Document Score Distribution Models for Query Performance Inference and Prediction

RONAN CUMMINS, University of Greenwich

Modelling the distribution of document scores returned from an information retrieval (IR) system in response to a query is of both theoretical and practical importance. One of the goals of modelling document scores in this manner is the inference of document relevance. There has been renewed interest of late in modelling document scores using parameterised distributions. Consequently, a number of hypotheses have been proposed to constrain the mixture distribution from which document scores could be drawn.

In this article, we show how a standard performance measure (i.e., average precision) can be inferred from a document score distribution using labelled data. We use the accuracy of the inference of average precision as a measure for determining the usefulness of a particular model of document scores. We provide a comprehensive study which shows that certain mixtures of distributions are able to infer average precision more accurately than others. Furthermore, we analyse a number of mixture distributions with regard to the recall-fallout convexity hypothesis and show that the convexity hypothesis is practically useful.

Consequently, based on one of the best-performing score-distribution models, we develop some techniques for query-performance prediction (QPP) by automatically estimating the parameters of the document score-distribution model when relevance information is unknown. We present experimental results that outline the benefits of this approach to query-performance prediction.

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

General Terms: Measurement, Theory, Performance

Additional Key Words and Phrases: Score distributions, query performance

ACM Reference Format:
Cummins, R. 2014. Document score distribution models for query performance inference and prediction. *ACM Trans. Inf. Syst.* 32, 1, Article 2 (January 2014), 28 pages.
DOI:http://dx.doi.org/10.1145/2559170

1. INTRODUCTION

The analysis of document scores returned from an information retrieval (IR) system in response to a given query is an important consideration in both theory and practice. Since the original work [Swets 1963] that proposed modelling relevant and nonrelevant documents as parameterised score-distribution (SD) models, various practical and theoretical works have further developed this area. Correctly modelling document score distributions is important for many IR tasks. For example, if the distribution of relevant scores could be accurately inferred from the entire distribution of document scores, it would be particularly useful for automatic query-performance prediction (QPP) and/or meta-search (fusion) tasks [Baumgarten 1999; Yom-Tov et al. 2005]. Regardless of the practical applications, correctly modelling the distribution of

Author's address: R. Cummins, Department of Computing and Information Systems, School of Computing and Mathematical Science, University of Greenwich, London, SE10 9LS; email: ron.cummins@gmail.com.

DOI:http://dx.doi.org/10.1145/2559170

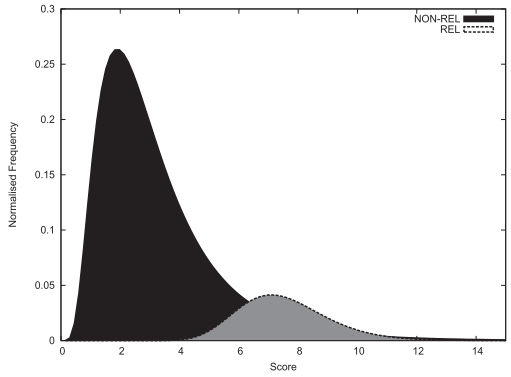

Fig. 1. A typical distribution of scores returned from a classical IR system.

relevant and nonrelevant documents remains an open, and theoretically important, area in IR. For illustration purposes, Figure 1 shows the document score distribution of a typical query on a standard IR system.

Using distributions to model the document scores provides a theoretically clean and practically useful approach for many IR tasks. Modelling an entire ranked list of document scores for a query using a mixture of distributions (i.e., one for the relevant document scores and one for the nonrelevant scores) allows for the compression of relevance information into a relatively small number of parameters. Over the last decade, the predominant distributions [Arampatzis and van Hameren 2001; Arampatzis et al. 2009a; Collins-Thompson et al. 2002] for modelling relevant and nonrelevant document scores have been normal and exponential, respectively. More recently, it has been suggested that the normal-exponential mixture has theoretical limitations [Robertson 2007], and in fact, a more theoretically valid approach is to model the scores using two gamma distributions [Arampatzis and Robertson 2011].

The first part of this article deals with determining the best distributions for use in an SD model by using the inference of average precision as a measure of *goodness*. We conduct a comprehensive empirical study of six SD models and show that a mixture of two log-normals is one of the best mixtures for modelling score distributions in terms of both *goodness-of-fit* and the inference of average precision for short queries. Furthermore, we present an empirical comparison of the mixtures with regard to the recall-fallout convexity hypothesis (RFCH). This analysis shows that the RFCH is a practically useful constraint in this area.

The second part of this article deals with the prediction of query performance using SD models when no relevance information is available. We present approaches that aim to directly predict the performance of a query using SD models. We show that the best method is comparable in performance to some current prediction methods. However, we provide experimental evidence which shows that the newly developed predictor has some novel features. We show that it is better normalised across different IR models and that it can be updated in a principled manner in light of labelled relevance data.

The remainder of the article is organised as follows: Section 2 reviews related work on modelling document score distributions and query performance prediction. Section 3 outlines the SD models used in this work and also presents the formulae used for inferring average precision from an SD model. Section 4 presents empirical results comparing the SD models using labelled data (i.e., relevance information) for a number of different metrics. In Section 5, we present an empirical comparison of

SD models with regard to the RFCH. In Section 6, we outline a heuristic approach to estimating the parameters of an SD model using unlabelled data (when relevance information is unknown). Section 7 presents comparative results for the newly developed QPP approach. We also present experiments that show the usefulness of the new predictor in different scenarios. Finally, Section 8 concludes with a discussion and an outline of future work.

2. RELATED RESEARCH

In this section, we review related work in SD models and outline relevant work in the area of QPP. Furthermore, we outline the contribution of this work.

2.1. Score Distributions

Early work into SD models investigated the use of normal distributions [Bookstein 1977; Swets 1963]. More recent work has shown that modelling relevant and nonrelevant document scores using a normal and an exponential distribution, respectively, fits the scores at the head of the ranked list (i.e., top-1000 documents) [Arampatzis and van Hameren 2001]. This SD model has become the predominant model used in the literature over recent years. We can see that if the distribution of scores in Figure 1 was truncated at a score of 4.0, for example, the higher score range (i.e., documents above 4.0) might appear to be comprised of an exponential distribution (for nonrelevant) and a normal distribution (for relevant).

Others have addressed more theoretical aspects of the underlying distributions and have developed hypotheses under which certain distributions can be theoretically rejected [Arampatzis and Robertson 2011; Robertson 2007]. Robertson developed a recall-fallout convexity hypothesis (RFCH) which states that the recall-fallout curve for good systems should be upper convex and has shown that if the probability ranking principle [Rijsbergen 1979] holds, then certain distributions should be rejected on theoretical grounds. Essentially, the convexity hypothesis postulates that the receiver operating characteristic (ROC) curve of an IR system with an infinite number of documents does not intersect the line of no discrimination (i.e., is strictly upper convex). In other words, for an infinite collection, as one encounters documents in order of decreasing score, the fraction of relevant documents encountered should always be greater than the fraction of nonrelevant documents encountered. It is worth remembering that it is only the SD model that adheres to the RFCH, and a ranking drawn from that model could by chance be nonrepresentative of the model.

The RFCH implies the probability ranking principle and can be seen as extending it to the continuous domain. Interestingly, the hypothesis is consistent with, and may help explain, certain phenomena. It successfully explains the reason that precision-based effectiveness measures (e.g., P@10) tend to increase as collection size increases [Hawking and Robertson 2003; Madigan et al. 2006]. Madigan et al. [2006] show, via simulation, that for a number of mixture distributions under certain conditions (i.e., the conditions that ensure that the RFCH is adhered to), precision-based effectiveness measures increase as the sample of documents increases.

Further contributions [Arampatzis and Robertson 2011; Arampatzis et al. 2009b] have developed a 'strong SD hypothesis' that suggests that each separate distribution in the SD model should be able to approach the Dirac delta function (i.e., it must be able to approach an impulse under which the entire mass of documents can reside) at different scores s on the score line. It is also proposed that under extreme circumstances, the distributions of both relevant and nonrelevant document scores should separate [Arampatzis and Robertson 2011]. The conclusion of much of this research has been that a mixture of two gamma distributions currently constitutes the most theoretically sound SD model.

Some of the theoretical problems associated with the normal-exponential SD model were addressed recently [Arampatzis et al. 2009a] using truncated forms of distributions. Some novel approaches [Dai et al. 2011; Kanoulas et al. 2009] to modelling the score distribution have used multiple normal distributions for the relevant documents and a gamma distribution for the nonrelevant ones. Important work in analysing the generation process (i.e., ranking functions) of document scores and their resultant distributions has also been conducted [Kanoulas et al. 2010]. Indeed, it is shown in that work that the score distribution should look positively-skewed due to constituent parts of the generating process. Some researchers [Baumgarten 1999; Manmatha et al. 2001; Wilkins et al. 2010] have used SD models in practical data fusion approaches. More recently, an extended expectation-maximisation approach has also been developed [Dai et al. 2012] that deals with the problem of combining document scores returned from different IR systems in response to the same query. They show that improved inference of document relevance can be achieved by combining multiple IR models using SD models.

2.2. Query Performance Prediction

Query performance prediction (QPP) aims to automatically estimate the performance of a query [Hauff et al. 2010a; He and Ounis 2006; Yom-Tov et al. 2005] when relevance judgments are unknown. The performance of these 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 [Hauff and Azzopardi 2009]. One of the main motivations for this area of research is that if good estimates of query performance are available to an IR system, the system can apply different query augmentation strategies in dealing with these different types of queries. Many recent approaches to query reformulation [Balasubramanian et al. 2010; Dang et al. 2010] use QPPs to estimate if an initial query can be perturbed effectively.

One of the earliest QPP approaches has been that of the clarity score [Cronen-Townsend et al. 2002], which measures the KL-divergence between the query and collection model in a language modelling framework. There have been many improved versions of this [Cronen-Townsend et al. 2006; Hauff et al. 2008b]. Some early research in the area [Yom-Tov et al. 2005] developed a learning approach for the task and showed that it could be used in a number of application areas. Much research [Hauff et al. 2008a; He and Ounis 2006] has been carried out into using pre-retrieval predictors to predict query performance. Others [Tomlinson 2004] have shown that the score of the highest ranked document is positively correlated with query performance. Performance predictors based on the robustness of a ranked list returned from a query have been developed [Zhou and Croft 2006]. The same authors developed predictors based on weighted entropy [Zhou and Croft 2007]. The WIG (weighted information gain) predictor, developed in that work, effectively measures the difference between the score of the K top-ranked documents and the average document returned in response to a query. Others [Lang et al. 2008] have shown that the performance of a ranked list is correlated to the ability of the top-ranked documents to cover all aspects of a query. Some relevant work [Vinay et al. 2008] investigates different document score normalisation techniques and aims to estimate query performance based on the these normalised scores.

