Predicting Query Performance Directly from Score Distributions

Ronan Cummins

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

Abstract. The task of predicting query performance has received much attention over the past decade. However, many of the frameworks and approaches to predicting query performance are more heuristic than not. In this paper, we develop a principled framework based on modelling the document score distribution to predict query performance directly.
In particular, we (1) show how a standard performance measure (e.g. average precision) can be inferred from a document score distribution. We (2) develop techniques for query performance prediction (QPP) by automatically estimating the parameters of the document score distribution (i.e. mixture model) when relevance information is unknown. Therefore, the QPP approaches developed herein aim to estimate average precision directly. Finally, we (3) provide a detailed analysis of one of the QPP approaches that shows that only two parameters of the five-parameter mixture distribution are of practical importance.

1 Introduction

Query performance prediction (QPP) has become an important problem in the area of information retrieval (IR). These predictors aim to automatically estimate the performance of queries so that different strategies (e.g. query expansion or reduction) can be applied based on their estimated performance. The performance of these predictors are usually compared by measuring the correlation between the output of the predictor and query performance (e.g. average precision). However, many approaches to QPP are unprincipled, and it is unclear how to improve their performance, or if their performance can even be improved.

In this paper, we develop a principled framework based on modelling document score distributions that aims to predict query performance directly. Fig. 1 shows an example of a document score distribution returned for a query (when relevance information is known). We (1) develop formulae that directly infer average precision from a document score distribution, (2) develop simple heuristics that can estimate the *important* parameters of the score distribution when relevance information is unknown, and (3) provide an analysis that informs us of the most important parameters in the distributional model. This analysis helps in narrowing the focus of future research.

The remainder of the paper is organised as follows: Section 2 reviews related work on score distributions and query performance prediction (QPP). Section 3 outlines our principled model, before the formulae for calculating the average precision from a mixture are introduced. In section 4, we outline three approaches to automatically predict the performance of a query from the score distribution when relevance information is unknown. Furthermore, we present an analysis that shows that only two parameters of the model are crucial in the estimation of average precision. Section 5 presents comparative results of the newly developed QPP approaches versus existing predictors. Finally, section 6 outlines our conclusions and future work.

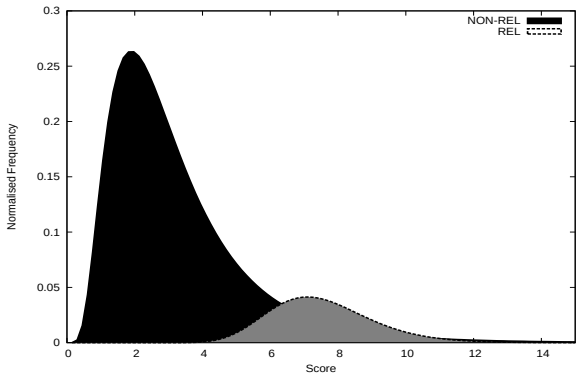

Fig. 1. A Typical Distribution of Scores Returned from a Classical IR System

2 Related Work

Modelling the distribution of document scores returned from IR systems has been studied from a theoretical perspective since the early days of IR [13]. More recently renewed interest has led to research that uses score distributions for data fusion [14]. Other researchers have modelled document score distributions for threshold filtering [1]. Others [9] have studied the generation process of the score distribution and have provided reasons for the typical shape (Fig. 1) of the distribution.

Automatically predicting query performance can aid information retrieval systems by enabling these systems to apply different strategies (e.g. query expansion) to queries of varying difficulty. One of the earliest approaches to QPP has been that of the clarity score [4], which measures the KL-divergence between the query and collection model in a language modelling framework. Some approaches [15] have measured the robustness of a ranking to perturbations and

have developed novel predictors from this, while others [7] have investigated the clustering ability of similarly ranked documents to develop predictors.

