Modelling Score Distributions Without Actual Scores

Stephen Robertson
University College London
Gower Street
London WC1E 6BT, UK
stephenerobertson@hotmail.co.uk

Evangelos Kanoulas
Google Inc.
Zurich
Switzerland
ekanoulas@gmail.com

Emine Yilmaz
University College London
Gower Street
London WC1E 6BT, UK
emine.yilmaz@gmail.com

ABSTRACT

Score-distribution models are used for various practical purposes in search, for example for results merging and threshold setting. In this paper, the basic ideas of the score-distributional approach to viewing and analysing the effectiveness of search systems are re-examined. All recent score-distribution modelling work depends on the availability of actual scores generated by systems, and makes assumptions about these scores. Such work is therefore not applicable to systems which do not generate or reveal such scores, or whose scoring/ranking approach violates the assumptions. We demonstrate that it is possible to apply at least some score-distributional ideas without access to real scores, knowing only the rankings produced (together with a single effectiveness metric based on relevance judgements). This new basic insight is illustrated by means of simulation experiments, on a range of TREC runs, some of whose reported scores are clearly unsuitable for existing methods.

Categories and Subject Descriptors

H.3.4 [**Inf Storage & Retrieval**]: Systems and Software

Keywords

effectiveness, models, score distributions

1. INTRODUCTION

Models of the output of information retrieval systems based on the notion of scoring and ranking, combined with the evaluation notion of relevance, and treating the scores of relevant/non-relevant documents as arising from statistical distributions (one for each class), have been around for half a century [22]. Such models can potentially explain aspects of the behaviour of systems, in particular the shape of effectiveness curves such as recall-precision curves (e.g. [8]). Purely as abstract models of system behaviour they may help to explain and understand aspects of effectiveness (such as the effect of test collection size on effectiveness metrics, [12]). They make it possible to do certain kinds of simulation experiment (for example to improve understanding of system behaviour or that of effectiveness metrics under different conditions – e.g. [17]). They also have potential application, and indeed have been applied, directly to various practical tasks in information retrieval, most obviously to threshold setting and to merging results lists from different systems (e.g. [7, 15, 1]).

This paper is not about any of these practical applications, but about the models themselves. We develop a fundamental aspect of this approach, implicit in the original work in the 1960s but ignored in all subsequent work. We make the approach much more robust, by removing its dependence on essentially arbitrary aspects of any scoring formula which do not affect the ranking. Even better, we make it possible to apply score-distribution models to systems which rank documents without producing or revealing scores at all, or whose scores are unsuitable for distributional modelling. Thus we achieve a considerable extension of both the scope and the robustness of this modelling approach to IR systems.

ICTIR'13, September 29–October 02, 2013, Copenhagen, Denmark.

http://dx.doi.org/10.1145/2499178.2499181 .

2. BACKGROUND

Many systems rank search results by means of some form of scoring: each candidate item is given a score measuring in some sense how closely it fits with the query, and the items are ranked in descending score order. A significant literature has built up around the analysis of scores, in particular on modelling score distributions, from Swets in the 1960s [22, 23]. It is generally assumed in this work that these scores are available, and actual numerical scores are usually used. This is somewhat problematic, because the relationship between scoring and ranking has a lot of redundancy. Given a scoring method which produces a certain ranking, we could derive any number of different scoring functions which produce exactly the same ranking: any monotonic transformation of the original score will have this property. Even if a system does actually reveal a numerical score, we often have no knowledge of what transformations may have been applied before the score reached us.

An example is provided in the files recording submissions to TREC. Generally these include for each topic a ranked list of items and a score for each one. The only requirement from the TREC side on this score is that it should be consistent with the ranking.[1] In some cases it looks like a score which might have been generated by one of the familiar scoring algorithms (e.g. BM25 or the query-likelihood language model), but in others it looks quite different. Sometimes it

[1]trec_eval actually ignores the presented ranking and uses the ranking implied by the scores instead.

is clear that system output was just the ranking of the top 1000 items, and a score has been generated by subtracting the rank-position from 1001 (entirely acceptable by TREC rules). If indeed such a system ranks on the basis of an internal scoring function, then this rank-based score is actually a monotonic transformation of the internal score – but of a very peculiar kind, and making any kind of inference about the characteristics of the internal scoring, or any kind of distributional analysis, impossible.

Further, there are systems (particularly web search engines) which make an initial Boolean selection and then apply a scoring function within the selected set. The scoring function is often undefinable for items outside the selected set, using features of the documents which can be derived only for those satisfying the Boolean requirement. There are other possible configurations – for example, a system which performs some initial retrieval, perhaps with a standard type of scoring, and then re-ranks the top-ranked documents based on some criterion which only applies to these documents; or a system which fuses the top-ranked items from multiple searches; or a system which computes a score for every document, chooses the single top-ranked item, and then recomputes the score for each successive choice. Thus the notion of a single scoring function applicable (in principle) to all documents in the collection, which defines the final ranking, may be quite far from the truth.