Research has also shown that the standard deviation (σ) of scores in a ranked list is a good predictor of query performance [Cummins et al. 2011; Pérez-Iglesias and Araujo 2010; Shtok et al. 2009]. Some of these approaches [Cummins 2012a; Shtok et al. 2009] outline useful arguments as to why the deviation in the head of a ranked list is a

good estimator of performance. The latter of these works uses Monte-Carlo simulations using score distributions to investigate two hypotheses. Others [Diaz 2007] have used ideas based on the clustering hypothesis and the similarity of document scores to develop system-performance predictors. Recent work using statistical decision theory [Shtok et al. 2010] has led to some significant improvements to the state-of-the-art of QPP. This work constructs a general framework in which a number of QPPs are used to estimate the quality of a pseudo-perfect reference ranking. A given ranking is then compared to the reference ranking using a measure of similarity to produce effectiveness estimates for the given ranking. Even more recently [Kurland et al. 2011], a general framework has been constructed with the aim of unifying a number of QPPs developed from apparently different frameworks.

2.3. Contributions

This work has a number of contributions. First, we conduct an extensive evaluation over numerous IR systems of several SD models for the task of inferring average precision. An extensive empirical study that compares several binary mixture SD models has not been conducted. We find that the best SD model for short queries is a two-lognormal model. However, for longer queries, the two-gamma model more accurately infers average precision. We show that the best method for estimating parameters for this task is the method of moments (MME), rather than maximum likelihood (MLE). We present experiments that show that adhering to the RFCH is useful as it reduces the number of parameters in the SD model but does not reduce the models' ability to infer average precision accurately. We then apply the two-lognormal SD model to the QPP task. We develop heuristic approaches to estimate the parameters of the SD model without labelled data. Finally, we show that the best approach developed is comparable, in terms of performance, to many currently used predictors. However, we show that the new approach developed has useful normalisation properties that allow the predictor to be compared across different IR systems. A further novel feature is that it is easily updated when partial relevance information is known.

3. SD MODELS

In this section, we introduce a number of assumptions that underpin this work. We then present several SD models that are used in this work to model the scores of relevant and nonrelevant documents.

3.1. Assumptions and Restrictions

Typically, a user with an information need (IN) in mind submits a query Q to an information retrieval (IR) system M. The system, which has an index of N documents, scores each document D according to some scoring, or ranking, function $S(Q, D)$. The system then returns all documents in decreasing order of $S(Q, D)$. We assume a binary view of relevance, and as such, the SD models only consider relevant and nonrelevant documents. Furthermore, we only consider IR systems that rank documents independently of each other, in accordance with the probability ranking principle (PRP) [Rijsbergen 1979].[1] While many current Web-based IR systems use the linked structure of the Web to derive importance scores for each document, these are not considered in this work.

[1]This does not hold for all IR systems (e.g., for those that rerank the top documents to promote diversity), but it is a widely held ranking principle.

Table I. Distributions Considered

Distribution	# of parameters	MME	MLE
Normal	2	$\hat{\mu} = m$ $\hat{\sigma^2} = v$	See MME
Log-Normal	2	$\hat{\mu} = ln(m) - 0.5 \cdot (1 + v/m^2)$ $\hat{\sigma^2} = ln(1 + v/m^2)$	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} ln(x_i)$ $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(ln(x_i) - \hat{\mu})$
Gamma	2	$\hat{k} = m^2/v$ $\hat{\theta} = v/m$	See footnote 2
Poisson	1	$\hat{k} = m$	See MME
Exponential	1	$\hat{\beta} = m$	See MME

3.2. Distributions

Following a review of the literature, we comprise SD models of the distributions listed in Table I, where m and v are the sample mean and sample variances of a sample $x_{i..n}$.

$$N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

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

$$G(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k \Gamma(k)}, \tag{3}$$

$$P(x;k) = \frac{k^x \cdot e^{-k}}{\Gamma(x+1)}, \tag{4}$$

$$E(x;\beta) = \frac{1}{\beta} e^{-x/\beta}. \tag{5}$$

Equations (1) to (5) [Evans et al. 2001] are the probability density functions (pdf) of the normal, log-normal, gamma[2], continuous Poisson, and exponential distributions, respectively, where Γ is the gamma function. With regard to the Poisson distribution, it has been shown that a combination of two Poisson distributions unconditionally adheres to Robertson's recall-fallout convexity hypothesis [Bookstein 1977; Robertson 2007]. However, due to the discrete nature of the Poisson distribution, it is unclear how it can be used to model document scores, which can be assigned any positive real number by the IR systems used in this work. Therefore, we use the continuous Poisson distribution [Marsaglia 1986], where a simplification of the pdf, which has been used in the literature [Kim et al. 2010], is as in Eq. (4).

3.3. Mixtures

We model both sets of documents, relevant and nonrelevant, using the same type of distribution (except for the normal-exponential model), where $f(s|1)$ is the pdf for the scores (s) of relevant documents and $f(s|0)$ is the pdf for the scores of nonrelevant documents, the general binary SD model is as follows:

$$f(s) = (\lambda) \cdot f(s|1) + (1 - \lambda) \cdot f(s|0), \tag{6}$$

where λ is the proportion of relevant documents drawn. λ can be viewed as the prior probability of relevance for each document prior to being ranked. In this work, no form

[2]The MLEs can be found by solving $ln(\hat{k}) + \psi(\hat{k}) = ln(\frac{1}{n}\sum_{i=1}^{n} x_i) + (\frac{1}{n}\sum_{i=1}^{n} ln(x_i))$ using Newton's method and subsequently solving $\hat{\theta} = 1/(kn)\sum_{i=1}^{n} x_i$.

Table II. Composition of SD Models and Adherence to RFCH and Strong SD Hypothesis

Label	Relevant	Non-Relevant	# of parameters	RFCH	Strong SD
N_1E_0	Normal	Exponential	4	No	Yes
N_1N_0	Normal	Normal	5	When $\sigma_1 = \sigma_0$	Yes
L_1L_0	Log-Normal	Log-Normal	5	When $\sigma_1 = \sigma_0$	Yes
G_1G_0	Gamma	Gamma	5	When $k_1 = k_0$ or $\theta_1 = \theta_0$	Yes
P_1P_0	Poisson	Poisson	3	Yes	No
E_1E_0	Exponential	Exponential	3	Yes	No

of document score normalisation is performed or needed for the upper limit for any of the distributions. Negative values are not supported by the log-normal or gamma distributions, and the IR models considered in this work only return positive scores. Table II outlines the mixtures and the parameters that need to be estimated for each SD model.[3] Two of the SD models (P_1P_0 and E_1E_0) only contain three parameters, one contains four parameters (N_1E_0), while three SD models contain five parameters (N_1N_0, L_1L0, and G_1G_0). Therefore, some models have more flexibility in terms of their ability to model scores from different systems. We have included the normal-exponential (N_1E_0) model, as it has been used in many studies to model score distributions for various tasks. For the N_1E_0, N_1N_0, P_1P_0, and E_1E_0 mixtures, the MME and MLE estimates are equivalent. However, for the L_1L_0 and G_1G_0 mixtures, the MME and MLE estimates will lead to different parameter settings.

3.4. Inferring Average Precision

Average precision is an informative measure used widely in the field of IR. Average precision can be viewed geometrically, as the area under the precision-recall curve [Aslam and Yilmaz 2005], and it conveys a broad view of the effectiveness of a query. It is also a stable measure [Buckley and Voorhees 2000] and is probably the most prevalent metric of both query and system performance used in the literature. Further discussions on the many useful properties of average precision are discussed in detail in some recent research [Robertson et al. 2010]. 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:

$$recall(s) = \int_s^\infty \frac{\lambda \cdot f(s|1) \cdot ds}{\lambda} = \int_s^\infty f(s|1) \cdot ds. \tag{7}$$

In a similar manner, the precision at s (i.e., the proportion of relevant returned documents over the number of returned documents) can be defined as follows:

$$precision(s) = \frac{\int_s^\infty \lambda \cdot f(s|1) \cdot ds}{\int_s^\infty f(s) \cdot ds} \tag{8}$$

Now that we can estimate the precision and recall at any score s in the range $[0 : \infty]$, we can infer a precision-recall curve and, therefore, estimate the average precision ($AuPR()$) of a query as follows:

$$AuPR() = \int_0^1 precision(s) \cdot dr_s, \tag{9}$$

[3]For the parameters of each model, we use the subscript 1 to imply that the parameter is used to model relevant document scores, whereas we use the subscript 0 to imply that the parameter is used to model nonrelevant document scores.

where $r_s = recall(s)$. This formulation is an intuitive way of calculating average precision using the score distributions. Other approaches to estimating average precision from a score distribution have been used in the literature. Expected average precision could be calculated from the probabilities of document relevance (which could be derived from the SD model using the original observed samples) [Aslam and Yilmaz 2006]. However, our approach estimates an SD model from the sample and in turn estimates average precision directly from the model. This method has been shown to be a more effective estimate of average precision [Cummins 2012b] than the aforementioned approach.

3.5. Parameter Estimation

Now that the basic model has been outlined, we turn to the problem of parameter estimation. In our experiments, we only use documents that have been 'returned' from an IR system to estimate the parameters of the model outlined in Equation (6). We define the *returned set* (RET) as the set of documents that match at least one query term (all other documents are considered not returned). We will now outline the arguments for this and how it affects the inference of average precision. We also outline an adjustment to $AuPR()$ that deals with the issue.

First, as many systems assign a score of zero to documents that do not match any query term, many documents will be assigned a score of zero.[4] If a binary mixture model was used to model the entire set of documents, the *goodness-of-fit* of such a model would be poor. An obvious solution would be to introduce two further distributions to model the relevant and nonrelevant documents in the unreturned set of documents. For many IR models, these extra distributions would simply model a spike of document mass at some low score. Theoretically, these extra distributions could be modelled using two Dirac delta functions and some further mixing parameters. Dirac's delta function is a distribution that is zero everywhere except at zero. Excluding these unranked documents from a binary model is a simplification. However, we will see that the effects of this simplification on the inference of average precision can be overcome with a small modification to $AuPR()$.

Second, one of the main reasons for modelling document scores using a mixture distribution is to be able to infer the parameters of the model using an unsupervised (or semisupervised) learning method with unlabelled data. Without labels, the documents that are not returned (i.e., that contain no query terms) provide little information to an unsupervised learning approach. The information provided by these extra distributions is very limited (as the all the scores tend to be similar). In fact, the only useful information that these distribution would provide are their mass (which is trivial to calculate).

