On GMAP – and other transformations

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

ABSTRACT

As an alternative to the usual Mean Average Precision, some use is currently being made of the Geometric Mean Average Precision (GMAP) as a measure of average search effectiveness across topics. GMAP is specifically used to emphasise the lower end of the average precision scale, in order to shed light on poor performance of search engines. This paper discusses the status of this measure and how it should be understood.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*information search and retrieval*

General Terms

Measurement

Keywords

evaluation, effectiveness measures

1. INTRODUCTION

In the now fairly long history of evaluating information retrieval systems, we have accumulated a significant number of different measures of effectiveness. There are favourites (such as Mean Average Precision, MAP) which are very widely used, and there is some sense it is good to use a measure with which other researchers are familiar; but there is nothing close to agreement on a common measure which everyone will use.

Some of the differences between authors have to do with constraints external to the measure itself: for example, if there are multi-level relevance judgements, and the author wishes to make use of these, s/he is constrained in the choice of measure (e.g. MAP does not take account of multiple levels). Other differences have to do with transparency (e.g. it is very obvious what Precision at rank 5, P@5, means; MAP is less clear), with stability (e.g. MAP is known to be a more stable measure than P@5), with what the measure seems to be measuring or with a task or user model (e.g. Mean Reciprocal Rank, MRR, is only concerned with the first relevant document, which may be appropriate for factual questions; MAP is concerned with all relevant documents, which may be appropriate for a literature review). There is a whole raft of different criteria which may be applied.

One particular recent concern has been with 'difficult' topics: those queries or topics on which systems seem to do badly. The TREC Robust track [11] is devoted to this question. The measurement issue here is not so much about the primary measure, but about the way it is averaged or otherwise summarised over a set of topics. The track has used different measures, but the main measure used recently, following a suggestion from the present author, is Geometric Mean Average Precision, GMAP. The geometric mean is now being used on other measures too e.g. [4]; and other similar tranformations have been proposed [2].

This paper discusses the rationale for this choice. It does not make any original proposals for measures or measurement (except for trivial variations), nor does it present any data. Its contribution is to lay out the arguments systematically and reasonably comprehensively, and to help readers think about the choice and its implications. The central argument, however, is a polemical one: that in principle, the geometric mean (or mean of logs) is just as valid and useful a form of average as the (arithmetic) mean, and should be accorded a similar status.

2. SOME BACKGROUND

2.1 Average Precision

This paper is not primarily about the Average Precision measure as defined on an individual query or topic, but what happens after that. However, it is appropriate to revisit the basic ideas behind this measure, as a reminder.

We consider just one query at a time. We assume that the output of a search system in response to a query is a ranked list of items, and also that user judgements categorise documents in binary fashion as relevant or not to the topic or information need that gave rise to the query. The effectiveness of the system is generally measured in terms of how well the system succeeds in ranking relevant documents (according to the user's judgement) at or near the top of the list. If we think in terms of the traditional measures of re-

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.

call (proportion of the relevant documents that have been retrieved) and precision (proportion of the retrieved documents that are relevant), stepping down the ranked list from the top involves stepping up the recall scale, each time we encounter a relevant document in the list. Typically, as we increase recall in this way, precision tends to drop. Plotting precision at different recall levels gives us the familiar recall-precision curve. When comparing two systems, if system A has a higher curve than system B, we regard it as more effective.

Effectiveness measures may be broadly categorised into those that relate to a single rank position or point on the recall-precision curve, and those that attempt to describe the entire curve or ranking. $P@10$, precision at rank ten, describes a single point only. Average precision is essentially a measure of the area under the recall-precision curve, therefore reflecting the entire ranking. This 'area-under-the-curve' interpretation provides some sort of intuitive interpretation of what average precision is trying to measure.

Formally, we follow the usual 'non-interpolated' definition:

$$\text{Average Precision} = \frac{1}{|R|} \sum_{r \in R} P@r \tag{1}$$

where R is the set of relevant documents, r is a single relevant document, and $P@r$ is the precision at the rank position of r. In practical terms, Average Precision is usually calculated from a ranking which is truncated at some rank position (typically 1000 in TREC, maybe much lower in a live system). In these conditions, for any r which ranks below this truncation point or not at all, $P@r$ is treated as zero. This is usually a reasonable approximation.