Score distribution modelling is nevertheless a potentially useful device in a number of contexts, as discussed above. But the dependence of explicit score distribution modelling on specifically reported numerical scores, and the likely sensitivity to perhaps grossly non-linear transformations of these scores, limits its usefulness. It may, however, be possible to tap into the abstract theoretical model of IR embedded in the score-distribution notion, without actually using explicit scores. The present paper develops this notion. In the next section, we develop a theoretical analysis of the kind of smooth models of effectiveness which the score-distribution models attempt to supply, and how such models might be fitted to real data in the absence of scores. This leads to the main conclusions of the paper (Section 4). Some limited, initial experimental results using simulations follow (Section 5). As well as providing some empirical support for the theory, these serve to illustrate one way in which the models might be used.

3. SMOOTH MODELS

3.1 Swets re-examined

A half-century ago, Swets [22] proposed a way of looking at the results of a traditional Cranfield-style IR test, inspired by work on signal detection theory. The elements of this model, slightly re-interpreted for the present paper, were: 1. The system produces in response to a query a full or partial ordering of documents, which we can *model* as resulting from a scoring-and-ranking process; 2. We can then model the effectiveness curve as deriving from the distributions of scores of relevant and non-relevant documents respectively; 3. One hypothesis concerning these distributions is that both are Gaussian, with the same variance.

By *effectiveness curve*, we might mean for example a recall-precision graph showing effectiveness over the ranking. What Swets used was a recall-fallout graph, because this is easier to tie to the distributional model. Although in his second paper he made an empirical analysis using real test data, he did not use real scores. Instead, he drew inferences from his model about the shape of the recall-fallout curve, and tested these inferences – specifically that the recall-fallout graph, after suitable transformation of both metrics, would be a straight line.

We might see his model as follows: that systems actually generate scores, and that these scores actually follow a Gaussian distribution. However, given the redundancy between scoring and ranking, we can make a slightly more sophisticated version of the hypothesis as follows:

Swets hypothesis revised: If system scores are transformed so that the distribution of non-relevant scores is Gaussian, then the distribution of relevant scores will also be Gaussian, with the same variance.

This is as a much weaker hypothesis. Given a distribution of non-relevant scores, it takes only very weak assumptions to ensure that the scores *can be monotonically transformed* so that the distribution is Gaussian (see the Appendix). So the only remaining question is, what does this same transformation do to the relevant score distribution?

This version is effectively independent of actual scores: given some real system scores, we are free to transform them in any monotonic way (linear or not) to fit the hypothesis. Given a system which ranks without generating explicit scores, we can in principle assign our own scores consistent with the ranking, so as to fit the hypothesis. We could even throw away any explicit system scores and infer our own, again consistent with the ranking. As will be seen below, we can do without scores altogether.

Most work on score-distribution models (e.g. [9, 8, 7, 4, 15, 10, 1, 3, 2]) has used explicit scores provided by systems and fitted distributions to these scores. Some of this work has performed some form of (usually linear) score normalisation. To our knowledge, none has tried to improve the fit of any particular distributional model with any non-linear transformation of the scores, and none is able to work without scores.

3.2 Theoretical basis

Actual distributions of scores are messy, being both discrete and unsmooth in various ways. This is not only because we only ever observe a finite number of documents and their scores, but also because of how the scores are derived (e.g. a weighting scheme based on tf and idf has a discrete range of actual values, since both tf and df must be integers). The observed samples are usually small because for a given query, the total number of relevant documents in a collection is often small; and in the case of non-relevant documents, we are interested only in the extreme outliers, which also behave like a small sample (see [14]). Nevertheless, we are interested in models which smooth over this messiness. What might be the theoretical basis for introducing such a smooth model?

We might for example [17] regard the test collection as having been sampled from a large or infinite population of possible documents. In this view, the smooth model is assumed to apply to the infinite population; all messiness arises from the sampling. This is only a partial explanation: it deals with the messiness that arises from small-sample causes, but not with any that might arise from the actual scoring formula. We have *also* to assume that the actual scoring formula provides only an approximation to some no-

tional, truly continuous measure. For practical purposes, this assumption is probably not very onerous.

There may be other ways in which we could see our observations as arising from a smooth model. For example, we could see each document text as noisy, arising from language-model-style choices of words by the author, with a random element. Another possible source of noise lies in the relevance judgement – we may assume that the judgement, and therefore the resulting assignment of a score to one or other distribution, has a random element. Both of these may provide additional explanations for observed messiness, but neither (on its own or in combination) can fully bridge the gap between messy observations and a smooth model.

3.3 Smoothness assumptions

The usual kind of a smoothness assumption is to assume that each score distribution follows some known continuous form. But Swets does not attempt to fit the distributions to actual scores: he first infers a smooth effectiveness curve, and then observes empirical curves. As discusssed above, almost all later work (from Brookes [9]) uses scores directly. Here we follow Swets' example. (But we note a further difference, that Swets' curves are summaries over sets of topics, while the distribution-fitting approach is applied topic-by-topic. Swets' accumulation over topics is problematic; we follow the topic-by-topic approach.)

Effectiveness curves reflect the ranking rather than the actual scores – any transformation of the scores that preserves ranks also preserves the effectiveness curve. Thus a smoothness assumption might be expressed in terms of the effectiveness curve rather than the actual scores. Following previous work, we define recall and fallout as the cumulative distribution functions for the relevant and non-relevant score distributions at a score threshold (actually reverse cumulative functions, starting from the high-score end). Precision at a score threshold is a function of recall, fallout and generality. A recall-fallout or recall-precision effectiveness curve results from varying the score threshold. The curves are invariant under all monotonic transformations of the scores.

