On the Inference of Average Precision from Score Distributions

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

ABSTRACT

Modelling the document scores returned from an IR system for a given query using parameterised score distributions is an area of research that has become more popular in recent years. Score distribution (SD) models are useful for a number of IR tasks. These include data fusion, query performance prediction, determining thresholds in filtering applications, and tasks in the area of distributed retrieval. The inference of performance metrics, such as average precision, from these SD models is an important consideration. In this paper, we study the accuracy of a number of methods of inferring average precision from an SD model.

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

General Terms: Experimentation, Measurement, Performance

Keywords: Information Retrieval, Score Distributions, Inference

1. INTRODUCTION

Modelling the document scores returned from information retrieval (IR) systems using score distributions (SD) models [16] is both theoretically principled and practically useful [12]. A number of works [1, 6, 9] have attempted using expectation-maximisation to infer the distributions of relevant and non-relevant document scores from unlabelled data. However, there are a number of unanswered questions with regard to SD models even when labelled data (i.e. relevance judgments) are available. Before dealing with 'noisy' and unlabelled data, it is important that we know how to correctly model and accurately infer performance metrics from 'clean' labelled data. This paper deals with determining the best method of inferring average precision from SD models that use labelled data and are theoretically consistent with many known IR principles.

2. BACKGROUND AND RELATED WORK

In this section, we discuss related work regarding SD models and we outline the contribution made by this work.

2.1 Hypotheses

A number of recent works [13, 11, 2, 7] in this area have aided in constraining the search for models that are consistent with various theories and observed phenomena in the domain of IR. Robertson's Recall-Fallout Convexity Hypothesis [13] generalises previous work [4] and extends the probability ranking principle (PRP) to the continuous score domain where one ranking of the collection, given a query, is a sample ranking drawn from an infinite collection which is assumed to adhere to the PRP. This effectively invalidates the normal-exponential SD model used in many works. Others [2] have postulated both 'strong' and 'weak' SD hypotheses that help to constrain the choice of distributions that comprise an SD model. These hypotheses suggest that the SD model should be able to support perfect retrieval (i.e. full separation of relevant and non-relevant distributions).

2.2 Score Normalisation

Normalisation of retrieval scores prior to parameter estimation can affect the theoretical validity of an SD model. For example, min-max (0-1) normalisation used in many previous works [1, 9] effectively prohibits scores below 0 and above 1. In such circumstances, any distribution that supports values outside of that range (i.e. 0-1) would be invalid, as it would assign a non-zero probability to an observation that is unobservable (i.e. impossible) [13].

2.3 Composition

As it is the query that generates both relevant and non-relevant document scores, it would be unlikely that the pair of distributions that comprise a binary SD model are from two different families of distributions (e.g. it is unclear why relevant documents should be drawn from a normal, when non-relevant documents are drawn from an exponential distribution). In most cases, an IR system does not know the underlying relevance of each document and can only strive to separate the two distributions as much as possible (as implied by the 'strong' and 'weak' SD hypothesis). Therefore, it is likely that both distributions should be of the same family (type) of distribution (at least for an initial ad hoc retrieval run where no relevance information is available).

2.4 Truncation

Using only a portion of the ranked-list (i.e. the top 1000

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$15.00.

Table 1: Adherence of SD models to a number of SD Hypotheses

	RFCH	Strong SD	Weak SD	Score Range Supported	
				Rel	Non-Rel
Lognormal-Lognormal	when $\sigma_1 = \sigma_0$	yes	yes	$[0 : \infty]$	$[0 : \infty]$
Gamma-Gamma	when $\theta_1 = \theta_0$ or $k_1 = k_0$	yes	yes	$[0 : \infty]$	$[0 : \infty]$
Poisson-Poisson	yes	no	yes	$[0 : \infty]$	$[0 : \infty]$
Normal-Normal	when $\sigma_1 = \sigma_0$	yes	yes	$[-\infty : \infty]$	$[-\infty : \infty]$
Exp-Exp	yes	no	yes	$[0 : \infty]$	$[0 : \infty]$
Exp-Normal	no	no	yes	$[0 : \infty]$	$[-\infty : \infty]$

document scores) from which to infer the SD model parameters may negatively affect the performance of a model. Truncating the list at arbitrary points may ignore useful information regarding the document scores of a large number of documents. Although most of the documents that are discarded below the truncation point may not be relevant, they are important for the inference of the non-relevant document score distribution. This can, in turn, affect the inference of performance metrics. Conversely, modelling the entire collection of documents as a ranked-list leads to problems where the actual 'fit' of the model is poor, due to the number of documents that receive no score (due to not matching any query-term). This would manifest itself as a large probability mass (spike) at a score of zero. In fact, for many models of retrieval, documents that do not match a query-term are excluded from the ranking (i.e. they are not ranked), rather than receiving a score of zero[1].