The estimate used for λ in the experiments in Section 4 is $\frac{|REL \cap RET|}{|RET|}$ (i.e., a maximum likelihood estimate that uses the number of relevant documents $|REL \cap RET|$ in the entire returned set). This is in fact the probability of relevance for a document given that it has been retrieved ($p(REL|RET)$). This estimate will tend to be biased, as $p(REL|RET) > p(REL)$ for any good system (where $p(REL)$ is the probability of relevance of a randomly selected document in the collection). This in turn will affect the inference of average precision from the model. For short queries, the number of actual relevant documents that are not returned (and are therefore not included in the inference of recall from the model) can be quite high. This will tend to lead to an

[4]It is noted that not all retrieval models assign the same score to documents that do not get returned, but many of those considered in this article do. In practice, the ranking functions generated from IR models tend not to score documents that do not contain any query terms, regardless of the probabilistic smoothing that is used in some language models.

over-estimation of actual average precision. In general, for longer queries, our estimate of λ will tend to be closer to the true prior probability of relevance. Regardless, $AuPR()$ can be adjusted as follows to account for the overestimate of recall that is brought about by our estimate of λ:

$$AuPR()' = \phi \cdot AuPR() + (1 - \phi) \cdot \varepsilon, \tag{10}$$

where $\phi = \frac{|REL \cap RET|}{|REL|}$ is the recall of the returned set and ε reflects the average precision in the set of unreturned documents (we assume that $\varepsilon = 0$ for our experiments). If we were to model the unreturned set of documents using two Dirac delta functions as mentioned earlier, this adjustment of $AuRP()$ would arise naturally from the more complex four-distribution model. Although the adjustment of $AuPR()$ increases the number of parameters to be estimated, for simplicity, we treat this adjustment as separate from the SD models.

Furthermore, when an IR system returns a ranked list (RET) for a query, and when relevance judgements are known, the mean (m) and variance (v) of the relevant and nonrelevant document scores in the returned set can be calculated using this labelled data. Therefore, by using the MME (or MLE) equations from Table I, we can estimate the parameters in each SD model. Each parameterised model is specific to each ranked list, and therefore, it models the returned set of document scores from a system given a query. For the experiments in the next section, we estimate the parameters of the SD models using relevance (labelled) data and subsequently compare the performance of the models using different measures of goodness.

4. EMPIRICAL STUDY OF MIXTURE PERFORMANCE

In this section, we perform a comparative analysis of the six SD models using labelled data across a number of different IR systems (e.g., vector space, classic probabilistic, language model, and axiomatic model).

4.1. Measures of Goodness

For different fields of study and problems, different measures may be applicable. Herein, we conduct a comparative analysis using the following measures of goodness.

— We use the correlation between the average precision inferred by an SD model and the actual average precision over a set of queries as a measure of the usability of the model in an IR setting.
— The root mean squared error (RMSE) of the average precision inferred by an SD model compared to the actual average precision, is used to measure the interpretability of the output of a particular SD model as an actual average precision value.
— We use the Kolmogorov-Smirnov's D-statistic to measure the relative goodness of fit of SD models over the entire returned set.

Usually, goodness-of-fit tests are used to either accept or reject certain models as a 'good fit'. It is well known in IR that documents, and therefore document scores, at the head of a ranked list are more important than those further down the list. An intuitive way of measuring the usability of a specific SD model is by trying to infer the average precision of a query using the model, and its parameters as estimated from labelled data. Therefore, over a set of queries, the correlation between the inferred average precision from the SD model and the actual average precision of the query from the IR system affords us a measure of the usefulness of different SD models.

Table III. Test Collection Details

Collection		# docs	# topics	topic range	Query Length	
					title	desc
Test	AP	242,918	149	051–200	3.6	10.1
	FT	210,158	188	251–450	2.5	7.6
	WT2G	221,066	50	401–450	2.3	6.3
	WT10G	1,692,096	100	451–550	2.6	6.7

We compare the six SD models introduced earlier in Table II over a range of IR systems. Different distributions may be better at modelling different IR systems, and so we compared the SD models across 11 IR systems. We chose the vector space model using TFIDF [Salton and Buckley 1988] and pivoted document-length normalisation (with three parameter settings of $s = 0.01$, $s = 0.05$, and $s = 0.2$) [Singhal et al. 1996], the probablistic model BM25 (with three parameter settings of $k_1 = 1.2$ and $b = 0.25$, $b = 0.5$, and $b = 0.75$) [Robertson et al. 1994], divergence-from-randomness model ($I(n)L2$ with three parameter settings of $c = 1$, $c = 2$, and $c = 5$) [Amati and Van Rijsbergen 2002], a language modelling (LM) approach (Jelinek-Mercer smoothing with the smoothing parameter set to 0.2) [Zhai and Lafferty 2004], and the axiomatic approach (F2EXP) [Fang and Zhai 2005], as these represent a broad range of classical and more modern ranking functions. Table III shows the test collections[5] used in this research.

4.2. Linear Correlation

We measure the correlation of the inferred average precision (estimated from Equation (9)) for an SD model using MME with the actual average precision, over a set of queries. We report this correlation for the 11 systems on a number of test collections. Figure 2 shows the linear correlation of the six SD models for all of the 11 systems when using the entire returned set as a sample. First, it is worth noting that the correlation coefficients are strong (some above 0.8), indicating that much of the information regarding average precision can be accurately modelled by some of the SD models.[6] Experiments using only the top 1,000 documents as a sample (not shown) produce higher absolute correlations, but reflect the same relative ranking of the SD models.

For short queries, the best-performing model across all systems is the two-lognormal model. For longer description queries, the best-performing model is the two-gamma model. Rather surprisingly, we can see that the normal-exponential model is generally the worst-performing model for inferring average precision on short queries. It should be noted that in this work, we do not normalise document scores, unlike other works that use the normal-exponential model, nor do we truncate the ranked list at 1,000 scores. Although this may hurt the performance of the normal-exponential model, score normalisation and the truncation of the ranked list lead to theoretical inconsistencies [Robertson 2007] (i.e., both affect the parameters estimated from the data). The two-poisson model performs well on some collections despite having fewer parameters than other models. All models on the larger Web collection (WT10G) show reduced performance in general, indicating that it is a harder collection from which to infer performance using SD models. This may be due to the more noisy nature of Web documents. Interestingly, the QPP task has also been shown to be more difficult on large-scale Web corpora [Hauff et al. 2008b].

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

[6] Using Kendall's tau correlation produces lower absolute correlation coefficients, while largely maintaining the relative ranking of the SD models.

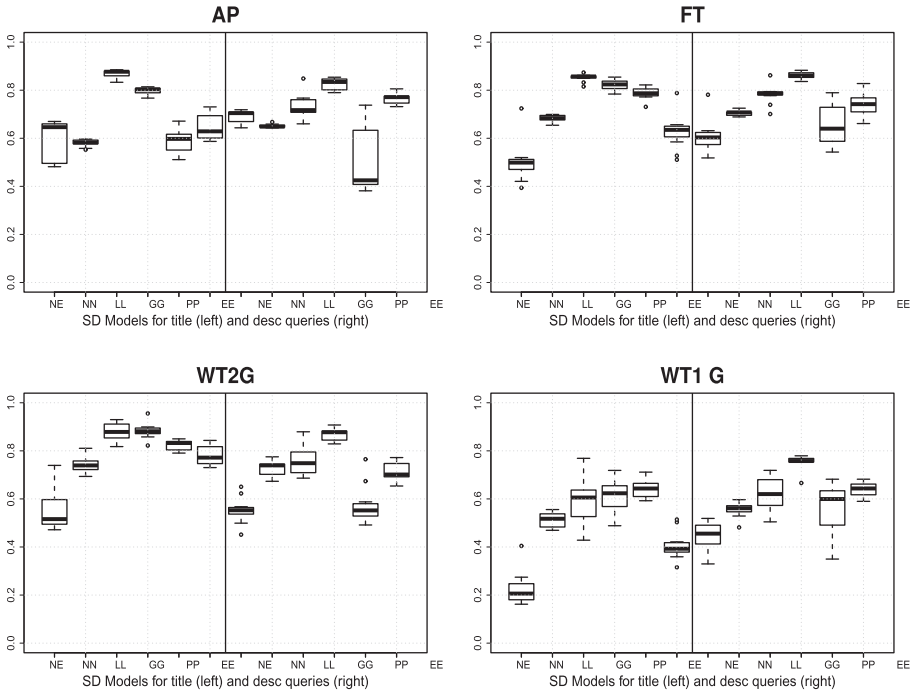

Fig. 2. Linear correlation of SD models inferred average precision and real average precision for AP and FT Newswire collections (top) and for WT2G and WT10G Web collections (bottom) on 11 systems.

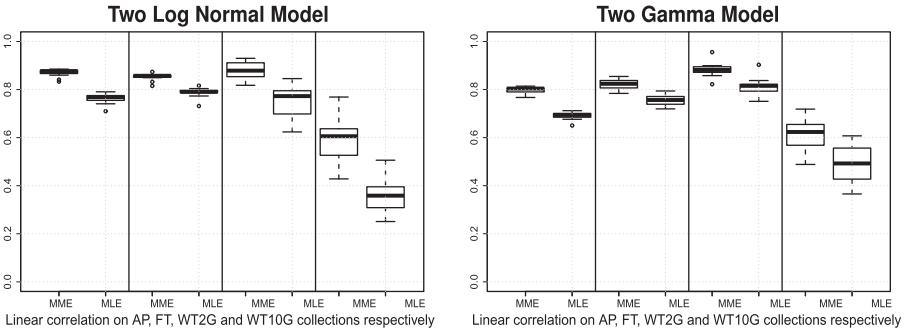

Fig. 3. Comparison of MME and MLE estimation techniques for a two-lognormal model (left) and a two-gamma model (right) using title queries.