A smooth distribution hypothesis will necessarily lead to a *smooth effectiveness curve hypothesis* (SECH), but we may also formulate a SECH directly. We note two recent uses of something like a direct SECH. The first is the maximum entropy approach [6]. Here, a recall-precision curve is derived from a single per-query measurement of average precision, and an assumption that the curve is the one which maximises the entropy. The second [5] explicitly assumes that the R-P curve is one of a family of curves controlled by a single parameter, which is fitted again by means of a single per-query measurement, this time of R-Precision. (This is the AY family, below.)

These arguments suggest three possible ways to approach smoothness, apart from the usual score-distribution-fitting: 1. We use a distributional model to derive a SECH (e.g. the Swets two-Gaussian-equal-variance model); 2. we come up with a SECH directly (e.g. the AY family of R-P curves); 3. we use a principle which allows us to infer a smooth model (e.g. the maximum entropy principle). In any of these cases, we can then investigate the SECH empirically, including fitting any necessary parameters and testing goodness of fit, without making any use of the scores themselves. We note further that [13] takes a step in the reverse direction, by assuming a SECH and a distribution of non-relevant scores, and inferring the score distribution for relevant documents. This step is central to the arguments and to the simulation experiments described below.

3.3.1 Model parameters

One consequence of moving from explicit scores to SECHs is a reduction in the number of parameters that we need to estimate. For example, in the original Swets model (two Gaussian distributions of equal variance) we need three parameters (two means μ_R, μ_N, for relevant and non-relevant respectively, and one variance σ). However, the effectiveness curve is fully determined by a single parameter which can be expressed as $\frac{(\mu_R - \mu_N)}{\sigma}$ [20]. The redundancy in the actual scoring allows us to throw away two parameters (scale and absolute location). (We note also that some inferences from the orginal Swets hypothesis require a fourth paramer, generality (= prior probability of relevance). The SECH version may still require this parameter, see below.)

The two papers which use something like a direct SECH fit a single measurement of retrieval effectiveness with a smooth model. In [6] this measurement is average precision; in [5, 13] it is R-Precision. In principle, a single measurement should be enough to fit any single parameter model. As pointed out in [5], if the smooth model takes the form of a single-parameter family of recall-precision curves, then it is easy to fit the parameter from R-Precision, which defines a specific R-P point on the positive diagonal through which the curve must pass. In principle it should be possible to fit a single parameter to average precision (interpreted as the area under the R-P curve), but it may be intractable to express this in analytical form.

Some further distributional models reduce to single-parameter families of R-P curves, and can therefore be treated in a similar way; two examples are given below. Other distributional models require more parameters: for example, the Swets unequal-variance-Gaussians model has four parameters, which should reduce to two for a SECH. We do not pursue any such models here. The maximum entropy principle is more flexible: we can (in principle, again) introduce any number of constraints within which the entropy has to be maximised. Thus we can require it to fit average precision only, or include some other parameter(s) as constraints. The use of this principle requires further exploration, not attempted in this paper. Here we discuss some possible SECHs and the estimation of parameters. In section 5 we report on some experiments, illustrating the estimation process and also one example of a use to which these models might be put.

3.4 Some SECHs

Given (a) a distribution of non-relevant scores, and (b) a SECH with estimated parameter values, we can infer the distribution of relevant scores, as shown in [13]. We can now take this principle a stage further. We ignore actual scores, and rely on the SECH to give us all the information that we need to know about the ranking effectiveness of the system. Now we are at liberty to choose *any* reasonable non-relevant distribution and infer the corresponding relevant distribution. In effect we infer the behaviour of some scoring function which would be consistent with what we know about the ranking. The scoring function whose behaviour we infer could be any monotonic transformation of

the original scores – it matters not what this transformation is, nor does it need to be made explicit.

But we do need a good SECH for this purpose. Here we discuss three smooth models for recall-precision curves. The first is based on a simple, direct assumption about the curves themselves; the other two derive from assumptions about score distributions. However, we need first to introduce some useful relationships. The basic ideas are taken from [18], with some development from [13]. Given a pair of score distributions, $f_R(x)$ and $f_N(x)$, for relevant and non-relevant documents respectively, and given a score threshold t, we identify recall and fallout with the cumulative distribution functions (cumulated from the right) as follows:

$$r(t) = \int_t^\infty f_R(s)ds \quad \text{and} \quad n(t) = \int_t^\infty f_N(s)ds$$

Now we can derive a formula for precision at t, $p(t) = \frac{r(t)}{r(t)+O\cdot n(t)}$, where O is the prior odds of non-relevance in the collection (that is, another way of expressing generality, see section 3.3.1).

3.4.1 Family 1: the AY family

A single-parameter family of smooth recall-precision curves is defined in [5] by the equation $p(r) = \frac{1-r}{1+\alpha_1 r}$. This curve starts at $(0, 1)$ and ends at $(1, 0)$, and is symmetrical about the line $y = x$. We can estimate an appropriate α_1 for a particular outcome (ranked and relevance-assessed documents for a given system and topic) by fitting the curve using the value of R-Precision, rp. This metric defines a point (rp, rp) through which the curve should pass; solving for α_1 we get

$$\alpha_1 \approx (1/rp - 1)^2 - 1 \tag{1}$$

We note, however, that this is not the only way to estimate α_1 to fit a particular outcome; it may suffer from the fact that it tries to fit only a single point on the curve. Below we discuss a different method using average precision. Nevertheless, this method fits actual recall-precision curves surprisingly well [5].