2.5 Contribution

Table 1 lists a number of SD models used in the literature and outlines the hypotheses that they each adhere to, when assuming that the expected score of relevant documents is greater than the expected score of non-relevant documents [13]. Inferring average precision from these distributions is an important consideration in determining the practical usefulness of the model and the practical usefulness of any task using these models. However, average precision has been inferred using two different methods thus far in the literature [8, 9]. In this paper, we determine empirically which of these methods is more accurate. In doing so we use SD models that are consistent with all of the aforementioned hypotheses. Furthermore, we do not apply score normalisation or ranked-list truncation.

3. SCORE DISTRIBUTIONS

In this work we model a document ranking using an SD model that adheres to the recall-fallout convexity hypothesis (RFCH) [13], where $f(s|1)$ and $f(s|0)$ are the probability density function of the relevant and non-relevant document scores respectively, and where λ is the mixing parameter. Therefore, the scores returned in a ranked list can be modelled as $f(s) = \lambda \cdot f(s|1) + (1 - \lambda) \cdot f(s|0)$. Furthermore, we use two gamma distributions[2] to model the scores of both relevant and non-relevant documents.

Given a ranked list of all document scores $s_1, s_2, s_3, ...s_{|ret|}$ returned in response to a query Q (i.e. where ret is the returned set of documents that match at least one query-term) and the known binary relevance labels for the documents at each of those scores, we can estimate the SD model parameters for that ranking. For simplicity, we use method-of-moment estimates and ensure that the model adheres to the RFCH by equating the scale parameters of both distributions using the values obtained from the non-relevant document scores following previous research [8]. We estimate λ using only the relevance labels in the returned set (i.e. $\lambda = |rel \cap ret|/|ret|$ and where rel is the set of relevant documents). It is important to note that we do not perform any score normalisation on the output of the scores from any of the IR systems used. We now review two methods that can be used to infer average precision from an SD model and the available ranking.

3.1 Expected Average Precision

Research using the Maximum Entropy Method to analyse performance metrics [3] has shown that the expected average precision **E[ap]** of a ranking can be calculated as follows:

$$E[ap] = \frac{1}{R} \cdot \sum_{i=1}^{N} (\frac{p_i}{i} \cdot (1 + \sum_{j=1}^{i-1} p_j)) \qquad (1)$$

where R is the number of relevant documents for a query, N is the number of documents in the ranking, and p_i is the probability that a document at rank i is relevant. The probability of relevance at rank i can be inferred from the SD model as $p_i = \frac{\lambda \cdot f(s_i|1)}{f(s_i)}$ following Bayes' rule used in previous research [9]. Adherence of the SD model to the RFCH ensures that this probability of relevance decreases as i increases, and therefore, adheres to the PRP. Given that documents that do not match a query-term are not modeled in our approach, $N = ret$ is the number of returned documents and therefore, R can be estimated as $\sum_{i=1}^{|ret|} p_i$. This method is used in recent research [9] and can be computed in $O(N)$ time.

3.2 Area Under the PR Curve

Average precision corresponds to the area under the precision recall curve [15]. This can be expressed in terms of a score s as follows:

$$AuPR = \int_0^1 prec(s) \cdot drec(s) \qquad (2)$$

where $prec(s)$ is the precision at score s on the score line and $rec(s)$ is the recall at score s [5]. $rec(s)$ is calculated as

[1] For models of retrieval that allow negative scores, assigning a score of zero to documents that match no query-terms is not a practical, or theoretically sound, solution. Deeming the documents not-returned would seem a practical and more theoretically sound approach.

[2] A two-lognormal model that adheres to the RFCH yields higher actual correlation values, but comparatively similar results as regards the accuracy of the two approaches to inferring average precision studied in this paper.

