Reprinted from: G. Amati, C. Carpineto, G. Romano (eds), Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007. Springer, 2007. (pp 40–51)

On score distributions and relevance

Stephen Robertson

Microsoft Research, 7 JJ Thomson Avenue, Cambridge CB3 0FB, UK
ser@microsoft.com

Abstract. We discuss the idea of modelling the statistical distributions of scores of documents, classified as relevant or non-relevant. Various specific combinations of standard statistical distributions have been used for this purpose. Some theoretical considerations indicate problems with some of the choices of pairs of distributions. Specifically, we revisit a generalisation of the well-known inverse relationship between recall and precision: some choices of pairs of distributions violate this generalised relationship. We identify the choices and the violations, and explore some of the consequences of this theoretical view.

1 Introduction

The idea of modelling the distributions of scores of relevant and non-relevant documents in an information retrieval system has been around for a long time (see Swets [1]), but in recent years has taken a new lease of life [2–4]. Various combinations of statistical distributions have been proposed, for example two normal distributions of equal variance [1], two unequal variance normals or two exponentials [5], two Poisson distributions [6], two gamma distributions [2], an exponential for non-relevant and a normal for relevant [3, 4, 7], an exponential and a gamma [4].

Clearly a strong argument for choosing any particular combination of distributions is that it gives a good fit to some set of empirical data, and some of the above authors address this question in various ways. However, we do not attempt in this paper any such empirical analysis. Nor does it claim any fundamentally new theoretical results. Rather, it revisits old work [8, 9] in order to consider some theoretical properties which might be desirable for such distributions. The primary argument of the paper is that, putting aside considerations of empirical fit, some combinations of distributions exhibit undesirable or anomalous features which reduce their theoretical value. This argument generalises a point made by Bookstein [6] about the Swets unequal variance model. Some of these considerations were also aired in [10] in the context of an analysis of the relation between performance and collection size. The contribution of the paper is to bring together and clarify the theoretical issues, and to connect them with the recent work on score distributions.

Note that many other authors model or analyse score distributions without reference to relevance. This work is not discussed here. Also, the work depends on

an assumption of the binary nature of relevance; a different approach would have to be taken to take account of degrees or grades or ranks of assessed relevance. The link between relevance and ranking is assumed to be the probability ranking principle [11], which asserts that a search system should rank output in order of probability of (assumed binary) relevance.

In the next section, we introduce the main theoretical argument of the paper. In Sect. 3 we analyse in detail one of the early suggestions, the case of two normal distributions of unequal variance. Then in Sect. 4 we define a simple test and apply it to five different sets of distributional assumptions that have been suggested in the literature. Finally we discuss some further issues and conclude.

2 Recall and fallout

We consider the output of a retrieval system, as a result of a search query, to be a list of documents ranked by score or retrieval status value, and the user action to involve reading down the list until some stopping point. This stopping point then corresponds, explicitly or implicitly, to a threshold on the score: everything above this threshold has been retrieved, i.e. seen by the user; everything below has not. We further model the situation in terms of the distributions of scores in the populations of relevant and non-relevant documents: a signal detection (SD) theory view of retrieval. In this case, the two natural parameters for evaluation are recall, which corresponds to the proportion of the relevant distribution exceeding the threshold, and fallout (the same for the non-relevant distribution). We interpret these parameters in a probabilistic fashion as the values of the respective cumulative distribution functions, cumulated from the right (i.e. from the high-score end). In this case, the natural performance graph to consider is a graph of recall against fallout, referred to in the SD context as the receiver operating characteristic or ROC curve.

The recall-fallout graph is not normally used for real retrieval experiments, partly because real fallout values are typically so small, but also so unevenly distributed, that it is difficult to display such graphs in a reasonable way. One solution is to transform fallout in some way, e.g. by using a log scale. However, for the purpose of considering some theoretical characteristics, it is appropriate to think in terms of a recall-fallout graph on linear scales. We also present all such graphs with fallout on the x-axis and recall on the y-axis. All such curves may be presumed to pass through (0,0) (very high threshold, nothing retrieved) and (1,1) (very low threshold, everything retrieved). Going down the ranking from the top to the bottom, i.e. lowering the threshold, corresponds to traversing the curve from bottom left (0,0) to top right (1,1).