In any experimental evaluation of systems over many queries, there is considerable variation in the Average Precision of different queries. The usual approach is to take the (arithmetic) mean of Average Precision values over the queries. This is the measure known as Mean Average Precision. Because the per-query Average Precision is normalised to the range [0,1], irrespective of (say) the number of relevant documents in the query, the mean can be said to treat all queries equally.

Average Precision is sometimes criticised as an opaque measure which is difficult to understand, as compared to (say) $P@10$, whose meaning would be immediately apparent to a user – although as indicated above, the 'area-under-the-curve' interpretation provides some intuitive validation. As against this supposed opaqueness, it is known to be a much more stable measure than (say) $P@10$ [1], is more sensitive to differences between rankings, and tends to predict other measures well. It is also, despite being a measure on the whole curve, much more sensitive to differences at the top of the ranking than lower down – a characteristic which is absolutely necessary to ranking evaluation measures used in information retrieval. These reasons make it attractive to researchers, despite its lack of transparency.

2.2 Geometric mean as mean of logs

The definition of the geometric mean of n values is the nth root of the product of the n values:

$$\text{GM}(x_1 \dots x_n) = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

However, an alternative way to think about it which is actually better for our present purpose is that we take logs of the observations and average them. We may convert this back to the original scale by exponentiating, but this might be regarded as optional. Thus we might define the Average Log (AL):

$$\begin{aligned} \text{AL}(x_1 \dots x_n) &= \frac{1}{n} \sum_{i=1}^{n} \log x_i \\ \text{GM}(x_1 \dots x_n) &= \exp \text{AL}(x_1 \dots x_n) \end{aligned} \tag{2}$$

For this purpose the base of the logarithms is e. Note also that it is assumed here that the original values x_i are positive. Generally this will be applied only to variables that are constrained to be non-negative, but the possibility that one or more of the observed values is zero will cause problems; this is discussed further in section 4.1.

The log is a strictly monotonic transformation ($\log x_i > \log x_j \iff x_i > x_j$. Thus whatever measure the x values represent, the log preserves all its ordinal or preference qualities.

In the main application of geometric mean considered in this paper, the x_i values are the average precision values for individual queries, and the geometric mean is what is referred to in this paper as GMAP. Note that if the values x are always less than one, as in the average precision example, then the log is always negative, thus AP itself is negative. This is not a problem in a formal sense, however, it may be thought confusing. This might be taken as a reason for exponentiating back to the original scale.

Because of its monotonicity, for a single query, log(Average Precision) will display precisely the same preference order of systems as Average Precision itself. However, because of the non-linearity of the transformation, it will average differently over a set of queries. The same statements apply to any other measure of effectiveness.

2.3 Scales

There is a traditional way of classifying measurement scales which is useful in this context, due to Stevens [5]. Scales may be (in order of increasing constraint) Nominal, Ordinal, Interval or Ratio. Nominal scales have (necessarily discrete) unordered values: a set of exclusive categories is a nominal scale for the categorised objects. Ordinal scales have the additional requirement that there is a natural order on the values; an example might be a relevance scale (e.g. highly relevant, partially relevant, not relevant). Interval scales have the additional property that differences can be compared; for example, temperature is usually regarded this way (a temperature change from 10° to 15° is in some sense the same as a change from 30° to 35°). Ratio scales additionally require ratios to be comparable; many physical measurements such as length are ratio scales. Note however that ratio scales must have a meaningful zero; thus temperature (in most of the scales usually used) has an arbitrary zero, and does not have the ratio property. For example, it is meaningless to say that a temperature of 20° is 'twice as hot as' 10°.

