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 aim to explain the 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.

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 $\lambda$ 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 between the moments in a two-lognormal model:

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

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)

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 $\mu_1$ value (variable $\mu_1$) ranging uniformly from 1.5 to 2.5, while $\mu_0 = 1.5$, $\sigma_0 = 0.25$, and $\lambda$ remains fixed2. This experimental setting changes the distribution (i.e. the mean score) of relevant documents while keeping the other parameters constant. The Kendall-$\tau$ 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-$\tau$ 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 $\mu_0$ value ranging uniformly from 1.5 to 2.5 (variable $\mu_0$) while $\mu_1 = 2.5$, $\theta_0 = 0.25$, and $\lambda$ remains fixed. This experimental setting changes the distribution (i.e. the mean score) of non-relevant documents

2These 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)
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-\(\tau\) correlation of average precision with the standard deviation-at-\(k\) documents and average the correlation for 50 simulations.

We perform both of these experiments, (\textit{variable} \(\mu_1\)) and (\textit{variable} \(\mu_0\)) for different values of \(\lambda\) (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

![Varying SD Models](image)

Figure 1: Average precision of rankings drawn from SD models with \(\sigma_0 = 0.25\), \(\mu_0 = 1.5\) and \textit{variable} \(\mu_1\) 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 \(\mu_1\) and \(\lambda\) values. Intuitively, we can see that the average precision of the rankings increase as both \(\lambda\) and \(\mu_1\) increase. This shows that our simulated rankings cover a large range of average precision values.

Table 1 shows the average Kendall-\(\tau\) correlation of average precision with the standard deviation-at-\(k\) for the simulated rankings for both experimental settings (\textit{variable} \(\mu_1\) and \textit{variable} \(\mu_0\)). We can see that when only \textit{varying} \(\mu_1\), the correlation coefficient is high and significant for all values of \(\lambda\). 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 \textit{varying} \(\mu_0\) for most settings of \(\lambda\). 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 the 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 \(\mu_1\) and \(\mu_0\) 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 \textit{varying} \(\mu_0\). 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 \(\lambda\) is low and \(k\) is high.

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

<table>
<thead>
<tr>
<th>(R = 50)</th>
<th>variable (\mu_1)</th>
<th>variable (\mu_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>(k = 100)</td>
<td>(k = 400)</td>
</tr>
<tr>
<td>250</td>
<td>0.2009</td>
<td>0.838</td>
</tr>
<tr>
<td>500</td>
<td>0.1005</td>
<td>0.833</td>
</tr>
<tr>
<td>1000</td>
<td>0.0500</td>
<td>0.828</td>
</tr>
<tr>
<td>2000</td>
<td>0.0250</td>
<td>0.803</td>
</tr>
<tr>
<td>4000</td>
<td>0.0125</td>
<td>0.780</td>
</tr>
<tr>
<td>8000</td>
<td>0.0063</td>
<td>0.762</td>
</tr>
<tr>
<td>16000</td>
<td>0.0031</td>
<td>0.729</td>
</tr>
<tr>
<td>32000</td>
<td>0.0016</td>
<td>0.708</td>
</tr>
<tr>
<td>64000</td>
<td>0.0008</td>
<td>0.686</td>
</tr>
<tr>
<td>128000</td>
<td>0.0004</td>
<td>0.660</td>
</tr>
</tbody>
</table>

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