3.4.2 Family 2: an exponential family, E

If we assume (one of the possible models originally discussed by Swets) that both distributions are exponential, we can derive a simple relationship between recall and fallout: if the exponential rates (inverse of mean) are λ_R and λ_N respectively, then $n(t) = r(t)^{\frac{\lambda_N}{\lambda_R}}$. Now we obtain $p(r) = \frac{1}{1+O\cdot r^{\alpha_2}}$, where $\alpha_2 = \frac{\lambda_N}{\lambda_R} - 1$, normally positive. This curve starts at $(0, 1)$, but ends at $(1, \frac{1}{(1+O)})$, which is actually the correct place (precision is then equal to generality, the prior probability of relevance in the collection, rather than zero). Again, we can estimate α_2 from rp:

$$\alpha_2 \approx \frac{\log(\frac{1}{rp} - 1) - \log O}{\log rp}. \tag{2}$$

3.4.3 Family 3: a logistic family, L

If we assume two logistic distributions of equal variance, we have a simple linear relationship between $\text{logit}\, r$ and $\text{logit}\, n$. This translates to $n(t) = \frac{r(t)}{\alpha_3 - r(t)(\alpha_3 - 1)}$ for some α_3, which will normally be greater than 1. From this, $p(r) = \frac{\alpha_3 - r(\alpha_3 - 1)}{\alpha_3 + O - r(\alpha_3 - 1)}$. This curve does not start at $(0, 1)$, but at $(0, \frac{\alpha_3}{(\alpha_3+O)})$ – somewhere between 0 and 1 on the precision axis. This is consistent with theoretical results reported in [14]. Again, we can estimate α_3 from rp:

$$\alpha_3 \approx \frac{rp(rp + O - 1)}{(1 - rp)^2}. \tag{3}$$

Graphs of all three families are shown in Figure 1. We note also that the score-distribution models on which the E and L families are based both satisfy the convexity hypothesis [18].

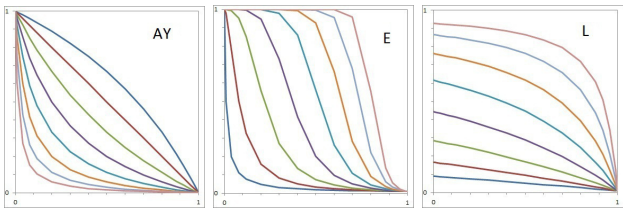

Figure 1: R-P graphs for three families of SECHs: AY, E and L. x = Recall, y = Precision.

3.5 Estimation of alpha

For all the families described, the suggested estimation of the relevant α parameter from rp is based on ensuring that the fitted curve passes through the single point (rp, rp). It might be better to fit the curve parameter α to the observed average precision, avp, rather than to the observed rp, because avp is based on the entire curve. We note that average precision is the area under the recall-precision curve. In principle we could integrate over the theoretical curve and fit the parameter by equating this area with the observed avp. However, none of the three families discussed above results in a tractable analytical formula for the integral, to allow us to express the α parameter in terms of observed avp. We present an alternative method, based on numerical integration

The principle of the method is to compute upfront the avp values associated with a grid of α values, and then, for each system-topic pair, to find the nearest avp in the resulting table. This would give us an estimated α for this case. However, the method needs to take into account two more factors: 1. We have already seen that the odds O is involved in the relationship, so we need to include O in the analysis. 2. Observed avp is calculated to some threshold recall, namely the recall attained by the system at whatever rank threshold is used (as opposed to recall=1); we therefore need some way to deal with this incompleteness.

Another aspect that we need to think about, specifically for simulation, is whether we attempt to model the entire collection of documents (with a corresponding extremely large O value), or whether we attempt to limit to some set of top-ranked documents. Most work on explicit score distribution modelling tries to model only the retrieved set, though it is also possible [1] to regard the retrieved set as a sample from a true distribution that has been truncated. Simulation of the entire collection is probably intractable on any scale. However, the method suggested here allows an intermediate approach. We assume an arbitrary threshold some way down the ranking (which we would never normally reach), by which point it is likely that all relevant documents would have been retrieved – in the simulations reported below, this

limit is set at 5000. This defines the full length of the recall-precision curve. Then we estimate the curve parameter α by fitting *avp*, observed at an earlier threshold, in an appropriate way. This gives us a model which extends (courtesy of the smooth curve assumption) well beyond where we have observed it. We can now construct two different ways of fitting observed *avp*. One is to make use of the observed value of recall at rank 1000. We tabulate the values of the area under the R-P curve for a grid of the following, each divided into bins: $\alpha \times O \times$ recall. That is, we numerically integrate the assumed smooth R-P curve for the given combination of α and O, from zero to the given recall value. Then, for each topic (with its odds O and its observed *avp* and recall at rank 1000), we can find the α value that gives the closest match. The second method is to project a corrected *avp* value for the whole curve, and then use an $\alpha \times O$ grid. In the experiments reported below, this method was used, and a heuristic correction was found to give reasonable results.