This traditional classification is fairly simple, missing out a number of subtleties, and has been extensively criticised (e.g. [7]). Nevertheless, it does provide some useful insights. In particular, certain standard statistical operations make assumptions about the scales, which are sometimes questionable. The prime example is averaging (arithmetic

mean). Taking a mean of a set of observations of a variable implicitly assumes that the variable is on an interval scale: for example, if one observation is raised by 5 units, and another is lowered by 5 units, the mean operation cancels these changes out. Nevertheless, means are used for many variables whose adherence to the interval property is at best questionable. Similarly, the most obvious interpretation of GM would probably involve an assumption of the ratio property – which is equivalent to the interval property on the log values.

In the case of a variable which cannot be assumed to be interval, or for which the assumption is dubious, we might take different views. A measure of central tendency which is strictly suitable for ordinal-only scales is the median; this can be useful, but is somewhat limited. Alternatively, we might take means anyway, but look for ways to investigate the effects of making the interval assumption, and the sensitivity or robustness of the results to this assumption.

3. GMAP

3.1 Nonlinear monotonic transformations

If we have any reason to doubt that a difference of 5 points at one end of the scale is comparable to a difference of 5 points somewhere else, then we might investigate sensitivity/robustness by putting our values through a nonlinear but monotonic transformation before averaging them. Note that if the interval assumption is not valid for the original measure nor for any specific transformation of it, then *any* monotonic transformation of the measure is *just as good a measure* as the untransformed version. If we believe that the interval assumption is good for the original measure, that would give the arithmetic mean some validity over and above the means of transformed versions. If, however, we believe that the interval assumption might be good for one of the transformed versions, we should perhaps favour the transformed version over the original. But if there is no particular reason to believe the interval assumption for any version, then all versions are equally valid. If they differ, it is because they measure different things.

We might therefore go through this process repeatedly, with any number of different transformations. Good robustness would be indicated if the conclusions looked the same whatever transformation we used; if we found it easy to find transformations which would substantially change the conclusions, then we might infer that our conclusions are sensitive to the interval assumption, and that the different transformations measure different things in ways that may be important to us.

It will be clear now that the IR evaluation field has indeed been taking a step along this path. The transformation concerned is the log; in the Robust Track, we have discovered that some of our conclusions are indeed sensitive to the interval assumption. Specifically, for example, on the original MAP, we can get significant gains through blind feedback; on the transformed measure (or equivalently on GMAP), blind feedback is often found to be detrimental. Depending on our view of the relative importance of differences in different parts of the scale, a GMAP-based conclusion that blind feedback is detrimental is *just as valid* as a MAP-based conclusion that blind feedback is beneficial. (See the next section for a further discussion on this point.)

The conclusion must be that MAP and GMAP measure somewhat different things. Each reveals different aspects of system effectiveness, in much the same way that MAP and P@5 reveal different aspects. The difference between MAP and P@5 has to do mainly with emphasis on the different parts of the recall-precision curve; the difference between MAP and GMAP has to do solely with emphasis on different parts of the effectiveness distribution over topics.

In section 4.2 below, we discuss a different transformation that may be useful for some measures.

3.2 Example

To illustrate the differences between the arithmetic and the geometric mean, we generate some random data: 50 data points between zero and one, with deliberately high variance but biased towards the low end. They may be thought of as average precision observations for 50 topics. A histogram of the distribution of data points used is shown in Figure 1. Note also that there are three zeros in the set, and also three observations of 0.01.

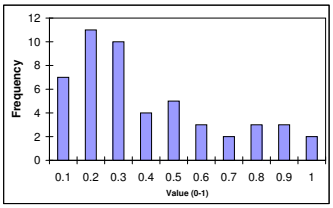

Figure 1: Distribution of values in the example.

The arithmetic mean of these values (MAP) is 0.309; the geometric mean (GMAP) is 0.117 (see section 4.1 below for how we deal with the zeros). Note that we expect GMAP to be less than MAP.

We indicate the way in which GMAP is sensitive to the smaller observations as follows: we boost each value by 0.05. For obvious reasons, MAP increases by the same amount, to 0.359; GMAP however increases much more, to 0.263. The reason for this is that the boost of 0.05 has proportionately more effect on the very small observations (zero increases to 0.05 and 0.01 to 0.06). If we split the observations into the lowest and highest 25, and apply the boost to one set only, the effect on MAP is the same for both sets: an increase of 0.025. On the other hand, GMAP increases to 0.249 when we boost the lowest 25, but only to 0.123 when we boost the highest 25.