Recent research has shown that the standard deviation (σ) (i.e. dispersion) of scores in a ranked-list is a good predictor of query performance [10, 12, 5]. These approaches are more heuristic based and lack a deeper theoretical understanding. The performance of predictors are usually measured by calculating the correlation (i.e. linear and/or non-parametric) between the output of the predictor and the performance of the query (i.e. usually average precision) over a set of queries.

However, to the authors knowledge, to date there has been no research that has directed aimed to estimate the performance of a query (either using score distributions or other methods). While some predictors use document scores returned from a system, and use various measures of the dispersion of such scores to develop their predictors, the methods are unprincipled and do not aim to directly predict performance, rather some surrogate of performance.

3 Explicitly Modelling Query Performance

In this section, we present a mixture distribution that is used in this paper to model the scores of both relevant and non-relevant documents.

3.1 Assumptions and Mixture Model

Consider an IR system that retrieves a returned set of N documents, and thus N scores given a query (Q). We assume that a system ranks documents independently of each other, in accordance with the probability ranking principle (PRP) [11] and that the relevance judgments are binary.

The log-normal distribution has been used successfully [14] to model scores for fusion tasks in IR, and therefore, we adopt this distribution[1]. The probability density function (pdf) of the log-normal distribution is as follows:

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}} \tag{1}$$

where μ and σ are the parameters. This distribution is supported from 0 to ∞ and the cumulative density function (cdf) is again simply the integral of this function from 0 to ∞. The mean of the distribution is $e^{\mu+\sigma^2/2}$, while the variance is $(e^{\sigma^2}-1)\cdot(e^{2\mu+\sigma^2})$. Therefore, by rewriting these equations the method-of-moments estimates (MME) are as follows:

$$\hat{\mu} = ln(m) - \frac{1}{2}(1+\frac{v}{m^2}) \qquad \hat{\sigma}^2 = ln(1+\frac{v}{m^2}) \tag{2}$$

where m and v are the sample mean and variance respectively. Therefore, similar to previous approaches, the document score distribution can be thought of as a mixture of relevant and non-relevant documents as follows:

[1] Noting that any reasonable choice of distribution can be substituted into the mixture

$$P(s) = (\lambda) \cdot P(s|1) + (1 - \lambda) \cdot P(s|0) \tag{3}$$

where $P(s|1)$ is the probability density function (pdf) for the scores (s) of relevant documents, $P(s|0)$ is the pdf for the scores of non-relevant documents, and where $\lambda = \frac{R}{N}$ is the proportion of relevant documents R in the entire returned set N.

3.2 Inferring Average Precision

We will now show how average precision (a standard metric for the effectiveness of a query) can be calculated directly from the mixture of continuous distributions. As recall is the proportion of relevant returned documents compared to the entire number of relevant documents, the recall at score s can be defined as follows:

$$recall(s) = \int_s^\infty \frac{\lambda \cdot P(s|1) \cdot ds}{\lambda} = \int_s^\infty P(s|1) \cdot ds \tag{4}$$

which is the cumulative density function (cdf) of the distribution of relevant documents (viewed from ∞). Under the distributions outlined earlier for our model, we know that $recall(s)$ will vary between 0 and 1, (i.e. when $s = 0$, $recall(s) = 1$ as ensured by the cdf). Similarly, the precision at s (the proportion of relevant returned documents over the number of returned documents) can be defined as follows:

$$precision(s) = \frac{\int_s^\infty \lambda \cdot P(s|1)}{\int_s^\infty (\lambda) \cdot P(s|1) + (1 - \lambda) \cdot P(s|0)} \tag{5}$$

Now that we can calculate the precision and recall at any score s in the range $[0 : \infty]$, we can create a precision-recall curve. Furthermore, as average precision can be estimated geometrically by the area under the precision-recall curve [2], the average precision ($avg.prec$) of a query can be calculated as follows:

$$avg.prec() = \int_0^1 precision(s) \cdot dz(s) \tag{6}$$

where $z(s) = recall(s)$ which is in the range [0:1]. As these expressions are not closed-form, they can be calculated using relatively simple geometric numerical integration methods. It is worth noting that the formulae given for calculating average precision can over-estimate the actual average precision value calculated from TREC runs. This is due to the fact that recall is calculated as the number of relevant documents in the returned set, rather than the total number of relevant documents in the collection.