4.3. MME v MLE

Figure 3 shows the performance of two different parameter estimation techniques (for two SD models) for the task of inferring average precision on short queries. The MME approach to parameter estimation consistently outperforms the MLE approach for the task of inferring average precision as measured by a linear correlation. The results for longer description queries (not shown) show a similar trend, although the difference in performance is not as pronounced. This result is of importance for applications that might wish to use SD models for various IR tasks. Interestingly, we determined that the location parameters estimated from the MME approach are consistently higher than those estimated by MLE, especially for the relevant distributions. The MMEs

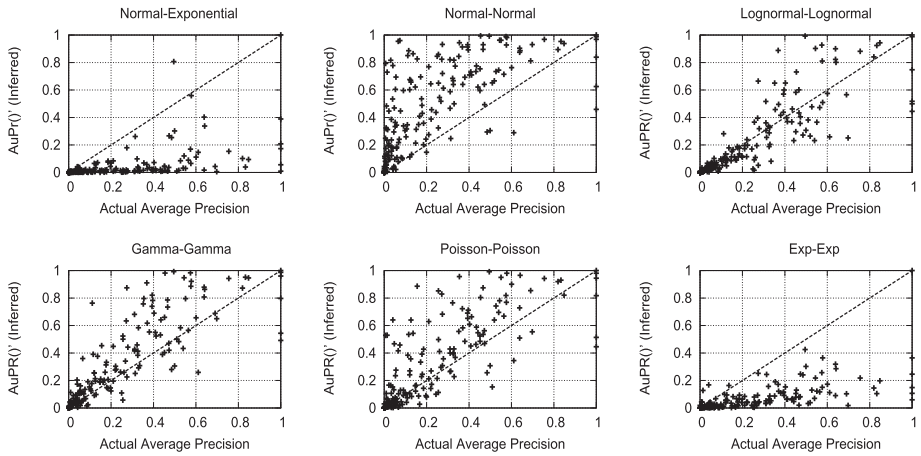

Fig. 4. Average precision versus inferred average precision for default BM25 function on FT Newswire collections for Topics 251–450 using title queries.

tend to be biased toward the higher-ranked scores, as these scores tend to be outliers. This is due to the positive skew of the score distribution created by typical IR systems. Although MLEs do tend to be better estimates for parameters that model the entire data, much of the data are irrelevant (i.e., documents below rank 1,000) for the task of inferring average precision. Therefore, this would seem to suggest that the MMEs are better for modelling the outliers at the head of a ranked list, and also suggests why the inference of average precision is superior for these estimates.

4.4. RMSE

Figure 4 shows the actual average precision versus the inferred average precision for all six SD models for a typical system (BM25) on the FT collection. We can see for two of the models (normal-exponential and two-exponential) that the inferred average precision is underestimated. For the two-normal model (and two-Poisson model), we can see that average precision is overestimated. For both the two-lognormal and two-gamma model, the inferred average precision is closer to the actual average precision. These results are typical across the systems and collections tested.

Figure 5 shows the average RMSE of the inferred average precision compared to the actual average precision for a set of queries on all of the systems. We can see that the RMSE of the log-normal model is lower on all collections and queries sizes. The normal-exponential model outperforms the two-normal model on all collections. In fact, the worst-performing model is the two-normal model. This is because it severely overestimates average precision, as seen in Figure 4. Underestimating average precision will tend to lead to a lower RMSE due to the fact that the mean average precision of a set of queries tend to be between 0.2 and 0.3 on many collections. While the RMSE is not a normalised measure of the mutual information between two variables, it does inform us that the raw average precision value inferred by the log-normal model is closer to the actual average precision of a query, and therefore, that the output of the model is more interpretable as an actual average precision value.

4.5. Goodness-of-Fit

The Kolmogorov-Smirnoff D-statistic measures the maximum distance between the cumulative density function of the theoretical distribution (i.e., one of our SD models) and the empirical distribution (i.e., the actual scores). We use the measure as a

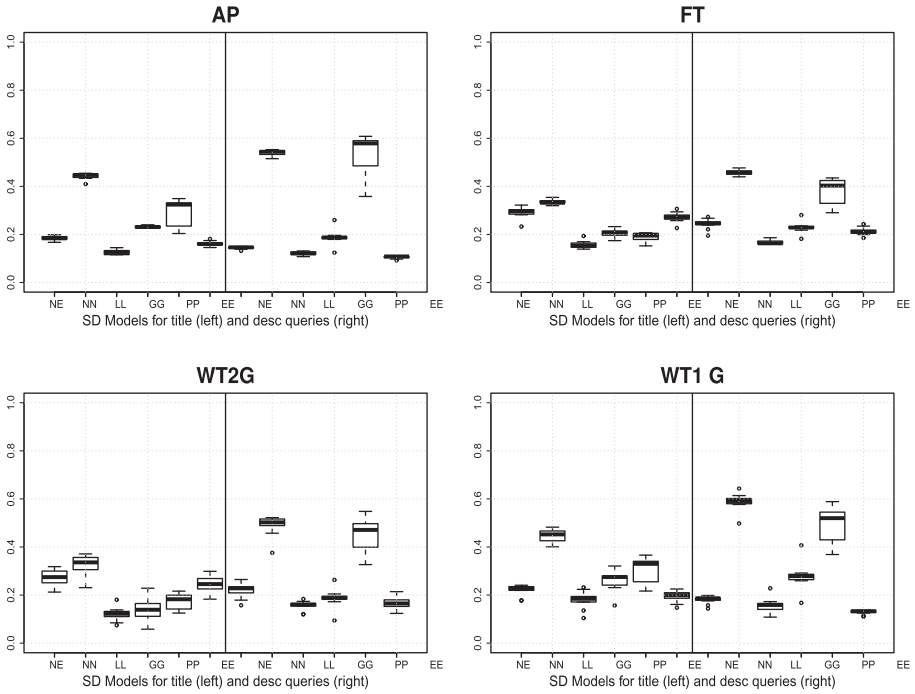

Fig. 5. Average RMSE between SD models' inferred average precision and real average precision on a set of topics for FT and AP Newswire collections (top) and for WT2G and WT10G Web collections (bottom).

relative measure of goodness-of-fit.[7] Figure 6 shows that the two-lognormal model tends to have a better fit compared to the two-gamma model across systems and query lengths. The gamma model has a better fit for some IR systems for short queries on Web collections. The three models with the best fit (i.e., two-normal, two-lognormal, and two-gamma models) are those with the largest number of parameters. They are also the only SD models that can possibly adhere to both the RFCH and the 'strong SD' hypothesis. The two-lognormal and two-gamma models are the best-fitting models when using only the top 1,000 documents of a ranking (not shown), with the former being the best fit. The actual D-statistics of all SD models are higher (indicating a worse fit) when using only the top 1,000 samples (due to the smaller sample sizes).

5. USEFULNESS OF RFCH

The recall-fallout convexity hypothesis (RFCH) [Robertson 2007] has been proposed as a possibly useful constraint for valid SD models. Assuming document scores can be modelled using SD models, this hypothesis constrains the parameters of certain models of score distributions. Therefore, in this section, we aim to investigate the practical usefulness of the RFCH.[8] 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. In this section, we again use

[7]As the parameters of the model are estimated from the observed samples, the critical values of the Kolmogorov-Smirnoff test are invalid. However, we use the D-statistic as a relative measure to compare the mixtures, and not as a statistical test to accept or reject the validity of the distribution given the data.
[8]Much of this section is updated from a recent short paper but is included, as it leads to simplified models that are used in the latter part of the paper [Cummins and O'Riordan 2012].

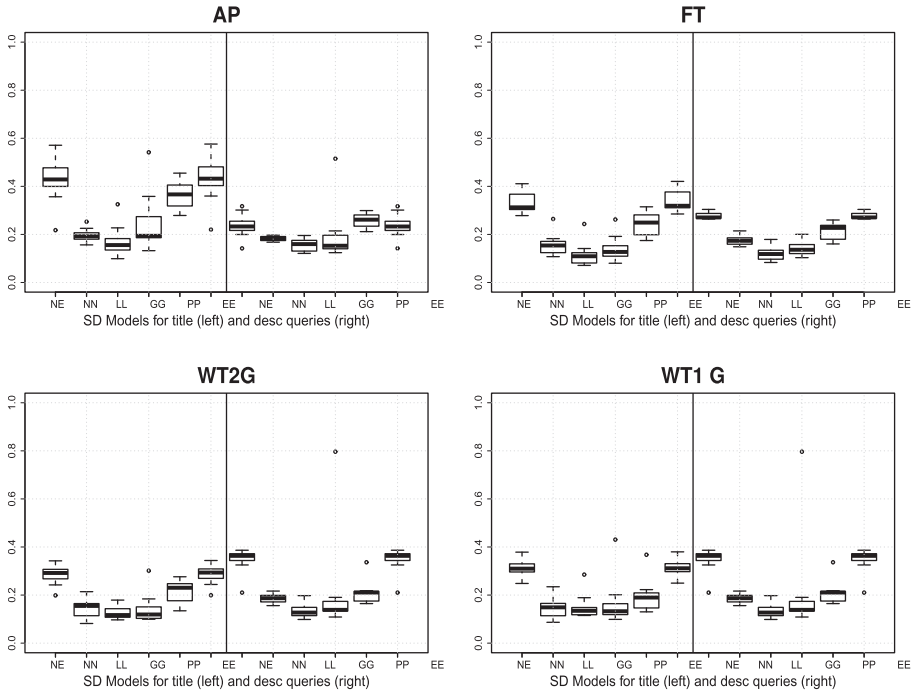

Fig. 6. Average D-statistic of SD model's on a set queries for all systems for FT and AP Newswire collections (top) and WT2G and WT10G Web collections (bottom).

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 nonrelevant 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, it is more 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 section, 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 nonrelevant).[9] We modify the variance of the relevant document scores, as it has fewer degrees of freedom than the mean and therefore will inherently be the less accurate estimator.

The *two-normal (N_1N_0) model* has been shown to adhere to the RFCH only when the variances of both relevant and nonrelevant distributions are equal (i.e., $\sigma_1 = \sigma_0$) [Robertson 2007]. These variance parameters are very rarely equal when the variances are estimated from the sample variances. Therefore, to force this mixture to conform to the RFCH, the sample mean for both relevant and nonrelevant documents (m_1 and m_0) are used as the mean of both distributions, respectively, (μ_1 and μ_0), and the sample variance of the nonrelevant documents (v_0) is used as both variance parameters ($\sigma_1^2 = \sigma_0^2$).

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

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 ($\theta_0 = \theta_1$)[10] [Robertson 2007]. 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 can be modified to $v_1 = v_0 \cdot m_1/m_0$ before the method-of-moments estimates are calculated to ensure that $\theta_1 = \theta_0$. This constrains the two-gamma SD model and implies that $\frac{E[s_1]}{E[s_0]} = \frac{Var(s_1)}{Var(s_0)}$ for this model.