3.3 The Robust track and difficult topics

The emphasis of the Robust Track at TREC can be described in two slightly different ways: either it is primarily about a class of topics which might be termed 'difficult', or it is about the distribution of effectiveness over topics, and in particular the lower end of the effectiveness scale. In part it has revolved around identifying difficult topics and biasing measurement towards those topics; in other words, regarding (for this purpose) some topics as more important than others. Initially, the methods of measurement were designed explicitly with this in mind [9].

The shift to GMAP actually changes the emphasis very slightly. If we think of GMAP as the average of logs (whether

or not we transform it back to the AP scale after averaging), then it becomes clear that there is no emphasis on particular topics. The average of logs (of AP) is a straight arithmetic mean over topics; each topic has exactly the same influence, there is no weighting based on topics. The emphasis is instead based on the scale of the effectiveness measure. That is, GMAP treats a change in AP from 0.05 to 0.1 as having the same value as a change from 0.25 to 0.5. MAP would equate the former with a change from 0.25 to 0.3, regarding a change from 0.25 to 0.5 as *five times* larger. These two statements effectively exemplify the interval assumptions applied to GMAP and MAP respectively. In the opinion of the present author, there is no reason to believe that one of these is correct and the other incorrect, or even that one is better than the other.

The distinction between emphasising different topics and emphasising different parts of the scale may seem slight, but is important. It reinforces the point that a conclusion based on GMAP has exactly the same status as a conclusion based on MAP. In both cases we are giving equal weight to all topics in the test set. We might for external reasons prefer one to the other (one possible reason is discussed in section 4.4 below), but there is little intrinstic difference.

3.4 The interval assumption revisited

There is one argument that might be used about the validity of the interval assumption, by reversing the example above. If we consider a change in AP for one topic of 0.5 to 0.25, what would be a corresponding change for a topic with a starting AP of 0.1? Using untransformed AP the change would have to be from 0.1 to -0.15 – but this of course is impossible. In this sense, we might argue that the interval assumption applied to a measure which is constrained to be positive does not make sense. On the other hand, this example works perfectly sensibly in the log AP scale.

However, the log AP scale does not deal with a similar problem at the other end: AP is also constrained to a maximum of 1. In order to deal with the problems at both ends, we could use another transformation, the logit, as discussed in section 4.2 below. This argument seems to suggest not that all transformations have equivalent validity, but rather that the logit transformation actually produces a measure in which the arithmetic mean operation has *more* validity than it does in the original scale.

4. SOME OTHER CONSIDERATIONS

4.1 Zeros

The log transformation maps from the strictly positive reals to the entire real line. It is not defined for zero or negative numbers. As implied above, it is usually applied to quantities which are constrained in some way to be positive. If the variable is a proportion, or perhaps a probability estimated by a proportion, then it can in principle and may in practice sometimes take the value zero. In the case of a probability, we might reasonably assume that the true value is never zero; nevertheless, it may be that in the sample from which we are estimating, zero can arise.

Similarly Average Precision as usually calculated may be estimated as zero for a few individual topics. In principle Average Precision defined over a ranking of the entire collection cannot be zero. That is, the measure is undefined if there are no relevant documents; if there are relevant documents, and we rank the entire collection, then at the rank position of each relevant document there must be a non-zero precision. Nevertheless, as indicated above, the usual way of calculating Average Precision is to truncate the ranking at some point. A relevant document which is not retrieved before the truncation point is assumed to contribute zero precision to the calculation. This means that a query which fails to retrieve any relevant document at all before the truncation point will be assigned zero average precision. Thus we need to deal with this situation in practice.

A simple pragmatic solution to this problem is to add a small quantity to the estimate before taking the log, and removing it again afterwards, thus:

$$\begin{aligned} \underset{\epsilon}{\mathrm{AL}}(x_1 \dots x_n) &= \frac{1}{n}\sum_{i=1}^{n} \log(x_i + \epsilon) \qquad (3) \\ \underset{\epsilon}{\mathrm{GM}}(x_1 \dots x_n) &= \exp \mathrm{AL}(x_1 \dots x_n) - \epsilon \end{aligned}$$

where ϵ is an arbitrary small number.