4 Estimating Parameters Without Relevance Information

In this section, we develop approaches to automatically estimate (i.e. when no relevant information is known) the five parameters of the mixture model (i.e. λ,

μ_1, σ_1, μ_0, σ_0) using a number of different methods. The section is comprised of three approaches to estimating the parameters of the mixture models. The first two approaches are based on heuristics and the MME of parameters. The third approach makes use of the standard EM algorithm for mixture models. We perform an analysis to find the most important parameters in one of the new parameter estimation approaches. Firstly Table 1 outlines the TREC[2] datasets used in this paper.

Table 1. Test Collection Details

Collection		# docs	# topics	range
Tuning	LATIMES	131,896	144	301-450
Test	AP	242,918	149	051-200
	FT	210,158	188	251-450
	WT2G	221,066	50	401-450
	WT10G	1,692,096	100	451-550

4.1 Estimating Moments and Mixture

In this section we aim to estimate the sample moments so that the parameters of the model can, in turn, be automatically calculated using method of moment estimates (MME) from equations (2) (i.e. Section 3.1). Therefore, to estimate the five parameters of the log-normal model using MME, we must estimate the sample mean (m_1 and m_0) and variances (v_1 and v_0) for the relevant and non-relevant document scores and the mixture parameter (λ).

Firstly, as the number of non-relevant documents (NR) is usually much larger than the number of relevant documents (R) in the entire returned set (i.e. $NR \gg R$), we can estimate the sample mean (m_0) and sample variance (v_0) of the non-relevant documents by using the mean and variance of the scores in the entire returned set (i.e. $N \simeq NR$), as this seems a rather sound heuristic. However, the estimation of the mean and variance (m_1 and v_1) of relevant documents is more problematic.

Recent research has posited that a theoretically valid distribution should be able to approach Dirac's delta function under the strong SD hypothesis [1]. Fundamentally, as IR systems are striving to separate the set of relevant documents (R) from the set of non-relevant documents (NR), we estimate the mean (m_1) and variance (v_1) of the relevant set by assuming that all documents over a certain threshold score (min-max normalised for convenience) are relevant. Fig. 2 shows the tuning[3] of this threshold on a separate tuning collection collection (i.e. the LATIMES for both title and desc queries) averaged over five different IR systems (i.e BM25, LM, Pivoted Normalisation, F2EXP [8] and ES [6]). We

[2] http://trec.nist.gov/

[3] During this tuning process, the actual mixture value (λ) is assumed to be known.

can see that a common stable performance (i.e. average Spearman correlation with average precision) for both title and desc queries, can be achieved at a min-max normalised score of around 0.5 (i.e. midway between the minimum and maximum score of a ranked list). Therefore, the sample mean (m_1) and variance (v_1) of the relevant document scores are estimated by calculating the mean and variance of all scores that lie in the top half of a min-max normalised score range.

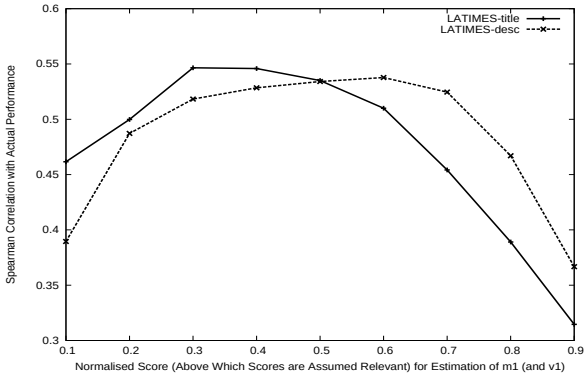

Fig. 2. Estimating m_1 (and v_1) threshold by tuning on LA Times collection averaged over five IR systems

At this stage, we have estimates of the mean and variance of both relevant and non-relevant document scores, and consequently, from these we can calculate four parameters of the mixture model using MME. However, the final parameter that needs to be estimated is the mixture parameter λ. We apply a similar approach as before and assume that all documents over a certain threshold normalised score are deemed relevant[4]. Similar to the previous tuning experiment, Fig. 3 shows the performance of the mixture model at various normalised threshold scores for estimating the mixture parameter λ. We can see that the best performance (i.e correlation with actual average precision) occurs when assuming that very few documents are relevant (i.e. only those scores that are at or above a normalised score of 0.95).