An example of an idealised smooth curve is shown in Fig. 1 (a detailed derivation of this curve is given in Section 3). We can also see in this figure two other properties of the recall-fallout graph on linear scales. Assuming that this curve represents a single request, the slope of the line OA from the origin to a point A on the curve is a monotonic function of the precision at point A. Also the slope of the tangent at A represents the 'instantaneous' precision –

that is, the probability that a document having exactly that score is relevant. A mathematical explanation of these points is given in Section 3. First, we formulate the Convexity Hypothesis, which provides a strong expectation on the shape of the curve.

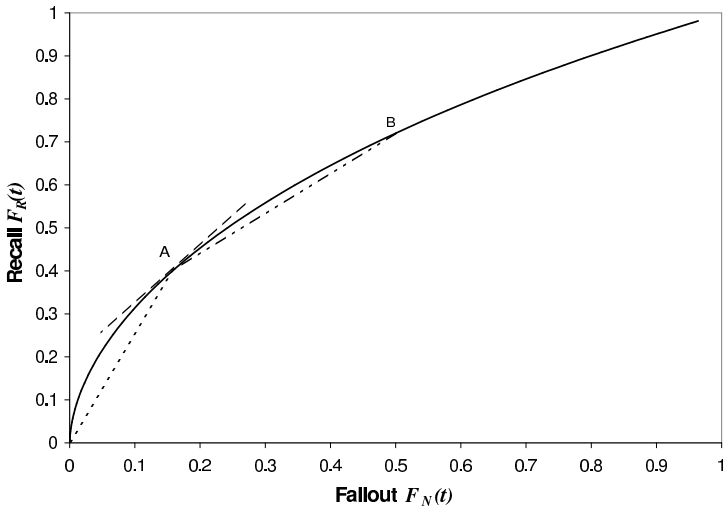

Fig. 1. Receiver Operating Characteristic (ROC) curve for the distributions discussed in Section 3.

2.1 The convex curve

The straight line on the linear recall-fallout graph, from point (0,0) to (1,1), represents a random ordering of the document collection – identical relevant and non-relevant score distributions [12]. Other straight lines may also be interpreted as random orderings of sets. For example, suppose that we have two points A, B on the recall-fallout graph, corresponding to two score thresholds t_A and t_B, with $t_A > t_B$. Then the straight line from A to B corresponds to retrieval of all documents at A (those whose scores exceed t_A), followed by a random ordering of the documents scoring between t_A and t_B.

It follows that we would in general expect the recall-fallout curve to be convex, when viewed from the top left (0,1). If we found a scoring function which generated a curve containing a concavity, we could improve upon it simply by means of a randomisation process on that section of the ranked list corresponding to the concavity in the curve – this operation would repace the concave section by a straight line, thus raising this part of the curve. (Actually, we could do better than this: the concave part represents scores which tell us something about

likely relevance, but in the reverse order – a suitable re-ordering of score values would get us as far above the straight line as the concavity is below it.) Thus even if it is not always the case that the curve is convex, we would certainly expect it of a good system, because a departure from convexity implies that the system can be very easily improved. We may therefore state the following hypothesis, with the support of the above arguments:

Convexity hypothesis For all good systems, the recall-fallout curve (seen from the ideal point of recall=1, fallout=0) is convex.

This result is related to, but somewhat stronger than, the usual inverse relationship between recall and precision – that is, the R-P relationship follows from convexity [12, 8]. We see the convexity hypothesis as a generalisation of the hypothesis of the inverse R-P relationship.

The same hypothesis can also be formulated as a condition on the instantaneous precision, or the probability of relevance of a document at an exact score. The condition is that this should be a monotonic increasing function of the score – the higher the score the higher the probability of relevance. This condition is assumed in [4]; their use of it will be discussed further below.