This is a little untidy, because it requires the specification of ϵ and allows comparisons only in the case where the same value has been used. One could perhaps discover a rationale or principle for the choice of ϵ – possibly on the basis of estimation arguments. One possibility for the Average Precision case is to change the rule about assigning zero to unretrieved relevant documents, and instead pretend that they are all ranked at the very end of the entire collection. Thus each would have a small but non-zero precision value of (counting backwards from the last) $\frac{|R|}{N}$, $\frac{|R|-1}{N-1}$, $\frac{|R|-2}{N-2}\dots$, where N is the size of the collection. This gives a deterministic lower bound Average Precision for a full-collection ranking. A variant would be to treat all documents below the truncation point (the entire remaining collection) as randomly ranked. This would give a slightly higher, stochastic lower bound.

However, these are probably unnecessary complications. If we were to observe that many queries had zero Average Precision under the usual method, we would probably want to do some very different failure analysis. If we are simply dealing with a few instances, the ϵ method is probably quite adequate. In the Robust track at TREC 2005, a variant on the ϵ method, still involving an arbitrary small number, was used [11]. The individual topic AP values were not modified unless zero; every zero value was reset to 0.00001.

In the example above, we used the simple addition method above with $\epsilon = 10^{-5}$; as indicated, this gives GMAP= 0.117 If we reduce ϵ to 10^{-6}, GMAP falls to 0.102; if we increase ϵ to 10^{-4}, GMAP rises to 0.134. Thus GMAP is somewhat sensitive to the chosen ϵ; it will also affect a little the relative sensitivity of GMAP to different small values.

4.2 Other transformations

Clearly the log transformation is not the only one that may be useful. The log is intended to deal with the situation where, for the effectiveness measure concerned, the region close to zero is critical. For probability or proportion measures, it is sometimes the case that the region close to 1 is also critical, and always the case that the measure is constrained to a maximum of 1. Thus for example in spam filtering, a success rate of 99.99% is really very different from a success rate of 99.9%. Most measures commonly used in IR do not normally approach so close to 1; however, the spam example indicates that there are exceptions.

A transformation which deals with both sensitivities simultaneously is the logistic or logit transformation (in the case of a probability, also known as log-odds):

$$\text{logit}\, p = \log \frac{p}{1-p}$$

This transformation is, like the log, strictly monotonic. It has been used in the Spam Track at TREC [2]; the reason for its use here is related to the arguments of this paper. It is not to do with averaging across queries, but with combining two different measures of performance, each of which is an error probability, into a single composite measure. Recently, Cormack [3] has shown that under some sampling assumptions, estimation errors in logit(Average Precision) approximately fit a Gaussian distribution; the fit is better than estimation errors for Average Precision itself.

As indicated above, it is at least arguable that the appropriate transformation of a proportion or probability (which is constrained to lie between zero and one) is the logit or log-odds, because this gives it a $(-\infty, +\infty)$ range, and therefore allows the application of the interval assumption (as well as the use of linear models) without built-in contradictions at the ends of the scale. Cormack's results might be taken to support the same conclusion. All the qualitative arguments in this paper apply just as well to the logit as to the log.

For this transformation, the zeros problem also potentially occurs at both ends: the logit is undefined if either $p = 0$ or $p = 1$. If the estimation process could yield such results, then some similar solution to the ϵ above is required to deal with both ends, for example by adding ϵ to both numerator and denominator in the definition of logit.

4.3 Application to other measures

As in the previous case, transformations such as the log or logit might potentially be usefully applied to other measures than average precision. An obvious example is NDCG, normalised discounted cumulative gain, for multi-level relevance judgements. This has similar distributional characteristics to MAP: that is, for a significant number of queries, it might take a value very close to zero. If we wish to examine that end of the distribution of effectiveness over queries, then it makes sense to consider geometric mean NDCG, as opposed to the usual arithmetic mean.