4. INSIGHTS

The core of the present paper is a theoretical insight into modelling effectiveness data: that the use of score distribution models need not be dependent on the availability of actual scores, or on any assumptions about actual scores if they are available. We have seen that we can in principle set up a smooth model such as a SECH, and fit it to data arising from a conventional IR experiment (i.e. estimate the parameters of the model for each system-topic pair). The smooth model may or may not derive from score-distribution assumptions, but in either case we can interpret the model as a score-distribution model. This is the primary conclusion of this paper.

We report some very limited experimental results in support of this insight, based on simulation, and using the results of a set of TREC runs as detailed below. Here we specify the purpose of the experiments. Simulations are based on models of the score distributions, but do not use actual scores. Instead, we

1. assume a one-parameter smooth effectiveness model;
2. fit this model using a single effectiveness measurement for each system-topic pair;
3. assume a fixed non-relevant distribution, the same for all topics and systems, with no estimated parameters;
4. infer from the fitted smooth model the corresponding relevant distribution for each system-topic pair;
5. run a simulation by sampling from these distributions.

In the case of a SECH model, fitting the model means estimating the appropriate α parameter; in the case of maximum entropy, steps 2 and 4 go together: fitting the model would be exactly inferring the relevant distribution. Our theoretical insight implies that (a) using actual scores is unnecessary, and (b) the results will be independent of the particular assumption made at item 3. The object of the experiments is to provide some preliminary empirical support for these assertions, and some comparison of the three different SECH models.

5. EXPERIMENTS

In the immediately preceding section, we set out the aims of the experiments. The basic method is to model (query-by-query) the output of a range of TREC systems using the methods discussed above, to run a series of simulation experiments based on the models, and to compare the simulated results of different models with each other and with the real results.

All data used in these experiments was taken from the TREC 2004 Robust track: 249 topics and approximately half a million documents, and also the reported run results (ranked lists for each topic) submitted to TREC. 110 runs were submitted, by 14 different research groups. In order to test the ideas of this paper on as full a range as possible of system types, without an explosion in the number of experiments, we chose 14 runs, the single best (by MAP) run for each group. In terms of the reported scores, these included a full range from runs based on a single scoring function potentially applied to every document, to those (two runs) where the scores appeared to be reverse-engineered from ranks, as discussed above.

5.1 A brief account of the simulation process

The purpose of the simulation, based on a model such as a score-distribution model, is to obtain multiple measurements of any particular evaluation metric, on each system-topic pair, as if we had multiple test collections sampled from a population whose characteristics are encapsulated in the model. We can then assess whether the observed results (one measurement for each system-topic pair) could plausibly have arisen from the model that we used: a good model is one from which the observed results could plausibly have come. In these experiments, we consider the three models arising from the three SECH families, together with (very briefly) one bootstrap model and one model involving fitting distributions in the conventional way. Plausibility is assessed in terms of an informal goodness-of-fit analysis, using both mean and variation. Similar simulation experiments are presented in [11, 17, 21]; the last citation discusses some of the general issues involved.

Given relevant and non-relevant score distributions, we draw a sample of appropriate size from each, jointly rank the results, and evaluate each ranked list in the usual way, using any of the usual metrics. As discussed in Section 3.5, we simulate the top 5000 ranked documents (1000 only for bootstrap and GMG). This process is applied to each system-topic pair, and repeated 1000 times. Thus for each single metric M, defined for a single system-topic pair, we have a single observation and 1000 simulated values. Each simulated value represents what might have been the result from a different sample of documents from the same population.

The detail of the simulation process is somewhat complex, and is not described here; there is some discussion in [21]. However, we need to discuss one issue, concerning the relevant distribution. In the two cases of the bootstrap and the directly fitted distributions, we have an explicit distribution from which to sample. In the SECH-based models, we infer the relevant distribution from a smooth model and an assumed shape for the non-relevant distribution, following [13]. But inferring an analytical form of the density which allows us to draw samples is not possible in all cases; instead, we use a slice-sampling method [16]. Slice-sampling is a Markov chain method; it depends on the principle that one can sample from a distribution by sampling uniformly from the region under the plot of the distribution's density function. This is achieved by alternating sampling in the vertical di-

rection with sampling from the horizontal "slice" defined by the current vertical position. That is, starting from an initial x value, the method evaluates $p(x)$ and samples u uniformly in the range $[0, p(x)]$. Then it samples x uniformly from the slice through the distribution defined by $x : p(x) > u$. Given that the method is a Markov chain method there are two important details using the method. First, the slice-sampling algorithm discards a number of samples to allow the Markov chain to approximately reach stationarity. Second, it continues discarding samples throughout the entire process so that sample points are not serially correlated.

5.2 Reporting results

Simulation will deliver results for any chosen evaluation metric, but the metric used here is yaAP, as discussed in [19]. This is a smoothed version of the logit transform of average precision, and is chosen because of its good distributional properties; it can be derived from average precision if the total number of relevant is also known. We report results in two main ways. In the first, we plot observed (x) against mean simulated (y) values of yaAP for each system, one point for each topic; in the y direction we also show a range, representing one standard deviation of the simulated values. Thus each plot shown represents a single system and all the topics used in the evaluation. This form of plot reveals obvious biases which may occur with a poor model. Some examples may be seen in Fig 2.