The convexity hypothesis is the basis for the theoretical arguments of this paper.

3 Score distributions: details and an example

Consider a pair of score distributions for relevant and non-relevant documents. In Fig. 2, we see an example of a pair of normal distributions. The normal is used as example only, but we will generally be using continuous distributions, although it is likely that the scoring function has some degree of granularity, and also we are dealing with finite collections of documents. These distributions are shown in the form of density functions (the usual bell curve). The x-axis is the score or retrieval status value, denoted v; the two distributions are denoted $f_R(v)$ and $f_N(v)$ for relevant and non-relevant documents respectively. All the equations in this section apply to any pair of continuous distributions, but the diagrams relate to the pair of normals.

As indicated above, we turn them into cumulative distributions *from the right* – see Fig. 3. These functions are defined as follows:

$$F_R(t) = \int_{v=t}^{\infty} f_R(v)dv$$

and similarly for F_N

At any given cut-off or threshold t (examples shown in the form of vertical lines), the cumulative distributions give the probability of retrieving a relevant or non-relevant document respectively at or above that threshold score. These two probabilities may be equated with the traditional measures of recall and fallout respectively. That is, the probabilities can be used as *definitions* of recall

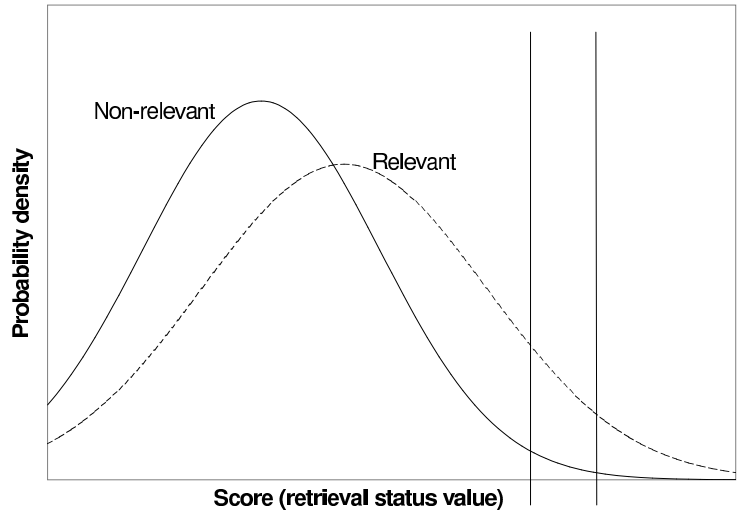

Fig. 2. SD model, normal distributions unequal variance: relevant mean 2.5 variance 1.2; non-relevant mean 1.8 variance 1

and fallout, and observed recall and fallout values are then estimates of these measures:

$$\begin{aligned}\text{Recall at threshold } t &= Pr(d \text{ retrieved at or above threshold } t | d \text{ relevant}) \\ &= Pr(v(d) \geq t | d \text{ relevant}) \\ &= F_R(t) \end{aligned} \tag{1}$$

(where d is a random document), and similarly for fallout and F_N. We can similarly define precision P, and identify it as a function of recall, fallout and generality G[1], as follows:

$$\begin{aligned} P &= Pr(d \text{ relevant} | v(d) \geq t) \\ &= \frac{GF_R(t)}{GF_R(t) + (1-G)F_N(t)} \end{aligned}$$

We reformulate this as odds:

$$\frac{P}{1-P} = \frac{G}{1-G} \frac{F_R(t)}{F_N(t)} \tag{2}$$

which gives us the monotonic relation between precision and the slope of the straight line OA of Fig. 1. Similarly we can define the odds that a document has a score between two limits:

$$Odds(d \text{ relevant} | t_1 \leq v(d) \leq t_2) = \frac{G}{1-G} \frac{F_R(t_1) - F_R(t_2)}{F_N(t_1) - F_N(t_2)} \tag{3}$$

[1] Generality is the proportion of documents in the collection that are relevant

which gives us the corresponding relation for the line AB. Furthermore, letting $(t_2 - t_1) \rightarrow 0$ gives the instantaneous precision result.