Now we have estimated, albeit heuristically, all the information needed to infer average precision without relying on relevance information. Furthermore, for all future experiments using this MMP1 (method of moments predictor) approach, the threshold for estimating m_1 and v_1 remain at normalised score of 0.5, and the threshold used for estimating λ is a normalised score of 0.95. As we will

[4] Preliminary experiments informed us that the number of documents above the normalised score of 0.5 grossly overestimated the number of actual relevant documents, although the estimates of m_1 and v_1 are suitable. Therefore, a separate and more stringent threshold is needed.

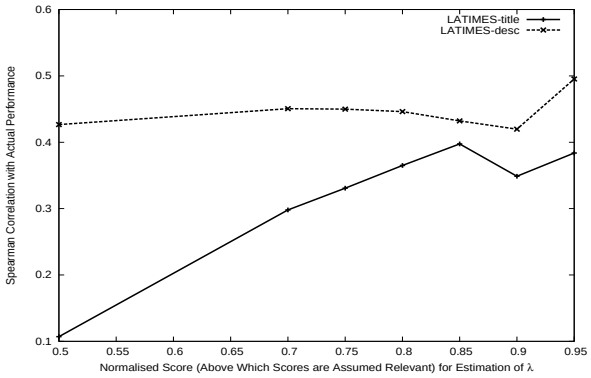

Fig. 3. Estimating λ threshold by tuning on LA Times collection averaged over five IR systems

see in the next section, the estimation of the mean and variance of non-relevant documents (m_0 and v_0) is based on a rather sound heuristic. However, the approach to estimating the mean and variance of the relevance document scores, and the mixture parameter, is where loss in predictive performance can be attributed. We shall see later in the results section (section 5.1) that the estimation from these heuristics yields very good performance compared to other predictors. However, in the next section, we analyse the most important parameters (i.e. m_1, v_1, m_0, v_0, λ) for the MMP1 approach outlined in this section.

4.2 Analysing Moments and Mixture

In this section, we aim to identify the parameters (i.e. m_1, v_1, m_0, v_0, λ) that contribute most to the performance of the model outlined in the previous section (i.e. MMP1). It would be beneficial to know which parameters are more important in terms of performance (i.e. correlation with average precision). We measure the amount of information contained in each parameter by initially assuming that all parameters (i.e. moments and mixture parameters) are as accurate as possible (i.e. using MME when relevance labels are known). We then substitute an estimated version of each parameter (i.e. estimated without relevance labels) and recalculate the performance of the model. At each stage of the process, a parameter is estimated using the heuristics in the previous section (section 4.1). Therefore, when the process is complete, all of the parameters of the model have been estimated without use of relevance information.

Fig. 4 show the results of such a process[5]. As we view the Figure from left to right, an estimate (i.e. without using labelled data) of each parameter is substituted into the model. It is clear from Fig. 4 that the estimation of the mean

[5] The results of all other collections show similar trends

and variance of non-relevant documents (m_0 and v_0), and the variance of relevant documents (v_1) can be accurately estimated using the approach previously outlined (section 4.1), as the correlation coefficient does not decrease. However, when the m_1 and λ parameters are estimated without using relevance information, the correlation coefficient decreases a significant amount. Therefore, the two most important parameters in the model are the mean of relevant document scores (m_1) and the mixture parameter (λ), the former being the most important. These results are averages across five different IR systems. We can report that all of the systems tested behaved very similarly.