Secondly, we compute for how many topics the observed value is at an extreme of the empirical distribution of simulated values: in the bottom or top 2.5% of the distribution, or outside the range of the simulated values. Extreme values are reported as counts in four categories: observed below all simulated / in bottom 2.5% / in top 2.5% / above all simulated. We interpret the extreme values is as follows: if we suppose that our simulated distributions represent correctly the results of sampling from the hypothetical infinite document population, and our actual test collection was indeed sampled from this population, then among the 249 topics, we might expect about 2.5% (≈ 6) to fall within each of the 2.5% tails of their respective distributions, and none or very few to give observations outside the range of their respective distributions – thus we expect to see something like 0/6/6/0. Much larger numbers would suggest that our models are not good; zeros might suggest overfitting. We report these results for some individual systems (all topics), and also averaged over systems.

5.3 Use of the data

The SECH-based models have no access to the reported scores of documents. For each simulation, each topic and each run, we use a single metric (either RPrecision or average precision) to fit the smooth model. Some versions of the average precision method require knowledge of the total number retrieved (which may be less than the usual 1000), in order to estimate a correction element to the observed average precision. We obtain from the qrels file the total number of relevant items for each topic, to get a measure of the topic's generality (= prior probability of relevance). The actual simulation process requires similar data: the total relevant for the topic, and both the total retrieved and the relevant retrieved in this run.

We compare with a bootstrap simulation and with one based on fitted distributions. In the bootstrap version, instead of sampling from smooth distributions, we sample with replacement from observed ranked lists. This is similar to the method used in [11], which is claimed to provide good confidence intervals for observed metric values. Clearly the bootstrap method needs to be provided with the actual ranked and labelled lists of items for each system-topic pair. In the fitted distributions version, we use the GMG (Gaussian Mixture and Gamma) model, developed by Kanoulas et al [13] – relevant distribution as a mixture of Gaussians and non-relevant as a Gamma. Estimated parameters provided by the fitting process are the means and variances of the Gaussians, their relative preponderance, and the shape and scale parameters of the Gamma. For this simulation, we need not only the ranked lists but the actual scores of ranked documents. It is not suitable for systems which do not generate reasonable scores, and in particular, we cannot use it for the two TREC runs where the reported scores are apparently reverse-engineered from ranks.

6. RESULTS

First we consider the three SECHs discussed above, all fitted with RPrecision, together with the assumption that the non-relevant distribution is gamma with fixed shape (1) and scale (0.2). All simulations here model the top 5000 documents, following the discussion above. The first three plots in Fig 2 are for run 29. (Note that the choice of run 29 is arbitrary – except where we note otherwise, all the characteristics which we note in the graphs are reproduced remarkably closely in all runs analysed. We have no evidence at all that different systems, or even different types of system, produce different patterns reflected in these graphs.

We see that for family 1, AY, there appears to be a substantial bias among poorly-performing topics. For family 2, E, there appears to be a small bias at the opposite end.

Family 3, L, looks better than either of the other families for this run. On the whole these patterns apply to other runs as well, although some of them seem less regular – family 3, for example, seems to have a bit more difficulty with run 17. Extreme values are broadly in line with these impressions: family 1 has 14/35/0/0 on average; family 2 has 1/3/2/0, so the bias is insufficient to show up. Family 3 has 3/12/6/1 on average, showing too many low-end extremes, apparently distributed over the range of the curve and therefore not quite so obvious from the plots. Note that this form of simulation is surprisingly robust – with the exception of the biases mentioned, all three SECHs seem to produce fair models, despite the considerable difference in the shapes of the three curve families, as shown in Figure 1.

The final graph of Fig 2 shows the family 3 results for run 88. This is one of the runs for which the reported scores are reverse-engineered from ranks, and therefore cannot be modelled using any of the standard score-distribution models. We see that while the fit is not perfect (with one obvious outlying topic), it is still fairly good. Extreme values for this run are 3/11/5/2, very close to the average over all runs reported above.

Next we consider the same three SECHs and the same non-relevant distribution assumption, but fitted with average precision, using the method described at the end of section 3.5. Results are not shown for space reasons; they look very similar to the ones shown for RPrecision fitting. In particular, the same systematic biases produced by families 1 and 2 are still present. The extreme value analysis