The *two-lognormal (L_1L_0) model* adheres to the RFCH when the variance parameters for both distributions are equal ($\sigma_0 = \sigma_1$) [Cummins et al. 2011]. 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 can be assumed to be $v_1 = v_0 \cdot m_1^2/m_0^2$ before the method-of-moments estimates are calculated. This constrains the log-normal SD model and implies that $\frac{E[s_1]}{E[s_0]} = \frac{\sqrt{Var(s_1)}}{\sqrt{Var(s_0)}}$.

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 the modified four-parameter models have any disadvantages in terms of their ability to correctly model relevance information (as measured by the ability of a model to infer average precision).

5.1. Experiments

To test the practical usefulness of our modified SD models, 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 a linear and Kendall's τ correlation coefficient to measure how well the output of a particular model (i.e., inferred average precision) agrees with the actual average precision. The IR systems used are the same as in the previous section. We only report results for both short title queries and longer description queries.

5.2. Results and Discussion

Figure 7 shows box plots of Kendall's τ correlation on the 11 systems for the three SD models that are not forced to adhere to the RFCH, and the modified version of the SD models (labelled '_v') that are forced to adhere to the RFCH on four collections. We can see that the results indicate that adhering to the RFCH is beneficial, as the ability of the modified SD models to infer average precision does not decrease. For some of the collections, the performance increases when using the less complex model. For all the models, in general, there is no loss in performance when the model is forced to adhere to the RFCH. These results are consistent when using a linear correlation as the measure of performance (not shown). This is an interesting outcome, as each modified SD model is less complex than its five-parameter counterpart. Overall, the four-parameter two-lognormal model is the best performing model for short queries on all collections.[11] Again, the two-gamma model is better for longer queries.

[10]We also conducted experiments that ensured that $k_1 = k_0$ and determined that setting both scale parameters ($\theta_1 = \theta_0$) to be equal was more beneficial for the inference of average precision.

[11]If comparing Figure 7 to Figure 2, it is worth remembering that the earlier figure used Pearson's correlation.

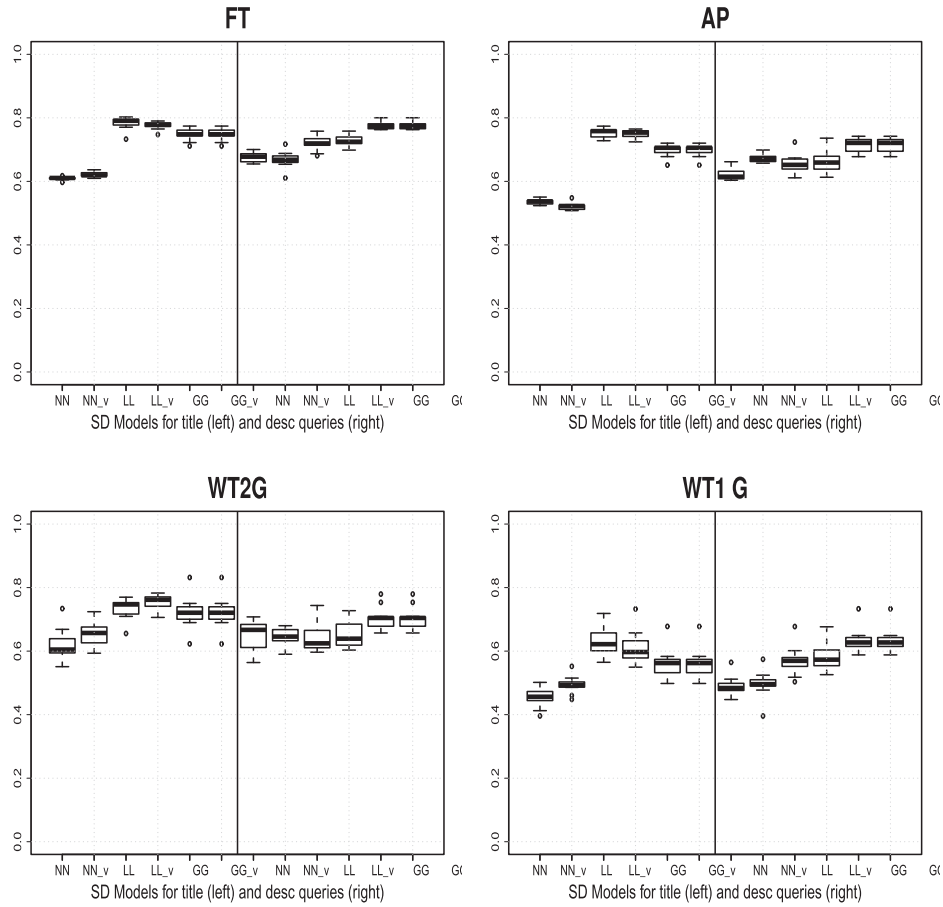

Fig. 7. Kendall's τ correlations for mixtures that violate the RFCH and those that adhere to the RFCH (labelled with '_v') for title queries on two Newswire and two Web collections.

5.3. Summary

We have presented an extensive empirical analysis of several SD models present in the literature for the task of inferring average precision. We have shown that an SD model consisting of two log-normal distributions outperforms many of the others in its ability to infer average precision for short queries across a number of different IR systems. The two-gamma model is a better choice for longer queries for this task. Furthermore, we have shown that MME outperforms MLE as the method for estimating parameters for the task of inferring average precision. Finally, an analysis using the RFCH has shown that it is a useful hypothesis, as it reduces the number of parameters in the SD model while maintaining its effectiveness of inferring average precision.

6. ESTIMATING PARAMETERS WITHOUT RELEVANCE INFORMATION

From the previous experiments, we have shown that the two-lognormal SD model is practically useful for modelling the scores returned from a variety of IR systems. Therefore, for the remainder of the article, we restrict ourselves to using the two-lognormal SD model. In this section, we develop approaches to automatically estimate (i.e., when no relevant information is known) the parameters of the SD model using

a number of different methods from unlabelled data. Consequently, we turn our attention to developing a principled query-performance prediction (QPP) approach using SD models. Given that we have modelled the scores in a theoretically principled manner, it would seem that approaches such as expectation-maximisation (EM) would provide a useful and clean solution to the problem of parameter estimation when dealing with unlabelled data. However, it has been shown that the standard EM algorithm is very sensitive to its initial starting parameters and performs poorly for this task [Arampatzis et al. 2009b; Cummins 2011] (i.e., inferring document relevance from scores). A possible reason for this is the class imbalance between the relevant documents and the nonrelevant documents. Given that there are so few relevant documents for many queries, it is extremely difficult to estimate the parameters for the relevant distribution given the noise from the nonrelevant document scores. It should be noted that an extended EM approach that uses multiple rankings does provide better performance [Dai et al. 2012]. However, without access to rankings provided from other systems, the problem of parameter estimation from unlabelled data remains extremely difficult for SD models. We do not attempt a theoretically principled and consistent approach to parameter estimation in this article. Rather, we will outline some heuristic methods to estimate the parameters of the SD model. We then outline the benefits of this approach in an experimental setting.

6.1. Estimation of Parameters

In order to infer average precision ($AuPR()'$), or any other metric of performance, we need to estimate the mixing parameter (λ) and the following three parameters of the two-lognormal model: $\{\sigma_0, \mu_0, \text{and } \mu_1\}$. Although there are extra parameters to estimate for $AuPR'()$ (i.e., ϕ), it can be seen that this can be compacted and treated as one parameter $\{\phi \cdot \lambda\}$ which is still in the range $[0 : 1]$. As we are aiming to estimate these parameters from unlabelled data, we simplify the model by ignoring the adjustment to $AuPR()$ and excluding the estimation of ϕ.

As we have previously described our parameters using moments, we aim to estimate the following three moments: $\{m_1, m_0, v_0\}$. These can be used, as in the previous section, to estimate the parameters of the SD model. It was noted that the RFCH implies $\frac{E[s_1]}{E[s_0]} = \frac{\sqrt{Var(s_1)}}{\sqrt{Var(s_0)}}$ for the two-lognormal model. Interestingly, we can see that this implies that the standard deviation of the relevant document scores is proportional to the mean of the relevant document scores. Furthermore, this suggests that methods [Cummins et al. 2011; Pérez-Iglesias and Araujo 2010; Shtok et al. 2009] that aim to estimate the performance of a query using measures of dispersion on the top-ranked documents are ultimately related to SD models. Regardless, if we assume that most of the documents returned for a query are nonrelevant, we can estimate m_0 and v_0 using all of the document scores that are returned for a query. It has been shown [Cummins 2011] that this heuristic is experimentally useful and that, not unsurprisingly, the parameters that are most difficult to estimate accurately are m_1 and λ.

6.1.1. Estimating the Mean of Relevant Scores. Previous research has shown that the standard deviation of document scores at the head of ranking is correlated to query performance. Given that the RFCH implies that the standard deviation of relevant scores is correlated to the mean of the relevant document scores for the log-normal model, we use one of these measures of dispersion to estimate this mean. The standard deviation of document scores from s_{max} down to a score of $s_{max}/2$ has been shown to be a good predictor of performance [Cummins et al. 2011]. We adopt this predictor as a basis by which we can estimate the mean score of relevant documents ($\hat{m_1}$). Furthermore, as the standard deviation is expressed in the same units as the data, the maximum value

of this standard deviation is $s_{max}/4$. As we know that the mean score of the sample m_1 must be less than s_{max}, we use the following adjustment of the standard deviation for our estimation of m_1:

$$\hat{m_1} = 4 \cdot stdev(s_{max}, s_{max}/2), \tag{11}$$

where $stdev(s_{max}, s_{max}/2)$ is the standard deviation of the ranked-list from s_{max} to $s_{max}/2$ and will vary from 0 to $s_{max}/4$. It is noted that this is a crude estimate based on a previous approach to QPP. However, experiments will show that the SD model approach to QPP has some novel features.

6.1.2. Estimating the Mixing Parameter. The mixing parameter is the only remaining parameter that needs to be estimated. This parameter controls the proportion of relevant documents that are drawn from the SD model compared to the total number of documents drawn. It can be viewed as the prior probability of relevance and does not depend on the scores produced by a specific IR system. Therefore, we aim to estimate this using pre-retrieval predictors similarly to Kurland et al. [2012]. It is expected that queries that are very specific would have a higher λ than very broad general queries. Query-specificity has been studied in many works [Arampatzis and Kamps 2010] and is known to be related to the performance of a query. Therefore, we use a number of different estimators of the mixing parameter λ as follows:

$$\hat{\lambda}_1 = \frac{10}{RET}, \tag{12}$$

where we simply use a fixed constant as the number of relevant documents. We also use the average idf of the terms in the query as follows:

$$\hat{\lambda}_2 = idf_{avg} \cdot \frac{10}{RET}, \tag{13}$$

where idf_{avg} is the average idf value of terms in the query. The idf value of each term is calculated as $idf = log(N/df)$, where df is the document frequency of a term and N is the number of documents in the collection. The similarity between the collection and query, as measured by the scq_{avg}, also measures the informativeness of the entire query and is adopted as follows:

$$\hat{\lambda}_3 = scq_{avg} \cdot \frac{10}{RET}, \tag{14}$$

where scq_{avg} [Zhao et al. 2008] is the average of the query term-weights w_t calculated for each term as $w_t = 1 + log(cf) \cdot log(1 + N/df)$, where cf is the frequency of a term in the entire collection. If during the estimation of λ, the estimate is assigned a value greater than 1.0, we simply assign it a value of 1.0 (and therefore the average precision inferred from such a model would also be 1.0).