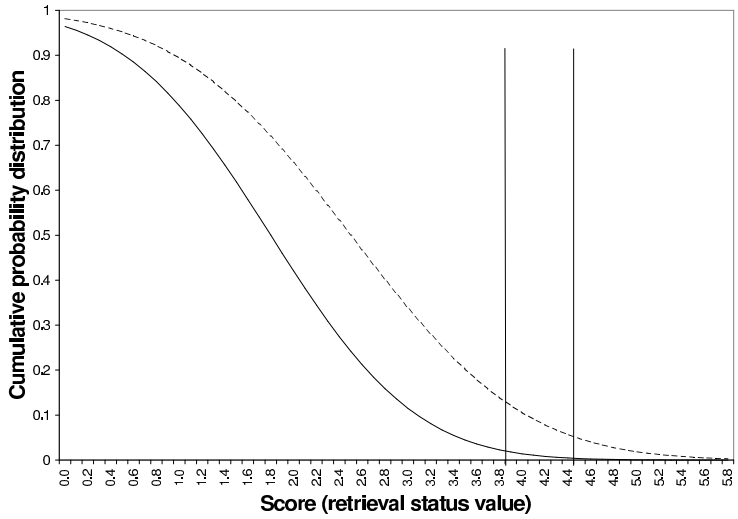

Fig. 3. SD model, normal distributions unequal variance, as Fig. 2, cumulative form

We now treat the score v as defining parametrically a relation between recall and fallout, and draw the ROC curve for these two parameters. The curve already presented in Fig. 1 is based on the distributions used here. It does not actually reach (1,1) because it was plotted only down to a threshold of zero; the assumed normal distributions both go below zero. The curve does indeed appear to be convex. However, in the full curve there would actually be a small concavity, at the right-hand end, invisible on the scale on which the graph is shown. This is because the relevant document distribution assumed, with a larger variance than the non-relevant, predicts a slightly larger number of documents with significant negative scores than the non-relevant. The curve is extended to (1,1) and the top right corner is blown up in Fig. 4; now the concavity is clearly seen.

In this case we may take this to be an artifact of the model, and of no practical significance whatever, because (a) the system probably does not calculate negative scores anyway, and (b) the number of documents in that range predicted by the distributions is probably measured in very small fractions of a document. It could be that the two normal model gives a fair approximation to real score distributions, and this theoretical anomaly is of no concern. However, the conclusion must be that the 2-normal (unequal variance) model is theoretically flawed, irrespective of its practical usefulness. It therefore seems useful to investigate the conditions under which a pair of distributions will violate the convexity hypothesis.

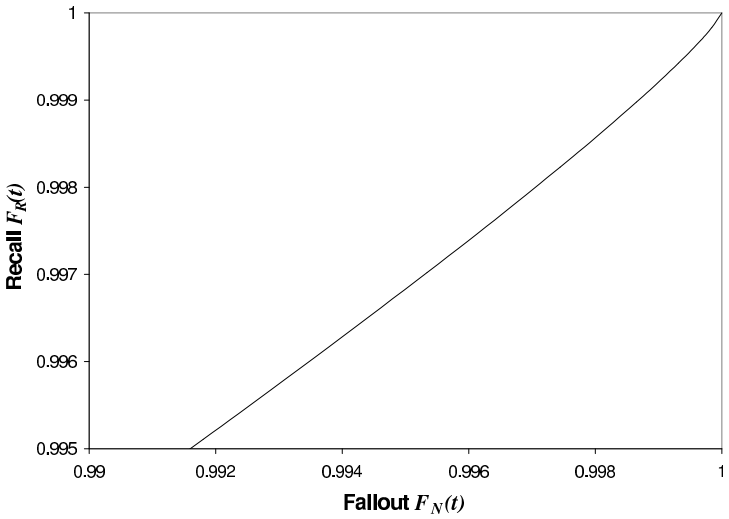

Fig. 4. Top end of the ROC curve for the distributions used in Figs 2 and 3.