$\int_s^\infty f(s|1) \cdot ds$, while $prec(s)$ is calculated as $\frac{\int_s^\infty \lambda \cdot f(s|1) \cdot ds}{\int_s^\infty f(s) \cdot ds}$. As s is supported on $[0 : \infty)$, we can approximate this using numerical integration. In particular, N uniformly spaced integration points on $[0 : B]$ can be used where B is some upper bound where $f(s) \approx 0$. We have found by experimentation that $2 \cdot s_1$ (i.e. two times the top score) is a suitable point[3] for the collections used here. **AuPR** can be calculated in $O(N)$ time where N is the number of samples used. By starting at B, we can estimate recall, fallout, and precision at each integration point down to 0. The estimate of **AuPR** can be calculated cumulatively as the algorithm progresses (as outlined in Algorithm 1). These samples can be viewed as documents that are uniformly dispersed on the score line s from 0 to B.

Algorithm 1 Calculate AuPR with N samples and mixing parameter λ

```
score = $2 \cdot s_1$
recall = 0
fallout = 0
ap = 0
ds = $score/N$
for i=0 to N do
  score = score - ds
  recall += rel_likelihood(score) · ds
  fallout += non_rel_likelihood(score) · ds
  prec[i] = ($\lambda$· recall) / ($\lambda$· recall + (1-$\lambda$) · fallout)
  rec[i] = recall
  if $i > 0$ then
    ap + = (rec[i]-rec[i-1]) · (prec[i]+prec[i-1])/2
  end if
end for
return ap
```

4. EXPERIMENTS

In the experiments that follow, we compare both methods of inferring average precision (i.e. **E[ap]** and **AuPR**) using a large number of queries on three test collections from TREC disks 1-5 (two Newswire collections and one Web collection[4]). We submit each query to an IR system and estimate the parameters of an SD model using the relevance judgments for that query. We then estimate the inferred average precision for each of the methods outlined. We report the Kendall-τ correlation and RMS (root mean square) error of real average precision to inferred average precision over a set of queries. We conducted these experiments on two IR systems (BM25 with default parameters and a language model LM using Jelinek-Mercer smoothing set to 0.2). It is important to note that both of these systems return positive scores as we used the modified BM25 [10] and the language model with JM smoothing from [17].

4.1 Performance Comparison

Table 2: Details of Collections Used

	# docs	# Topics	Range	Avg Qry Len
AP	242,918	149	051-200	3.6
FT	210,158	188	251-450	2.5
WT10G	1,692,096	100	451-550	2.5

We can see from Figure 1 for all three collections that **AuPR** outperforms **E[ap]** as there is a higher correlation with real average precision. Another point to note is that **AuPR** needs relatively few points (i.e. documents on the score line s) to reach maximum performance. On all three collections, after 64 points have been sampled, the correlation of **AuPR** with actual average precision is near its maximum. A similar trend is reported for the RMS error (Figure 2) and the results are consistent for both IR systems used. Overall **AuPR** is the more accurate approach of the two analysed and can be calculated using far fewer documents (N).

Using an SD model that adheres to the RFCH smooths the probabilities of relevance over the entire score range s in a principled manner. Using only a limited interval on the score line s, (as is the case for **E[ap]** which uses the initial discrete ranking) may ignore useful information at high scores. Approximating the integral (as is the case for **AuPR**) is a more accurate approach. We hypothesise that using only a part of this range effectively reduces the accuracy of the inference from the model. Furthermore, as score distributions effectively model an infinite collection of document scores in a principled manner, it is more intuitive to infer the effectiveness metric from the continuous domain. Modelling the actual ranking of documents as a sample drawn from an infinite collection was proposed recently [14] and it has been shown that average precision values, when smoothed appropriately, tend to follow a normal distribution. It must be noted that the **E[ap]** was not originally developed for the task of inferring average precision from SD models, and so this research does not dispute its' importance in other areas of IR.

5. CONCLUSION

We have presented an empirical study of two methods of inferring average precision from SD models. We have shown that approximating the area under the precision-recall curve directly from the SD model is the better of the two approaches in terms of accuracy.

Acknowledgments

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

6. REFERENCES

[1] Avi Arampatzis, Jaap Kamps, and Stephen Robertson. Where to stop reading a ranked list?: threshold optimization using truncated score distributions. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 524–531, New York, NY, USA, 2009. ACM.

[2] Avi Arampatzis and Stephen Robertson. Modeling

[3] While the integral can be transformed into an integral over a finite interval, the method used here produces a suitable approximate of average precision. It also enables us to control the number of uniformly sampled documents (N) on the score line.