7. COMPARATIVE RESULTS

In this section, we compare the various approaches outlined in the previous section. We conduct an extensive evaluation using many systems, collections, and QPP baselines.

7.1. Systems and Baselines

For broader comparison of predictors, we ran our experiments on four different IR systems. We used the default pivoted document-length normalisation [Singhal et al. 1996] ($s = 0.2$) ranking function, the default BM25 [Robertson et al. 1994] ranking function ($k_1 = 1.2$ and $b = 0.75$), a language model [Zhai and Lafferty 2004] with Jelinek-Mercer smoothing (set to 0.2), and an axiomatic term-weighting approach (F2EXP with

$s = 0.5$) [Fang and Zhai 2005] for our experiments.[12] It is important that results are not simply limited to one system, as we will see that predictors vary quite considerably on different systems. We only used title type queries for the experiments that follow.

We employed a variety of baselines against which to test our newly developed approach. As a weak baseline, we used the pre-retrieval predictor idf_{avg}, which is the average of the *idf* values of terms in the query [Arampatzis and Kamps 2010]. We used the standard deviation of the top-K documents (DEV) [Pérez-Iglesias and Araujo 2010] and query drift (NQC) [Shtok et al. 2009] at K documents. We used the normalised standard deviation of documents within a threshold (50%) of the top score (NDEV) [Cummins et al. 2011] and the weighted information-gain predictor (WIG). We also used a previous version of a predictor that uses score distributions MMP1 [Cummins 2011]. For the DEV and NQC approaches, we tuned K from 50–300 in increments of 25 on the test collections for each retrieval function. Similarly for WIG, we tuned K from 1–40 in increments of 4. We chose the best K averaged across all the collections for each of these three approaches. We can confirm that the typical recommended settings of 100 for DEV and NQC, and 5 for WIG perform within 85% of the most optimal settings for the approaches.

All the baselines mentioned thusfar are computationally inexpensive once the ranked list of scores is retrieved for a query. As a further baseline, we used query feedback (QF), a computationally more expensive[13] approach. We used the top-20 terms from the initial retrieval as feedback to the IR system [Zhou and Croft 2007]. The fraction of documents retrieved in the top-K documents that are common to both the initial query and the feedback query is used as the predictor. We set $K = \{20, 50, 100\}$ and determined that $K = 20$ was the best setting in general across the collections used in this work.

We report Kendall-τ, and a linear correlation where suitable, for the experiments that follow. Kendall's-τ is less effected by outliers, making it a more robust measure. It should be noted that Spearman's correlation tends to produce coefficients that are 50% greater that those of Kendall-τ [Fredricks and Nelsen 2007]. We have also observed this relationship during our experiments. Although we do not report Spearman's coefficient in this work, it is important to note when comparing the results in this work against others in the literature.

7.2. Experimental Setup

We conducted three experiments to demonstrate the usefulness of the developed approach. First, in Section 7.3, we compare the approaches developed in this article with the baselines over a number of different collections and systems by measuring the correlation between the average precision produced from the system and the output of the predictor. This is the most common evaluation approach used in the literature. In the second experiment (Section 7.4), we show the usefulness of our predictor in a meta-search scenario. In a meta-search scenario, a system may route different queries to different IR systems. Therefore, a meta-search engine may return rankings from different IR systems for different queries (i.e., it is not assumed that the same underlying function is generating the scores for each instance of a query). In the final experiment (Section 7.5), we show that our approach can incorporate relevance information in a principled manner. Although relevance information is not usually available for use with QPP approaches, we demonstrate how it can be incorporated if available.

[12]It should be noted that these systems only output positive document scores.
[13]It requires two retrieval runs and a term-selection stage.

Table IV. Kendall-τ Correlation of SD Predictors with Average Precision

		title		
M	Collection	SD_{λ_1}	SD_{λ_2}	SD_{λ_3}
PIV	AP	0.374	0.389	**0.425**
	FT	0.355	0.367	**0.375**
	WT2G	0.375	0.399	**0.433**
	WT10G	0.252	0.310	**0.320**
BM25	AP	0.416	0.413	**0.447**
	FT	0.321	0.337	**0.372**
	WT2G	0.280	0.295	**0.299**
	WT10G	0.186	0.259	**0.295**
LM	AP	0.318	0.333	**0.396**
	FT	0.357	**0.374**	0.362
	WT2G	0.327	0.330	**0.333**
	WT10G	0.237	**0.309**	0.294
F2EXP	AP	0.216	0.255	**0.300**
	FT	0.257	0.289	**0.320**
	WT2G	0.285	0.296	**0.301**
	WT10G	0.183	0.254	**0.271**

Table V. Linear Correlation of Predictor Output with Average Precision on Four IR Systems Using Title Queries on Two Newswire and Two Web Collections

		Linear								
M	Collection	SD_limit	idf_{avg}	QF	MMP1	DEV	NDEV	NQC	WIG	SD_{λ_3}
PIV	AP	0.88	0.371	0.453	0.489	0.486	**0.718**	0.387	0.643	0.539
	FT	0.86	0.407	0.356	0.477	0.557	0.505	**0.583**	0.360	0.539
	WT2G	0.92	0.548	0.458	0.253	0.577	**0.698**	0.486	0.649	0.545
	WT10G	0.66	0.195	0.254	0.311	0.403	0.452	0.260	**0.478**	0.371
BM25	AP	0.89	0.378	0.357	0.523	0.437	**0.702**	0.333	0.618	0.599
	FT	0.86	0.415	0.402	0.495	0.524	0.497	**0.570**	0.360	0.537
	WT2G	0.86	0.476	0.325	0.207	0.403	**0.588**	0.324	0.558	0.465
	WT10G	0.56	0.193	0.276	0.163	0.326	0.377	0.137	**0.479**	0.261
LM	AP	0.85	0.378	0.362	0.382	0.299	**0.636**	0.280	0.523	0.601
	FT	0.84	0.416	0.358	0.435	0.521	0.381	**0.570**	0.184	0.554
	WT2G	0.90	0.529	0.443	0.211	0.429	**0.642**	0.469	0.521	0.491
	WT10G	0.76	0.223	0.200	0.247	0.368	0.298	0.289	**0.434**	0.294
F2EXP	AP	0.88	0.354	0.347	0.214	0.255	**0.565**	0.098	0.442	0.573
	FT	0.83	0.401	0.282	0.367	0.480	0.462	0.402	0.367	**0.476**
	WT2G	0.92	0.411	0.356	0.317	0.144	0.243	0.252	**0.460**	0.423
	WT10G	0.60	0.189	**0.291**	0.208	0.060 $\downarrow$	0.181	0.051 $\downarrow$	0.235	0.263

7.3. Comparative Results

Table IV shows the performance of the predictors based on SD models. The best predictor is marked in bold, and all correlations are significant. On average, we can see that the predictor that uses scq_{avg} as the mixing parameter (i.e., SDλ_3) is more highly correlated (Kendall-τ) with average precision than the other approaches. We use this approach to compare to the baselines.

Table V and Table VI show the performance (both linear and Kendall-τ correlation) of SDλ_3 against the baselines outlined in Section 7.1. In Tables V and VI, most correlations are significant (those marked $\downarrow$ are not significant). The column labelled 'SD_limit' is the performance of the SD model using labelled data. It is the theoretical

Table VI. Kendall-Tau Correlation of Predictor Output with Average Precision on Four IR Systems Using Title Queries on Two Newswire and Two Web Collections

		Kendall-τ								
M	Collection	SD_limit	idf_{avg}	QF	MMP1	DEV	NDEV	NQC	WIG	SD_{λ_3}
PIV	AP	0.76	0.241	0.313	0.323	0.264	**0.510**	0.159	0.446	0.425
	FT	0.79	0.237	0.193	0.367	0.359	**0.380**	0.360	0.352	0.375
	WT2G	0.76	0.349	0.301	0.342	0.402	0.392	0.352	**0.446**	0.433
	WT10G	0.65	0.186	0.189	0.314	0.345	0.313	0.352	**0.356**	0.320
BM25	AP	0.76	0.247	0.269	0.366	0.219	**0.498**	0.122	0.427	0.447
	FT	0.80	0.234	0.202	**0.414**	0.360	0.371	0.361	0.334	0.372
	WT2G	0.71	0.254	0.265	0.320	0.306	**0.379**	0.266	0.351	0.299
	WT10G	0.60	0.190	0.181	0.211	0.289	0.321	0.267	**0.386**	0.295
LM	AP	0.74	0.245	0.269	0.255	0.121	**0.436**	0.086	0.353	0.396
	FT	0.76	0.240	0.190	0.309	0.317	0.286	0.338	0.252	**0.362**
	WT2G	0.75	0.287	0.312	0.319	0.324	**0.418**	0.297	0.348	0.333
	WT10G	0.71	0.228	0.256	**0.385**	0.291	0.326	0.259	0.339	0.294
F2EXP	AP	0.77	0.222	0.265	0.206	0.108	**0.373**	0.042	0.263	0.300
	FT	0.77	0.226	0.233	0.312	0.298	0.319	0.296	0.232	**0.320**
	WT2G	0.74	0.234	0.256	0.311	0.280	**0.333**	0.263	0.341	0.301
	WT10G	0.59	0.177	0.191	0.191	0.289	0.269	0.250	**0.313**	0.271

upper limit of the SD approach. On average, the best predictor is NDEV on these datasets. However, $SD\lambda_3$ is quite competitive and, on average, outperforms NQC, DEV, and QF. As the $SD\lambda_3$ uses the standard deviation to estimate m_1, it is not surprising that it should maintain a high performance. We also note that $SD\lambda_3$ is robust across collections and systems.