4 Convexity condition and distributional assumptions

The convexity condition is given in [8] as:

$$\frac{d^2(\text{Recall})}{dt^2} < \frac{d^2(\text{Fallout})}{dt^2} \frac{\frac{d(\text{Recall})}{dt}}{\frac{d(\text{Fallout})}{dt}} \tag{4}$$

for some controlling variable t. As above, we identify recall and fallout with $F_R(t)$ and $F_N(t)$ respectively. We note that

$$\frac{dF_R(t)}{dt} = -f_R(t)$$

As the density function f is always positive, this expression is negative. The condition can be expressed as:

$$\frac{1}{f_R(t)} \frac{df_R(t)}{dt} > \frac{1}{f_N(t)} \frac{df_N(t)}{dt} \tag{5}$$

throughout the range of t. We can now test this condition on a number of the pairs of distributions that have been proposed for modelling scores. For each distribution, we need the function

$$g(t) = \frac{1}{f(t)} \frac{df(t)}{dt} \tag{6}$$

derived from its density function f, and then we can compare $g_R(t)$ and $g_N(t)$.

4.1 Two exponential distributions

The case of two exponential distributions (one of the models suggested [5]) is simple. The exponential density function is

$$f(t) = \frac{1}{\mu} \exp\left(-\frac{t}{\mu}\right)$$

where μ is the mean. Thus

$$g(t) = -\frac{1}{\mu}.$$

Since μ_R would be larger than μ_N, the convexity condition holds for all t.

4.2 Two normal distributions

Here

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(t-\mu)^2}{2\sigma^2}\right)$$

where μ is the mean and σ^2 is the variance. Thus

$$g(t) = \frac{1}{\sigma^2}(\mu - t)$$

Again, we expect μ_R to be larger than μ_N. If the two variances are equal (the first model proposed in [1]), then the convexity condition always holds. But if $\sigma_R^2 > \sigma_N^2$ (as in the example above), there will be some small value of t (or perhaps a large negative value) below which the condition is not satisfied: the reverse is the case. If $\sigma_R^2 < \sigma_N^2$, the departure from convexity occurs at the other end.

4.3 Two Poisson distributions

This combination was suggested in [6], specifically in response to the kind of anomaly just observed. The Poisson distribution is discrete, so the analysis above based on continuous distributions does not apply. However, we can define a function analogous to $g(t)$ above, as follows:

$$g(k) = \frac{P(k+1) - P(k)}{P(k)}$$

for each integral threshold $k = 0, 1, ...$, where P is the probability of observation k. The probability function for the Poisson distribution is:

$$P(k) = \frac{\lambda^k \exp(-k)}{k!}$$

where λ is the Poisson mean, from which

$$g(k) = \frac{\lambda}{k+1} - 1$$

Once more, we expect λ_R to be larger than λ_N, so the condition is always satisfied. This is consistent with the argument in [6].

4.4 Two gamma distributions

This configuration is used in [2]. The density is:

$$f(t) = \left(\frac{t}{b}\right)^{c-1} \frac{1}{b\,\Gamma(c)} \exp\left(-\frac{t}{b}\right)$$

where b is the scale parameter and c is the shape parameter; the mean is bc. Thus

$$g(t) = \frac{c-1}{t} - \frac{1}{b}$$

Here if either c or b is the same for the two distributions, but the other varies in the way we would expect (higher mean for relevants), the condition is satisfied. The range of variations for which the condition is satisfied is in fact quite wide, although one could certainly construct examples which violate the condition for some t.

In fact Baumgarten's model is slightly more complex, involving shifted gamma distributions (i.e. shifted along the t-axis by a small amount).

4.5 Exponential non-relevants and normal relevants

This combination is used in [4], [3] and [7], making it currently the most popular model. If we examine the formulae for $g(t)$ in sections 4.2 and 4.1, we see that in the exponential case $g(t)$ is constant, while in the normal case it declines linearly with t. Therefore there is always a t above which the condition is not satisfied. This affects the bottom left end of the recall-fallout graph, whatever the parameter values.