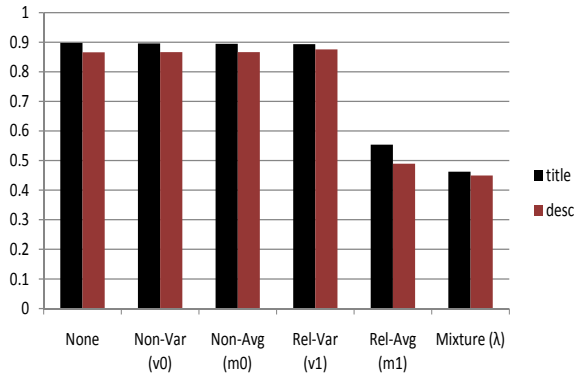

Fig. 4. Decrease in Spearman correlation with actual average precision as moments and mixture are estimated (FT Collection) from unlabelled data

4.3 Motivation and Improvement

While Fig. 4 (and the results later in Section 5.1) show that the first approach (MMP1) to estimating the mean and mixture of the set of relevant documents seems to be effective to some degree, there is little motivation as to why this may be so. The MMP1 approach estimated the mean and variance of relevant documents by using all of the scores above a normalised score of 0.5. Consider a system which returns N documents and where $K = \#(\|S(d)\| > 0.5)$ is the number of documents that are above a min-max normalised score of 0.5 (i.e. they have a score in the top half of the distribution). If K is small it implies that the system has also succeeded in promoting a relatively small number of documents, compared to the returned set N. Given the view of score distributions in Fig. 1, we can see that if the relevant and non-relevant scores are separated to a higher degree, the performance of the query will also be higher. Given that the distribution of document scores from systems is positively skewed, a smaller number of documents in this set of K documents will lead to a higher mean for

the relevant documents (m_1). This in turn is an indicator that there is a good separation between relevant and non-relevant documents (and subsequently an indiction of a good query).

The initial estimate of m_1 was calculated by averaging all the scores above a normalised score of 0.5. A subsequent analysis on the LATIMES tuning collection has informed us that for 80% of the queries, the mean of the set of relevant document scores lies in the top half of the score distribution. However, our initial method of estimating the mean score of relevant documents (m_1) cannot return an estimate below a normalised score of 0.5. Therefore, we now propose a small modification to the initial estimate of m_1 so that a score of below 0.5 can be achieved when it is detected that the distribution of relevant and non-relevant document have not been seperated to a sufficient degree. Given that a small value of K explicitly indicates good separation, the following formula give us an updated measure of the normalised mean score of relevant documents (m_1') using a simple linear combination with the original normalised m_1 estimate:

$$\|m_1'\| = \alpha \cdot (1 - \frac{log(K)}{log(N)}) + (1 - \alpha) \cdot \|m_1\| \tag{7}$$

where K is the number of documents above a normalised score of 0.5, N is the returned set, and α is a parameter we set to 0.5 for all subsequent experiments. The left-hand side of this equation will reduce the estimate of the normalised mean score ($\|m_1'\|$) when K is relatively large. Consider a query which returns $N = 10,000$ documents, for which a relatively large proportion $K = 4,000$ lie in the top half of the distribution. The left-hand side of the equation ($1 - \frac{log(4,000)}{log(10,000)} = 0.099$) will return a low value which can reduce the initial normalised estimate of $\|m_1\|$ below 0.5. The new estimate can be unnormalised to recover a new updated mean m_1'. This updated mean m_1' can be used in place of m_1 in the initial MMP1 approach to yield a second approach (MMP2). A further discussion of the comparative results of these approaches is undertaken in the results section.

4.4 Expectation Maximisation Approach

The EM algorithm is a popular unsupervised learning algorithm for estimating the parameters in mixture models [3]. We initialised the EM algorithm with the parameter estimates from the first MME approach (Section 4.1) that were generated using heuristics. We ran the EM algorithm for 50 iterations. Our initial experiments showed that the parameters converged prior to the 50^{th} iteration.