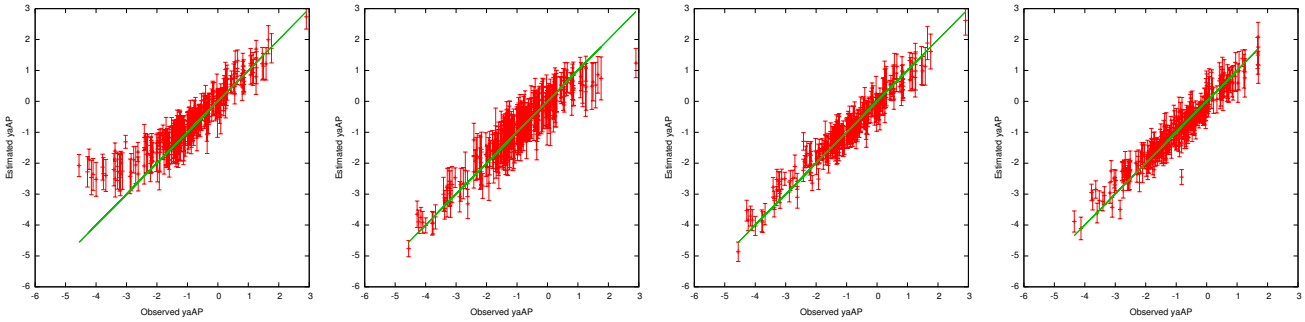

Figure 2: Three simulations of run 29, all based on SECH models fitted using RPrec, and showing yaAP: (a) the AY family of smooth R-P curves; (b) the E family; (c) the L family. (d) shows the L family simulation of run 88 (one of the two runs that does not have scores suitable for distributional analysis).

also looks similar: 14/30/0/0 (family 1), 0/0/5/3 (family 2), 0/8/2/0 (family 3); however, family 3 no longer seems to have difficulty with run 17. We note that the exact procedure for fitting curves to average precision is slightly more complex, and involves a heuristic correction to the observed average precision. The above results come after some fiddling with this heuristic correction. We conclude that the average precision method may be worth further exploration, but is not likely to result in substantially different results from the RPrecision version. For the remaining results we use family 3, the L family, as being marginally the best of the three considered, together with fitting based on RPrec.

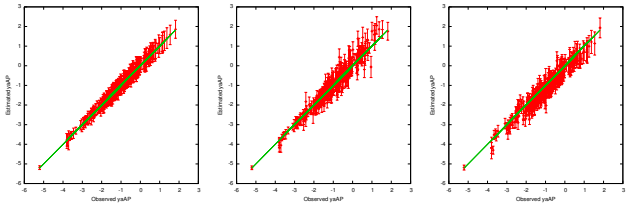

Figure 3: Simulations of run 109, comparing (a) bootstrap, (b) GMG, and (c) the fitted GMG non-relevant distribution, together with the L family SECH model fitted using RPrec. The corresponding simulation using a standard non-relevant distribution instead of the fitted one is indistinguishable from (c).

Next, we provide a more direct comparison of three models: (a) the bootstrap (unsmoothed) model; (b) the GMG model as described above; (c) the non-relevant distribution as fitted for GMG, together with the SECH using family 3; (d) a standard gamma distribution for the non-relevants, together with the SECH using family 3. These are shown in Fig 3; however, we have not shown (d) because it is indistinguishable from (c). In order to make these comparable, all simulations are based on the retrieved set only, so the (d) model differs slightly from that used for Figs 2, and on the 12 runs which appear to reveal suitable scores.

Extreme value counts for the bootstrap, for this run and for all other runs tested, are a clean sweep of zeros – this suggests strongly that the bootstrap simulation overfits the observed data. Extreme value counts for the GMG model for this run are good (0/3/4/0), so it seems that GMG is able to fit the distributions well. However, the averages over the 12 runs are 1/9/37/21, indicating that GMG has difficulty modelling at least some of the TREC runs. For (c) and (d) the averages are both the same at 0/0/22/4. Given this SECH, using the actual non-relevant distribution provides absolutely no advantage over using a standard one. This remains true even though the gamma distribution has two parameters to fit. Furthermore, the model fit as revealed by the simulation is as good as or better than with GMG. With a good SECH assumption, the *only* datum we need to simulate the full performance curve for a single topic is a single effectiveness measurement such as RPrec.

We also note that the extreme value counts for (d) are different from (worse than) those reported above for family 3. This is a function of the difference mentioned above, relating to the model – the present results are based on simulating the retrieved set only (normally 1000 documents), while the former results were based on simulating the top 5000 documents. We conclude, at least provisionally, that extending the simulation significantly further down the ranked list can improve it, which appears to be an advantage of the modelling approach of this paper – although the result needs much more exploration.

Finally we have conducted many other experiments which we have no room to discuss in full. Experiments using different shapes of non-relevant distribution confirm that the shape is largely irrelevant to the results – mean results are indistinguishable, variances sometimes differ slightly.

7. CONCLUSIONS

The primary focus of this paper is an insight about smooth models of the ranking effectiveness of search systems. All such models in current use are based on the scoring which many systems use to produce a ranking. As such, they can be used only with systems that reveal the actual scores of documents, and they have to make assumptions about the nature of these scores. We have now established that this use of actual scores is not necessary, at least for some purposes. Smooth models can be formulated, fitted and tested on the basis of the ranking only. More specifically, model fitting can be based on a single effectiveness parameter measured on each topic; this appears to be enough to provide an adequately fitted model for each topic. Furthermore, even if we abandon all use of actual scores, we can still make use of many of the score-distribution ideas of modelling IR effec-

tiveness. The simulation approach used for the experiments reported here is an example of such a use.

Experiments using these simulation models have provided some empirical support for these insights. Of the three families tested, at least the L family appears to give reasonable results over the range of TREC runs tested, despite the fact that these runs vary hugely in respect of the nature of their revealed scores, and include some that could not be modelled by present score-distribution-modelling techniques.

In this theoretical paper we have not attempted to derive any practical methods for the tasks to which score distribution modelling has been applied – this challenge awaits. There is much scope for development and potential use.