There is also a problem at the top right end (low t). Because the exponential is defined over the positive real numbers only, but the normal necessarily extends over the negatives as well, the curve hits the fallout=1 line below the recall=1 level. Thereafter it climbs straight up the fallout=1 line to the point (1,1). Thus this end also violates the convexity condition, again for all parameter values.

5 Discussion

5.1 Score range

In practice, score distributions may be truncated. It is common, for example, for scores to be constrained to be positive, either as a mathematical consequence of the scoring formula or as a matter of practical convenience. Indeed, most of the above theoretical distributions are also confined to the positive real numbers, although the normal is not. Many scoring systems also, however, constrain the scores below a maximum. For example, some produce scores that are normalized to the range [0,1]. All the above theoretical distributions extend to infinity in the positive direction. This fact produces its own theoretical problems: Should the fitted distribution be a truncated version of the theoretical one, i.e. normalised

so that its integral over the truncated range is unity? This is potentially problematic, because it affects such statistics as the mean. Many authors ignore this issue – e.g. [4] considers a scoring system which produces scores in the range [0,1], but does not worry about the implied truncation.

5.2 Non-convexity

This truncation might have the side-effect of resolving the non-convexity problem, by putting the non-convex part of the curve out of effective scoring range. In the case of [4], however, the non-convexity of the normal-exponential model does affect them, and they recognise it as a problem, at least at the high-threshold end (the non-convexity at the other end is avoided by the truncation at zero). They observe that for some of their topics, the non-convexity at the high t end falls within the scoring range [0,1]. In their terms, the probability of relevance as a function of score is no longer monotonic in these cases: after a certain point it declines. They resolve the problem by redefining the probability of relevance: when the predicted function starts declining, they replace it with a straight line from the maximum reached to the point (1,1) (that is, score=1 and probability of relevance=1). They do not give any justification for this procedure, other than that one would expect probability of relevance to be a monotonic function of score.

On the basis of the above analysis of the recall-fallout graph, one could devise an alternative procedure. Since a straight line on a recall-fallout graph represents a random ordering of some set of documents, we could perform a procedure similar to that of [4] but on the recall-fallout graph. We illustrate the procedure in Fig. 5. Replacing the concave section of the curve with the straight line is equivalent to randomly reordering all documents which score in that range. This is thus a well-founded form of extrapolation.

5.3 Monotonic transformations of the score

One characteristic of all the above analysis is that it assigns a status to scores which they might not possess. Systems produce scores in order to rank documents, and care not at all about the scale or shape of the scoring function. Thus any monotonic transformation of a score produces a new score which is indistinguishable from the old, in terms of ranking. Factors which do not affect the rank order may be arbitrarily included or removed at any stage. This fact is often used to simplify scoring functions, or their calculation.

Thus for example some scoring functions produce numbers that are restricted to the range [0,1] because they are intended to model probabilities. Independence assumptions lead one to multiply multiple probabilities; the result is another probability. On the other hand, it is often easier to use log-probabilities (or log-odds) and add them rather than multiplying them. The resulting logarithmic (or logistic) scale looks quite different, and belongs to the range $(-\infty, 0)$ (or to $(-\infty, \infty)$. But a system using such a scale might then decide to normalise back to