5 Results and Discussion

In this section we present comparative results of the two QPP approaches based on heuristics that estimate the model parameters via moments (MMP1 and MMP2), and the approach based on the EM algorithm (EM). We then discuss the contributions and limitations of the research undertaken. In the subsequent results we focus on the two most popular IR systems (i.e. BM25 and LM).

5.1 Comparative Results

In this section, we compare the performance of the new QPP approaches developed in Section 4 (labelled MMP1, MMP2 and EM) against other state-of-the-art post-retrieval approaches. The state-of-the-art baseline approaches that we use are the clarity score [4] (a principled approach using KL-divergence), the standard deviation of document scores at 100 ($\sigma(100)$) [10], and NQC [12] also at 100 documents. We also tested the automatically tuned version of the standard deviation [10], and the maximum retrieval score of a ranked list, and found that the baselines presented in Tables 2 and 3 are stronger.

Tables 2 and 3 show the Spearman correlation[6] of the output of each predictor and average precision, for the approaches on four test collections for two prominent IR systems (BM25 and LM). The column labelled 'OPT' is the theoretically maximum correlation of the mixture model, if the parameters could be predicted using the MME from labelled relevance data. We can see that the new MMP approaches outperform the clarity score on most of the collections and, in general, are comparable in performance to that of the best baselines for longer queries. In general, on short queries, the new MMP1 and MMP2 approaches outperform the baselines, with MMP2 noted as the best predictor. We performed statistical tests[7] on the correlation coefficients of the new MMP approaches against both baseline approaches for each collection, and found that on most of the collections, the correlation coefficients were not significantly higher. We can report that when any of the baselines outperformed the MMP approaches the result was not significant, but on some collections, the MMP approaches significantly outperformed one (always the lower) of the baselines (denoted by †). The MMP2 approach tends to outperform the MMP1 approach especially for longer queries. It should be noted that we have not tuned the linear combination (i..e $\alpha = 0.5$) parameter in this approach.

The results of the unsupervised EM learning approach are particularly poor. We analysed the parameters returned from the approach and determined that the EM algorithm tends to grossly over-estimate the mixture parameter (λ), while not estimating values that are close to the actual values for $\hat{\mu_1}$, $\hat{\sigma_1}$, $\hat{\mu_0}$, or $\hat{\sigma_0}$.

It is true that the methods for estimating the parameters of the distributions are heuristic, but these can be removed when more theoretically sound methods for estimating these are discovered. There are many approaches to query performance prediction that have not been evaluated against the new approaches developed here, but comparative studies [10] would tend to suggest that our approach is highly competitive. Furthermore, other approaches to QPP have not aimed to explicitly estimate the performance measure in question. One inter-

[6] Best results are in bold. Due to the sizes of the differences and the number of queries in some of the test collections, statistical tests tend not to find significant differences between most of the correlations. However, for all but the EM approach the individual correlations are significant.

[7] We transformed both coefficients to z-scores and tested whether the 0.95 confidence interval levels overlapped.

Table 2. Spearman correlation of output of various predictors vs average precision for title (top half of table) and desc (bottom half of table) queries for BM25

	BM25						
Collection	clarity	$\sigma(100)$	NQC	EM	MMP1	MMP2	OPT
AP	0.393	0.280	0.265	0.037	**0.511** †	0.495 †	0.87
FT	0.426	0.492	0.513	0.173	**0.596** †	0.590 †	0.88
WT2G	0.352	0.445	0.411	-0.125	0.423	**0.473**	0.82
WT10G	**0.357**	0.328	0.342	-0.031	0.298	0.344	0.74
Avg(title)	0.382	0.386	0.382	0.013	0.457	**0.475**	0.83
AP	0.508	**0.591**	0.543	0.060	0.513	0.571	0.84
FT	0.382	0.431	0.518	-0.025	0.519	**0.543** †	0.86
WT2G	0.321	**0.584**	0.592	-0.129	0.507	0.552	0.81
WT10G	0.400	**0.501**	0.491	-0.042	0.411	0.456	0.72
Avg(desc)	0.402	0.526	**0.536**	-0.034	0.487	0.530	0.81