It is also worth noting that the performance is quite different for different IR systems. It seems that it is more difficult to estimate the performance of a retrieval using scores from the F2EXP retrieval function. It is also more difficult to estimate the performance of a retrieval on the WT10G Web collection. We noted previously that Web documents create an inherently more noisy environment. Although, we have not seen an improvement over some of the baselines for our new approach, in the next sections, we will outline the major advantages of the new predictor $SD\lambda_3$.

7.4. System Independence

Some meta-search systems route queries issued to them to different IR systems and return the resultant ranking to the user. As the scores returned from a meta-search system may belong to different unknown IR systems, the problem of QPP is somewhat more difficult. We simulate such a scenario by simply pooling the data output by the four IR systems outlined in Section 7.1. Therefore, for each collection, we pooled the predictor output/average precision pairs across the four IR systems. This problem can be tackled as a score normalisation task [Arampatzis and Kamps 2009], but has not been addressed in the QPP area. This experiment tests whether the outputs of the predictors are comparable across systems. Table VII outlines the results of this experiment for the best QPP approaches. We included the query-feedback (QF) approach because it is normalised between 0 and 1 and therefore should be somewhat comparable across systems. We can see that the new $SD\lambda_3$ predictor outperforms a selection of predictors. This is because the $SD\lambda_3$ predictor aims to estimate average precision directly and therefore is correctly normalised. To negate the possibility that one system was causing this effect, we performed the same experiment over all four combinations of three systems (i.e., {PIV, BM25, LM}, {PIV, BM25, F2EXP}, {PIV, LM, F2EXP}, and

Table VII. Kendall-τ Correlation of Output of Predictor vs. Average Precision Independent of the Systems

		title			
M	Collection	QF	NDEV	WIG	SD_{λ_3}
POOLED	AP	0.287	0.327	0.258	**0.384**
	FT	0.209	0.265	0.210	**0.333**
	WT2G	0.275	0.215	0.133	**0.385**
	WT10G	0.154	0.210	0.243	**0.281**

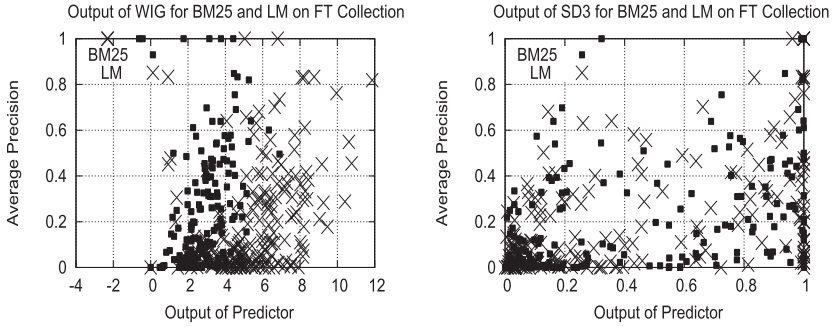

Fig. 8. Output of predictor vs. average precision for WIG and $SD\lambda_3$.

{BM25, LM, F2EXP}). The results from those experiments (excluded due to similarity) showed very similar results to those in Table VII. For each combination of three systems, SD_{λ_3} outperformed the other baselines.

Therefore, when given ranked lists that have been generated from any one of the four IR systems,[14] the $SD\lambda_3$ predictor produces a measure of performance which is independent of the system (i.e., an estimate of average precision). To illustrate this across system normalisation, Figure 8 shows the output of both WIG and $SD\lambda_3$ versus average precision for two systems (BM25 and LM) on the FT collection. We can see that the raw output of the WIG predictor is higher for the LM system than for the BM25 system which shows that across-system comparison is not possible. However, the output of the $SD\lambda_3$ predictor is normalised and therefore is more interpretable as a measure of performance, independent of the system which returned the list.

7.5. Updating with Relevance

While it is not usually the case that relevance information is known during the process of estimating query performance, the predictor developed herein provides a theoretically principled way of dealing with relevance information. Interestingly, Butman et al. [2013] also recently studied this problem in detail. When a labelled relevant document associated with a query Q is known to the system, the μ_1 parameter in the $SD\lambda_3$ predictor can be updated via updating the estimate of the sample mean score of relevant documents ($\hat{m_1}$) as follows when each new relevant document score is encountered:

$$\hat{m_1'} = \frac{S(Q, d_{R_i}) + \hat{m_1} \cdot (|R| - 1)}{|R|}, \tag{15}$$

[14]These arbitrary systems must still adhere to the assumptions outlined in Section 3.1.

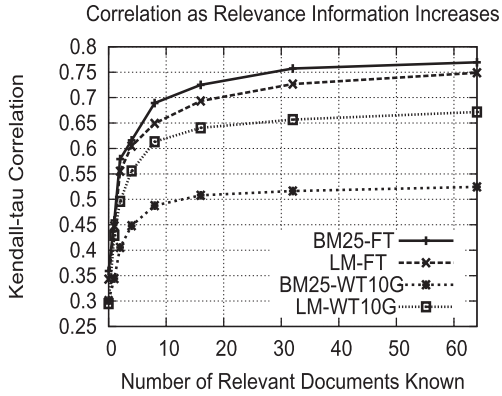

Fig. 9. Increase in performance on FT and WT10G as relevance information increases on BM25 and LM systems.

where $S(Q, d_{R_i})$ is the score of the known relevant document and $|R|$ is the number of known relevant documents. Consequently, we can update the mixing parameter as

$$\hat{\lambda} = \frac{|R|}{RET}. \tag{16}$$

The mean score of nonrelevant documents can be updated in a similar manner. These update formulas can be viewed as imposing weak priors on the parameters when no relevance information is available. We performed an experiment in which a varying number of randomly selected relevant documents were known to the system. We ran this randomised experiment ten times and averaged the results of all ten runs. Figure 9 shows the increase in the Kendall-τ correlation of the predictor and average precision as the number of relevant documents known to the system (i.e., $|R|$) increases. We can see that there are diminishing returns in terms of performance after approximately four documents are known to be relevant. It should be noted that previous research has investigated the minimum number of relevant documents needed in order to be able to infer average precision to a certain level of accuracy [Yilmaz and Aslam 2006]. The experiment outlined here shows the principled nature of the approach of the SD model outlined and show that modelling the QPP task in a more principled manner leads to novel features.

8. DISCUSSION AND CONCLUSION

We have conducted a study into SD models which determines the best composition to use when aiming to infer average precision. The two-lognormal model is the best performing for short queries for a number of performance metrics. By assuming Robertson's convexity hypothesis, we have demonstrated that the performance of several models can be maintained while reducing the number of parameters in the model. These experiments demonstrate the practical usefulness of the convexity hypothesis. Although there exists a myriad of distributions that can be evaluated for this task, we have only performed a comparison of six SD models. However, given that many retrieval functions (including all those in this article) are summations of term weights over the query [Lv and Zhai 2011], it can be seen that the resultant score distribution is a combination of the distributions of those term weights. In fact, aiming to

correctly model the individual term weights may be a promising avenue of future research.

Using the two-lognormal SD model, we have developed an approach for QPP and have evaluated the approach on two Newswire collections and two Web collections over four different IR systems. While the best predictor developed in this article does not outperform some of the best-performing QPP approaches, as commonly evaluated, it is comparable to many high-performing baselines. We have shown that one benefit of the approach is that it is comparable across IR systems that produce different scores (i.e., it has some normalising behaviour). We have not compared our approach against linear combinations of QPP approaches which have been shown to increase performance on certain collections [Kurland et al. 2011; Zhou and Croft 2007]. However, we feel that this is a useful contribution to the literature, as it creates a simplified model that can be built upon by more principled unsupervised learning approaches. In fact, more recent work [Kurland et al. 2012] outlines a probabilistic approach to QPP in which pre-retrieval predictors are seen as the prior probability of relevance. This is similar to the mixture parameter in our model which is the prior probability of relevance.[15]

Furthermore, there exist many other baselines against which we can compare our approach. For example, the improved query clarity predictor was not used in this work. While we did implement the original clarity measure [Cronen-Townsend et al. 2002], we did not include the results, as it performed poorly compared to WIG and DEV. It is also worth noting that the experiments in this article were performed on less noisy data than the newer larger Web collections. For example, the prediction of query performance for the ClueWeb collection, which is more representative of an online Web environment, has been shown to be more difficult for post-retrieval predictors [Hauff et al. 2010b]. Regardless, there are many application domains (such as digital libraries), where the document collection is inherently less noisy than the general Web.

Recently, a general framework has shown that many QPP approaches can be decomposed into two parts [Kurland et al. 2011]. The authors suggest that QPP approaches either measure the level of dissimilarity between the ranking produced by a query and poor-quality reference rankings, or they measure the similarity between the ranking produced by a query and good-quality reference rankings. The approach developed in this article also adheres to both of these intuitions. In fact, the separation of the distributions of relevant and nonrelevant documents in the SD model is a fundamentally similar concept. We have also shown that the SD predictor can be updated in a principled manner when relevance information is known. In a standard relevance feedback scenario, users might indicate relevance in the order in which they encounter documents. The update formulae presented herein are only valid when a random relevant document is known.

While the methods used in the latter part of this article for estimating parameters are heuristic, we have outlined a general model in which these parameters model meaningful aspects of a query. It is important to note that the estimation of these parameters is fundamental to the estimation of query performance (i.e., they are the same task). Future work aims to use principled machine learning approaches to more accurately estimate these parameters. This is not a trivial task due to a number of problems. First, it is important that the task is correctly modelled before attempting to use principled techniques like expectation maximisation (or other variants). Furthermore, there exists a large class imbalance when using the entire returned set of

[15]The mixture parameter in our model is estimated using the probability of relevance given it has been returned, as noted earlier.

document scores as samples. Future work will aim to modify certain machine learning techniques and use other sources of evidence to improve performance.

ACKNOWLEDGMENTS

The author would like to thank the reviewers for their insightful comments that greatly helped to improve the manuscript.

REFERENCES

G. Amati and C. J. Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst. 20*, 4, 357–389. ISSN 1046-8188.

A. Arampatzis and J. Kamps. 2009. A signal-to-noise approach to score normalization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. ACM, New York, NY, 797–806. ISBN 978-1-60558-512-3.

A. Arampatzis and J. Kamps. 2010. An empirical study of query specificity. In *Proceedings of the 32nd European Conference on Information Retrieval (ECIR)*. 594–597.

A. Arampatzis and S. Robertson. 2011. Modeling score distributions in information retrieval. *Inf. Retr. 14*, 1, 26–46.

A. Arampatzis and A. van Hameren. 2001. The score-distributional threshold optimization for adaptive binary classification tasks. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 285–293.