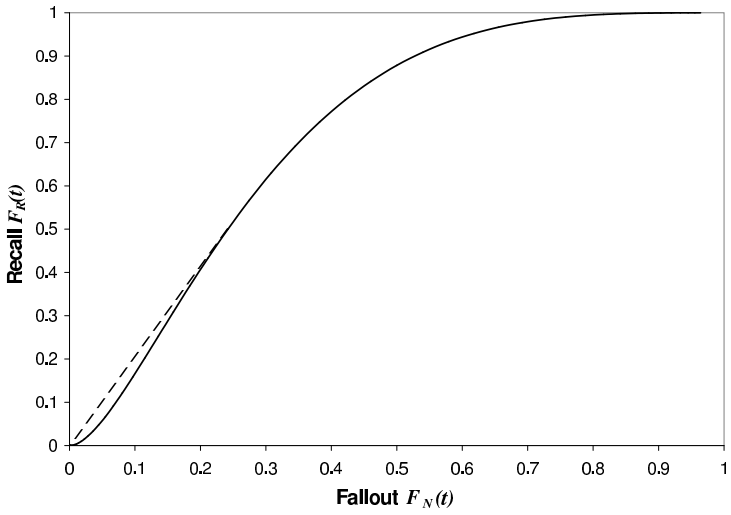

Fig. 5. Concavity at the high-threshold end.

[0,1] *linearly*, by taking account of the observed maximum and minimum values, rather than non-linearly, by reversing the logarithmic or logistic transformation.

All such operations will drastically affect the distributions of scores, while not at all affecting the resulting ranked output or any performance curve. Thus observed distributions might depend on, in effect, accidental characteristics of the system.

6 Conclusion

We have seen that we would normally expect the recall-fallout curve to be convex in the sense defined above. That is, if we find a system which violates this condition, then the system can be improved merely by adding a randomisation process. Therefore we would at least expect good systems to satisfy the condition already.

We have seen that under some models of the distributions of relevant and non-relevant scores. models which have been proposed and/or used by researchers, this convexity condition is violated. While the violation may relate to some part of the score range which is not normally encountered, any violation seems at least to raise questions about the general validity of the distributional model under consideration.

Specifically, the model that appears to be most frequently used at present, the normal/exponential mixture, *always* violates the convexity condition at both ends of the range of theoretically possible scores. While this result does not

invalidate the model as a reasonable approximation to the true distributions, it does put into question its general validity.

References

1. Swets, J.A.: Information retrieval systems. Science **141**(3577) (July 1963) 245–250
2. Baumgarten, C.: A probabilistic solution to collection fusion problem in distributed information retrieval. In Hearst, M., Gey, F., Tong, R., eds.: SIGIR'99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, ACM Press (1999) 246–253
3. Arampatzis, A., van Hameren, A.: The score-distributional threshold optimization for adaptive binary classification tasks. In Croft, W.B., Harper, D.J., Kraft, D.H., Zobel, J., eds.: SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, ACM Press (2001) 285–293
4. Manmatha, R., Rath, T., Feng, F.: Modelling score distributions for combining the outputs of search engines. In Croft, W.B., Harper, D.J., Kraft, D.H., Zobel, J., eds.: SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, ACM Press (2001) 267–275
5. Swets, J.A.: Effectiveness of information retrieval methods. American Documentation **20** (1969) 72–89
6. Bookstein, A.: When the most 'pertinent' document should not be retrieved – an analysis of the Swets model. Information Processing and Management **13** (1977) 377–383
7. Collins-Thompson, K., Ogilvie, P., Zhang, Y., Callan, J.: Information filtering, novelty detection and named page finding. In Voorhees, E.M., Harman, D.K., eds.: The Eleventh Text REtrieval Conference, TREC 2002. NIST Special Publication 500-251, Gaithersburg, MD: NIST (2003) 107–118
8. Robertson, S.E.: Explicit and implicit variables in information retrieval systems. Journal of the American Society for Information Science **26**(4) (1975) 214–222
9. van Rijsbergen, C.J.: Retrieval effectiveness. In Voigt, M.J., Hanneman, G.J., eds.: Progress in communication sciences. Volume 1., Ablex Publishing (1979) 91–118
10. Hawking, D., Robertson, S.: On collection size and retrieval effectiveness. Information Retrieval **6** (2003) 99–150
11. Robertson, S.E.: The probability ranking principle in information retrieval. Journal of Documentation **33** (1977) 294–304
12. Robertson, S.E.: The parametric description of retrieval tests. part 1: The basic parameters. Journal of Documentation **25**(1) (1969) 1–27