[4] Outlined in Table 2 and available from http://trec.nist.gov/

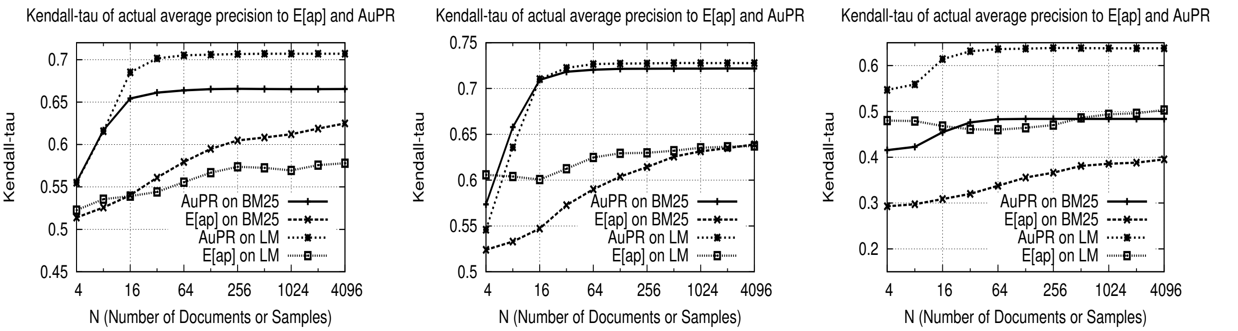

Figure 1: Kendall's τ Correlations of real average precision compared to inferred average precision for both E[ap] and AuPR on AP, FT, and WT10G collections respectively (left to right).

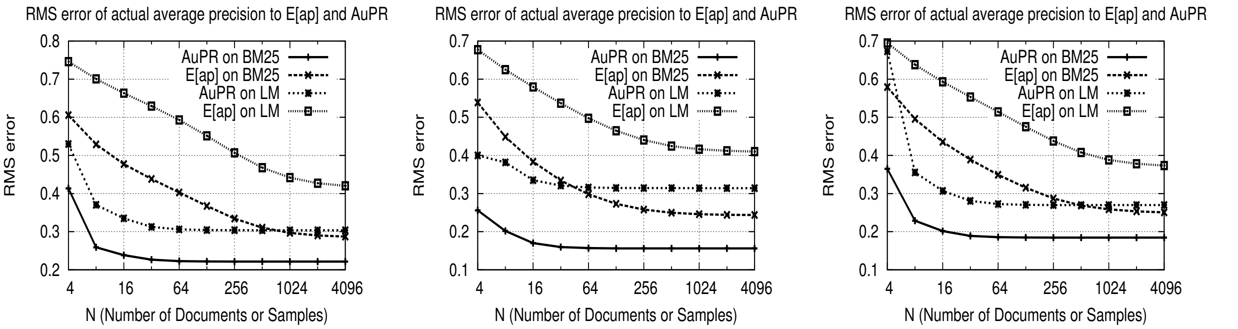

Figure 2: RMS (Root Mean Square) error of real average precision to inferred average precision for both E[ap] and AuPR on AP, FT, and WT10G collections respectively (left to right).

score distributions in information retrieval. *Inf. Retr.*, 14(1):26–46, 2011.

[3] Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. The maximum entropy method for analyzing retrieval measures. In *SIGIR 2005*, pages 27–34, New York, NY, USA, 2005. ACM.

[4] Abraham Bookstein. When the most pertinent document should not be retrieved - an analysis of the Swets model. *Inf. Process. Manage.*, pages 377–383, 1977.

[5] Ronan Cummins. Measuring the ability of score distributions to model relevance. In *AIRS*, pages 25–36, 2011.

[6] Ronan Cummins. Predicting query performance directly from score distributions. In *AIRS'11*, pages 315–326, Berlin, Heidelberg, 2011. Springer-Verlag.

[7] Ronan Cummins. Investigating performance predictors using monte carlo simulation and score distribution models. In *SIGIR*, 2012.

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

[9] Keshi Dai, Virgil Pavlu, Evangelos Kanoulas, and Javed A. Aslam. Extended expectation maximization for inferring score distributions. In *ECIR 2012*, Barcelona, Spain, 2012. ACM.

[10] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR*, pages 480–487, 2005.

[11] 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.

[12] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR 2001*, pages 267–275, New York, NY, USA, 2001. ACM.

[13] Stephen Robertson. On score distributions and relevance. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 40–51, Berlin, Heidelberg, 2007. Springer-Verlag.

[14] Stephen Robertson. On smoothing average precision. In *ECIR*, pages 158–169, 2012.

[15] Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. Extending average precision to graded relevance judgments. In *SIGIR 2010*, pages 603–610, New York, NY, USA, 2010. ACM.

[16] John A. Swets. Information retrieval systems. *Science*, 141(3577):245–250, 1963.

[17] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.