Table 3. Spearman correlation of output of various predictors vs average precision for title (top half of table) and desc (bottom half of table) queries on for a Jelinek-Mercer Language Model

	LM						
Collection	clarity	$\sigma(100)$	NQC	EM	MMP1	MMP2	OPT
AP	0.387	0.170	0.205	0.184	**0.389** †	0.378 †	0.89
FT	0.467	0.432	0.467	0.105	0.442	**0.469**	0.89
WT2G	0.335	0.467	0.428	-0.158	0.453	**0.514**	0.80
WT10G	0.246	0.276	0.253	0.040	0.523 †	**0.537** †	0.76
Avg(title)	0.358	0.336	0.338	0.042	0.451	**0.474**	0.83
AP	**0.525**	0.519	0.456	-0.038	0.430	0.499	0.86
FT	**0.414**	0.296	0.368	0.002	0.347	0.388	0.87
WT2G	0.249	0.533	0.517	-0.139	0.513	**0.577** †	0.82
WT10G	0.333	**0.567**	0.455	0.017	0.381	0.482	0.75
Avg(desc)	0.380	0.478	0.449	-0.039	0.417	**0.486**	0.83

esting practical advantage of the predictors developed here is that they can be easily modified to predict other performance measures.

6 Conclusion

In this work, we have developed new query performance predictors that explicitly aim to predict average precision. The new predictors (MMP1 and MMP2) based on estimating the moments and mixture parameter are comparable to state-of-the-art predictors. Furthermore, an analysis of the parameters of the predictor has determined that only two parameters (m_1 and λ) are of crucial importance to the performance of the predictor. This analysis aids in narrowing the focus

of future work. In a broader IR sense, it follows that only these two parameters are of importance to any IR application using score distributions. Future work, involves researching other unsupervised learning approaches to parameter estimation in the hope that they may yield higher performance predictors.

Acknowledgments Ronan Cummins is funded by the Irish Research Council (IRCSET), co-funded by Marie Curie Actions under FP7.

References

1. Avi Arampatzis and Stephen Robertson. Modeling score distributions in information retrieval. *Inf. Retr.*, 14(1):26–46, 2011.
2. Javed A. Aslam and Emine Yilmaz. A geometric interpretation and analysis of r-precision. In *CIKM*, pages 664–671, 2005.
3. Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
4. Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *SIGIR '02*, SIGIR '02, pages 299–306, New York, NY, USA, 2002. ACM.
5. Ronan Cummins, Joemon Jose, and Colm O'Riordan. Improved query performance prediction using standard deviations. In *SIGIR*, pages 524–531, 2011.
6. Ronan Cummins and Colm O'Riordan. Learning in a pairwise term-term proximity framework for information retrieval. In *SIGIR*, pages 251–258, 2009.
7. Fernando Diaz. Performance prediction using spatial autocorrelation. In *SIGIR*, pages 583–590, 2007.
8. Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR*, pages 480–487, 2005.
9. Evangelos Kanoulas, Keshi Dai, Virgiliu Pavlu, and Javed A. Aslam. Score distribution models: assumptions, intuition, and robustness to score manipulation. In *SIGIR*, pages 242–249, 2010.
10. Joaquín Pérez-Iglesias and Lourdes Araujo. Standard deviation as a query hardness estimator. In *SPIRE*, pages 207–212, 2010.
11. C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.
12. Anna Shtok, Oren Kurland, and David Carmel. Predicting query performance by query-drift estimation. In *ICTIR*, pages 305–312, 2009.
13. John A. Swets. Information retrieval systems. *Science*, 141(3577):245–250, 1963.
14. Peter Wilkins, Alan F. Smeaton, and Paul Ferguson. Properties of optimally weighted data fusion in cbmir. In *SIGIR '10*, SIGIR '10, pages 643–650, 2010.
15. Yun Zhou and W. Bruce Croft. Ranking robustness: a novel framework to predict query performance. In *CIKM '06*, CIKM '06, pages 567–574, New York, NY, USA, 2006. ACM.