevant documents but also includes the score of several non-relevant documents when $k > R$. Thus, the lower the score of these non-relevant documents compared to the relevant documents, the better the query (i.e. the degree of separation between μ_1 and μ_0 is measured when $k > R$). It can be seen that the correlation of standard deviation with average precision is lower when $k = 25$. Also shown is the correlation when $k = 400$, which shows that including too many documents in the standard deviation calculation leads to a lower correlation when only varying μ_0. A high deviation can in some cases indicate that the non-relevant documents have contaminated the relevant documents to a high degree, due to the right-skewed nature of score distributions. This can lead to a negative correlation between standard deviation-at-k and average precision when λ is low and k is high.

Table 1: Average Kendall-τ correlation of average precision of simulated rankings with the standard deviation of scores for top k documents

R=50		Ave. Kendall-τ			
		variable μ_1	**variable μ_0**		
N	λ	$k = 100$	$k = 400$	$k = 100$	$k = 25$
250	0.2000	0.838	-	0.663	0.238
500	0.1000	0.833	0.230	0.629	0.297
1000	0.0500	0.828	0.295	0.586	0.331
2000	0.0250	0.803	0.225	0.523	0.349
4000	0.0125	0.780	0.138	0.463	0.326
8000	0.0063	0.762	0.045	0.384	0.325
16000	0.0031	0.729	-0.058	0.312	0.298
32000	0.0016	0.708	-0.147	0.217	0.291
64000	0.0008	0.686	-0.250	0.144	0.251
128000	0.0004	0.660	-0.335	0.058	0.229

4. CONCLUSION

We have presented two hypothesis regarding the relationship between average precision and the standard deviation in the head of a ranking. Furthermore, we performed an analysis using SD models that indicates the conditions under which these hypotheses can be deemed true.

Acknowledgments

The author is funded by the Irish Research Council for Science, Engineering and Technology (IRCSET), co-funded by Marie-Curie Actions under FP7.

5. 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 *AIRS*, pages 25–36, 2011.

[3] Ronan Cummins, Joemon Jose, and Colm O'Riordan. Improved query performance prediction using standard deviation. In *SIGIR 2011*, pages 1089–1090, New York, NY, USA, 2011. ACM.

[4] Ronan Cummins and Colm O'Riordan. On theoretically valid score distribution in information retrieval. In *ECIR 2012*, pages 1089–1090, Barcelona, Spain, 2012. ACM.

[5] David Madigan, Yehuda Vardi, and Ishay Weissman. Extreme value theory applied to document retrieval from large collections. *Inf. Retr.*, 9:273–294, June 2006.

[6] Joaquín Pérez-Iglesias and Lourdes Araujo. Standard deviation as a query hardness estimator. In *SPIRE*, pages 207–212, 2010.

[7] Stephen Robertson. On score distributions and relevance. In *ECIR*, pages 40–51, 2007.

[8] Anna Shtok, Oren Kurland, and David Carmel. Predicting query performance by query-drift estimation. In *ICTIR*, pages 305–312, 2009.

[3] For TREC disks 4 and 5 and for many TREC collections, this is approximately the average number of relevant documents per topic.