8. REFERENCES

[1] A Arampatzis, J Kamps, and S Robertson. Where to stop reading a ranked list?: threshold optimization using truncated score distributions. In *SIGIR 2009*, pages 524–531. ACM, 2009.

[2] A Arampatzis and S Robertson. Modeling score distributions in information retrieval. *Information Retrieval*, 14(1):26–46, 2011.

[3] A Arampatzis, S Robertson, and J Kamps. Score distributions in information retrieval. In *ICTIR 2009*, pages 139–151, Berlin, 2009. Springer.

[4] A Arampatzis and A van Hameren. The score-distributional threshold optimization for adaptive binary classification tasks. In *SIGIR 2001*, pages 285–293. ACM, 2001.

[5] J A Aslam and E Yilmaz. A geometric interpretation and analysis of R-Precision. In *CIKM 2005*, pages 664–671, New York, 2005. ACM Press.

[6] J A Aslam, E Yilmaz, and V Pavlu. Maximum entropy method for analyzing retrieval measures. In *SIGIR 2005*, pages 27–34. ACM, 2005.

[7] C Baumgarten. A probabilistic solution to collection fusion problem in distributed information retrieval. In *SIGIR'99*, pages 246–253, New York, 1999. ACM Press.

[8] A Bookstein. When the most 'pertinent' document should not be retrieved – an analysis of the Swets model. *Information Processing and Management*, 13:377–383, 1977.

[9] B C Brookes. The measure of information retrieval effectiveness proposed by Swets. *Journal of Documentation*, 24:41–54, 1968.

[10] K Collins-Thompson, P Ogilvie, Y Zhang, and J Callan. Information filtering, novelty detection and named page finding. In *TREC 2002*, pages 107–118. Gaithersburg, MD: NIST, 2003.

[11] G V Cormack and T R Lynam. Statistical precision of information retrieval evaluation. In *SIGIR 2006*, pages 533–540. ACM, 2006.

[12] D Hawking and S Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6:99–150, 2003.

[13] E Kanoulas, K Desai, V Pavlu, and J A Aslam. Score distribution models: Assumptions, intuition and robustness to score manipulation. In *SIGIR 2010*, pages 242–249. ACM, 2010.

[14] D Madigan, Y Vardi, and I Weissman. Extreme value theory applied to document retrieval from large collections. *Information Retrieval*, 9:273–294, 2006.

[15] R Manmatha, T Rath, and F Feng. Modelling score distributions for combining the outputs of search engines. In *SIGIR 2001*, pages 267–275. ACM, 2001.

[16] M. Radford Neal. Slice sampling. *The Annals of Statistics*, 31(3):705 – 767, 2003.

[17] S Robertson. On document populations and measures of IR effectiveness. In *ICTIR 2007*, pages 9–22, Budapest, 2007. Foundation for Information Society.

[18] S Robertson. On score distributions and relevance. In *ECIR 2007*, pages 40–51, Berlin, 2007. Springer.

[19] S Robertson. On smoothing average precision. In *ECIR 2012*, pages 158–169. Springer, 2012.

[20] S E Robertson. The parametric description of retrieval tests. Part 2: Overall measures. *Journal of Documentation*, 25(2):93–107, 1969. `http://www.soi.city.ac.uk/~ser/papers/Parametric_part2.pdf`.

[21] S E Robertson and E Kanoulas. On per-topic variance in IR evaluation. In *SIGIR 2012*, pages 891–900. ACM, 2012.

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

[23] J A Swets. Effectiveness of information retrieval methods. *American Documentation*, 20:72–89, 1969.

APPENDIX
Distributions and monotonic transformations

Suppose we have relevant and non-relevant score distributions with densities $f_R(s)$ and $f_N(s)$, for some scoring function s. In the usual way, we define for any threshold t on s, recall $r(t)$ and fallout $n(t)$:

$$r(t) = \int_t^\infty f_R(s)ds \quad \text{and} \quad n(t) = \int_t^\infty f_N(s)ds$$

(i.e. the cumulative distributions from the right). We make the following two assumptions: 1. The support for these two distributions is exactly the range of the scoring function; 2. the densities are non-zero over the full range. ∞ is shorthand for the maximum of this range.

Following a similar argument used by [14], we transform the scores in such a way as to make one of these distributions uniform. Define a strictly monotonic transformation on s, taking it exactly from its range to the interval $(0, 1)$,

$$\phi : s \in (-\infty, \infty) \to x \in (0, 1), \quad \text{where} \quad x = \phi(s) = 1 - n(s)$$

(again, $-\infty$ is just shorthand for the minimum of the range). $x = \phi(s)$ is just another version of the score, having exactly the same ranking effect and therefore effectiveness as s itself. The non-relevant density on the transformed score x is uniform on the range $(0, 1)$. We observe that $\frac{dx}{ds} = f_N(s)$ The relevant density on x can be derived very simply, as

$$g_R(x) = f_R(s)\frac{ds}{dx} = \frac{f_R(s)}{f_N(s)}, \quad \text{where} \quad x = \phi(s)$$

We note that this transformation is reversible. It follows that *any* non-relevant distribution satisfying the assumptions can be transformed monotonically into *any* other, via the uniform distribution on $(0, 1)$.