A. Arampatzis, J. Kamps, and S. Robertson. 2009a. Where to stop reading a ranked list?: Threshold optimization using truncated core distributions. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 524–531.

A. Arampatzis, S. Robertson, and J. Kamps. 2009b. Score distributions in information retrieval. In *Proceedings of the 2nd International Conference on the Theory of Information Retrieval (ICTIR'09)*. Lecture Notes in Computer Science, vol. 5766, Springer-Verlag, Berlin, 139–151. ISBN 978-3-642-04416-8.

J. A. Aslam and E. Yilmaz. 2005. A geometric interpretation and analysis of r-precision. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*. 664–671.

J. A. Aslam and E. Yilmaz. 2006. Inferring document relevance via average precision. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 601–602.

N. Balasubramanian, G. Kumaran, and V. R. Carvalho. 2010. Exploring reductions for long Web queries. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. 571–578. ISBN 978-1-4503-0153-4.

C. Baumgarten. 1999. A probabilistic solution to the selection and fusion problem in distributed information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*. ACM, New York, NY, 246–253. ISBN 1-58113-096-1.

A. Bookstein. 1977. When the most pertinent document should not be retrieved—An analysis of the Swets model. *Inf. Process. Manage. 13*, 6, 377–383.

C. Buckley and E. M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in International Retrieval (SIGIR)*. 33–40.

O. Butman, A. Shtok, O. Kurland, and D. Carmel. 2013. Query-performance prediction using minimal relevance feedback. In *Proceedings of the Conference on the Theory of Information Retrieval (ICTIR'13)*. ACM, New York, NY. ISBN 978-1-4503-2107-5.

K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. 2002. Information filtering, novelty detection, and named-page finding. In *Proceedings of the 11th Text Retrieval Conference*.

S. Cronen-Townsend, Y. Zhou, and W. B. Croft. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*. ACM, New York, NY, 299–306. ISBN 1-58113-561-0.

S. Cronen-Townsend, Y. Zhou, and W. B. Croft. 2006. Precision prediction based on ranked list coherence. *Inf. Retr. 9*, 6, 723–755.

R. Cummins. 2011. Predicting query performance directly from score distributions. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology (AIRS'11)*. Springer-Verlag, Berlin, 315–326. ISBN 978-3-642-25630-1.

R. Cummins. 2012a. Investigating performance predictors using Monte Carlo simulation and score distribution models. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. ACM, New York, NY, 1097–1098. ISBN 978-1-4503-1472-5.

R. Cummins. 2012b. On the inference of average precision from score distributions. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 2435–2438. ISBN 978-1-4503-1156-4.

R. Cummins and C. O'Riordan. 2012. On theoretically valid score distributions in information retrieval. In *Proceedings of the 34th European Conference on Advances in Information Retrieval (ECIR'12)*. Springer-Verlag, Berlin, 451–454. ISBN 978-3-642-28996-5.

R. Cummins, J. Jose, and C. O'Riordan. 2011. Improved query performance prediction using standard deviation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information (SIGIR'11)*. ACM, New York, NY, 1089–1090. ISBN 978-1-4503-0757-4.

K. Dai, E. Kanoulas, V. Pavlu, and J. A. Aslam. 2011. Variational bayes for modeling score distributions. *Inf. Retr. 14*, 1, 47–67.

K. Dai, V. Pavlu, E. Kanoulas, and J. A. Aslam. 2012. Extended expectation maximization for inferring score distributions. In *Proceedings of the 34th European Conference on Advances in Information Retrieval (ECIR'12)*. Springer-Verlag, Berlin, 293–304. ISBN 978-3-642-28996-5.

V. Dang, M. Bendersky, and W. B. Croft. 2010. Learning to rank query reformulations. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 807–808. ISBN 978-1-4503-0153-4.

F. Diaz. 2007. Performance prediction using spatial autocorrelation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. ACM, New York, NY, 583–590. ISBN 978-1-59593-597-7.

M. Evans, N. Hastings, and B. Peacock. 2001. Statistical distributions, third edition. *Measure. Sci. Technol. 12*, 1, 117.

H. Fang and C. Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval (SIGIR)*. 480–487.

G. A. Fredricks and R. B. Nelsen. 2007. On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. *J. Stat. Plan. Inference 137*, 7, 2143–2150.

C. Hauff and L. Azzopardi. 2009. When is query performance prediction effective? In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 829–830.

C. Hauff, D. Hiemstra, and F. de Jong. 2008a. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*. 1419–1420.

C. Hauff, V. Murdock, and R. Baeza-Yates. 2008b. Improved query difficulty prediction for the Web. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, New York, NY, 439–448. ISBN 978-1-59593-991-3.

C. Hauff, L. Azzopardi, D. Hiemstra, and F. de Jong. 2010a. Query performance prediction: Evaluation contrasted with effectiveness. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR)*. 204–216.

C. Hauff, D. Kelly, and L. Azzopardi. 2010b. A comparison of user and system query performance predictions. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, New York, NY, 979–988. ISBN 978-1-4503-0099-5.

D. Hawking and S. E. Robertson. 2003. On collection size and retrieval effectiveness. *Inf. Retr. 6*, 1, 99–105.

B. He and I. Ounis. 2006. Query performance prediction. *Inf. Syst. 31*, 7, 585–594.

E. Kanoulas, V. Pavlu, K. Dai, and J. A. Aslam. 2009. Modeling the score distributions of relevant and nonrelevant documents. In *Proceedings of the 2nd International Conference on the Theory of Information Retrieval (ICTIR)*. Lecture Notes in Computer Science, vol. 5766, Springer-Verlag, Berlin, 152–163.

E. Kanoulas, K. Dai, V. Pavlu, and J. A. Aslam. 2010. Score distribution models: Assumptions, intuition, and robustness to score manipulation. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research Development in Information Retrieval (SIGIR)*. 242–249.

T. Kim, A. V. Nefian, and M. J. Broxton. 2010. Photometric recovery of Apollo metric imagery with Lunar-Lambertian reflectance. *Electron. Lett. 46*, 9, 63–633.

O. Kurland, A. Shtok, D. Carmel, and S. Hummel. 2011. A unified framework for post-retrieval query-performance prediction. In *Proceedings of the 3rd International Conference on the Theory of Information Retrieval (ICTIR)*. Lecture Notes in Computer Science, vol. 6931, Springer-Verlag, Berlin, 15–26.

O. Kurland, A. Shtok, S. Hummel, F. Raiber, D. Carmel, and O. Rom. 2012. Back to the roots: A probabilistic framework for query-performance prediction. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, 823–832. ISBN 978-1-4503-1156-4.

H. Lang, B. Wang, G. Jones, J.-T. Li, F. Ding, and Y.-X. Liu. 2008. Query performance prediction for information retrieval based on covering topic score. *J. Comput. Sci. Technol. 23*, 4, 590–601. ISSN 1000-9000.

Y. Lv and C. Zhai. 2011. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*. ACM, New York, NY, 7–16. ISBN 978-1-4503-0717-8.

D. Madigan, Y. Vardi, and I. Weissman. 2006. Extreme value theory applied to document retrieval from large collections. *Inf. Retr. 9*, 3, 273–294. ISSN 1386-4564.

R. Manmatha, T. Rath, and F. Feng. 2001. Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*. ACM, New York, NY, 267–275. ISBN 1-58113-331-6.

G. Marsaglia. 1986. The incomplete [gamma] function as a continuous poisson distribution. *Comput. Math. Appl. 12*, 5–6, 1187–1190. ISSN 0898-1221.

J. Pérez-Iglesias and L. Araujo. 2010. Standard deviation as a query hardness estimator. In *Proceedings of the 17th International Conference on String Processing and Information Retrieval (SPIRE)*. 207–212.

C. J. V. Rijsbergen. 1979. *Information Retrieval* 2nd Ed. Butterworth-Heinemann, Newton, MA. ISBN 0408709294.

S. Robertson. 2007. On score distributions and relevance. In *Proceedings of the 29th European Conference on Information Retrieval Research (ECIR'07)*. Springer-Verlag, Berlin, 40–51. ISBN 978-3-540-71494-1.

S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at trec-3. In *Proceedings of the 3rd Text REtrieval Conference (TREC'94)*. 109–126.

S. E. Robertson, E. Kanoulas, and E. Yilmaz. 2010. Extending average precision to graded relevance judgments. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 603–610. ISBN 978-1-4503-0153-4.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage. 24*, 5, 513–523.

A. Shtok, O. Kurland, and D. Carmel. 2009. Predicting query performance by query-drift estimation. In *Proceedings of the 2nd International Conference on the Theory of Information Retrieval (ICTIR)*. Lecture Notes in Computer Science, vol. 5766, Springer-Verlag, Berlin, 305–312.

A. Shtok, O. Kurland, and D. Carmel. 2010. Using statistical decision theory and relevance models for query-performance prediction. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 259–266.

A. Singhal, C. Buckley, and M. Mitra. 1996. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*. ACM, New York, NY, 21–29. ISBN 0-89791-792-8.

J. A. Swets. 1963. Information retrieval systems. *Science 141*, 3577, 245–250.

S. Tomlinson. 2004. Robust, Web and terabyte retrieval with Hummingbird Searchserver at TREC 2004. In *Proceedings of the 13th Text Retrieval Conference (TREC)*.

V. Vinay, N. Milic-Frayling, and I. Cox. 2008. Estimating retrieval effectiveness using rank distributions. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, New York, NY, 1425–1426. ISBN 978-1-59593-991-3.

P. Wilkins, A. F. Smeaton, and P. Ferguson. 2010. Properties of optimally weighted data fusion in CBMIR. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. 643–650. ISBN 978-1-4503-0153-4.

E. Yilmaz and J. A. Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'06)*. ACM, New York, NY, 102–111. ISBN 1-59593-433-2.

E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. 2005. Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*. ACM, New York, NY, 512–519. ISBN 1-59593-034-5.

C. Zhai and J. Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst. 22*, 2, 179–214. ISSN 1046-8188.

Y. Zhao, F. Scholer, and Y. Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proceedings of the 30th European Conference on Information Retrieval*

Research (ECIR'08). Lecture Notes in Computer Science, vol. 4956, Springer-Verlag, Berlin, 52–64. ISBN 3-540-78645-7, 978-3-540-78645-0.

Y. Zhou and W. B. Croft. 2006. Ranking robustness: A novel framework to predict query performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. ACM, New York, NY, 567–574. ISBN 1-59593-433-2.

Y. Zhou and W. B. Croft. 2007. Query performance prediction in Web search environments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. ACM, New York, NY, 543–550. ISBN 978-1-59593-597-7.

Received October 2012; revised April 2013, August 2013; accepted September 2013