4.4 Stability

As indicated at the beginning, one consideration in the choice of measure may be its stability or robustness under various conditions – for example, its statistical reliability when measured on limited numbers of topics [8]. There is evidence that MAP is rather more stable in this respect than many other measures. Such stability would not necessarily transfer directly to transformed versions of AP such as GMAP.

Analysis done for the TREC Robust Track indicates that GMAP, while not as stable as MAP, is still reasonably stable, more so than the measures previously used in the track [10].

4.5 Variety

To some extent, IR researchers are used to dealing with multiple measures. In the opinion of the present author, a significant contribution to the success of the TREC programme is the variety of measures in use. Many of the tracks report a whole range of different effectiveness measures, which may or may not be highly correlated; there may be a single specified main measure, but others may reflect different features. All this is good, both as a sanity check on the results and because different measures may show different things.

On the other hand, researchers would often prefer the simplicity of a single measure; those who are not interested in issues of measurement *per se* may find the variety confusing and hard to deal with. More particularly, in recent years, the problem of optimisation (in relation to a set of free parameters of a ranking function, say) has encouraged the choice of a single measure; it is hard to optimise more than one measure at once, particularly in the context of an automatic, iterative procedure.

Even in this case, however, variety may be good. There are many open issues associated with choice of measures for optimisation [6]. For one, it is not clear that the measure that you really want to optimise in the test set should be the one that you optimise in the training set. In this sense, having a variety of measures and understanding the characteristics of each and how they relate may be helpful.

5. CONCLUSIONS

MAP and GMAP may be seen as similar measures of average ranking effectiveness of a system across a test set of topics. They differ not in how much attention they pay to different topics in the set, but in how much attention they pay to different parts of the Average Precision scale. Given that a strong assumption about the interval nature of the AP scale is not justified, the two average measures might reasonably be regarded as having equal validity. Experimental results suggest that they measure somewhat different things.

They are, in effect, different members of the menagerie of measures developed over the last half-century for retrieval system effectiveness. The size of this menagerie may be problematic for researchers, but has some distinct advantages.

More generally, given any measure of search or ranking effectiveness, there are good reasons to consider also monotonic transformations of such a measure, which might place different emphasis on than the original on different parts of the scale. The two transformations considered in this paper, the log and the logit, have been shown to reveal interesting properties not apparent when we consider only the original measures to which they were applied. Thus they may be said to contribute to our understanding of ranking effectiveness.

6. REFERENCES

[1] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In N. J. Belkin, P. Ingwersen, and M.-K. Leong, editors, *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, New York, 2000. ACM Press.

[2] G. Cormack and T. Lynam. TREC 2005 spam track overview. In E. M. Voorhees and L. P. Buckland, editors, *The Fourteenth Text REtrieval Conference, TREC 2005*. Gaithersburg, MD: NIST, 2006.

[3] G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 533–540, New York, 2006. ACM Press.

[4] T. Sakai. Evaluating evaluation measures based on the bootstrap. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page forthcoming, New York, 2006. ACM Press.

[5] S. S. Stevens. On the theory of scales of evidence. *Science*, 103:677–680, 1946.

[6] M. Taylor, H. Zaragoza, N. Craswell, and S. Robertson. Optimisation methods for ranking functions with multiple parameters. Submitted for publication, 2006.

[7] P. F. Velleman and L. Wilkinson. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47:65–72, 1993.

[8] E. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. Jarvelin, editors, *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, New York, 2002. ACM Press.

[9] E. M. Voorhees. Overview of the TREC 2003 robust retrieval track. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text REtrieval Conference, TREC 2003*, NIST Special Publication 500-255, pages 69–77. Gaithersburg, MD: NIST, 2004.

[10] E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In E. M. Voorhees and L. P. Buckland, editors, *The Thirteenth Text REtrieval Conference, TREC 2004*, NIST Special Publication 500-261, pages 70–79. Gaithersburg, MD: NIST, 2005.

[11] E. M. Voorhees. Overview of the TREC 2005 robust retrieval track. In E. M. Voorhees and L. P. Buckland, editors, *The Fourteenth Text REtrieval Conference, TREC 2005*. Gaithersburg, MD: NIST, 2006.