Rank-Biased Precision for Measurement of Retrieval Effectiveness

ALISTAIR MOFFAT
The University of Melbourne
and
JUSTIN ZOBEL
RMIT University and NICTA Victoria Research Laboratory

A range of methods for measuring the effectiveness of information retrieval systems has been proposed. These are typically intended to provide a quantitative single-value summary of a document ranking relative to a query. However, many of these measures have failings. For example, recall is not well founded as a measure of satisfaction, since the user of an actual system cannot judge recall. Average precision is derived from recall, and suffers from the same problem. In addition, average precision lacks key stability properties that are needed for robust experiments. In this article, we introduce a new effectiveness metric, *rank-biased precision*, that avoids these problems. Rank-biased precision is derived from a simple model of user behavior, is robust if answer rankings are extended to greater depths, and allows accurate quantification of experimental uncertainty, even when only partial relevance judgments are available.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models, search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms: Experimentation, Measurement, Human Factors

Additional Key Words and Phrases: Recall, precision, average precision, relevance, pooling

ACM Reference Format:

Moffat, A. and Zobel, J. 2008. Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inform. Syst. 27, 1, Article 2 (December 2008), 27 pages. DOI = 10.1145/1416950.1416952 http://doi.acm.org/10.1145/1416950.1416952

This work was supported by the Australian Research Council.

Authors' addresses: A. Moffat, Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia; email: alistair@csse.unimelb.edu.au; J. Zobel, Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia; email: jz@csse.unimelb.edu.au.

1. INTRODUCTION

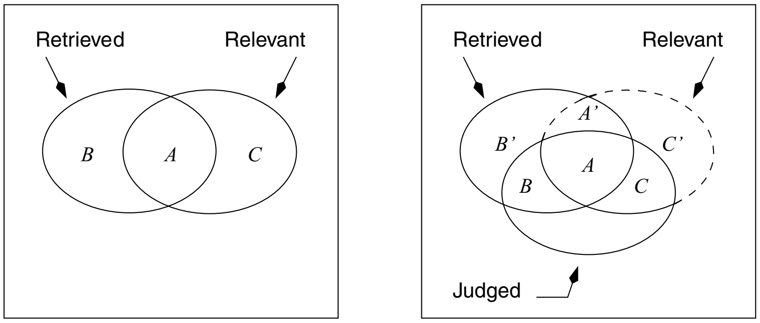
Information retrieval systems compute, for each document in a collection, a score that estimates the similarity between that document and a query. In typical systems, each score represents an estimated probability that the document is relevant to the information need expressed by the query. Once the process of scoring is complete, documents are presented to the user in decreasing score order, in the expectation that the user considers them in sequence until their information need has been satisfied.

A range of methods for measuring the effectiveness of information retrieval systems has been proposed. A key element of many of these measures—and certainly of those in the widest use—is the assumption of binary relevance, with human assessors asked to determine, for a set of documents, which members are relevant to the query and which are not. Given a ranking, each document is marked as relevant or irrelevant (or unjudged), and the sequence of decisions is then used as input to a quantitative measure of effectiveness. Two elementary measures are *recall* and *precision* [van Rijsbergen 1979, Chapter 7]. These can be combined to give a single value via mechanisms such as a 3-point or 11-point recall-precision average [Buckley and Voorhees 2005].

One of the most commonly used measures in recent IR research is *average precision* (AP), which does not directly use recall, but does require knowledge of R, the total number of relevant documents for the query in question. Other widely used measures are *precision at d documents retrieved* (P@d), where typically d is 10; *R-precision*, or P@R; and *reciprocal rank* (RR). However, all of these measures have failings. For example, it is not clear what user behavior is modeled by AP, and it has properties that render it volatile in typical experimental settings. In particular, using AP with incomplete relevance judgments typically leads to inflated effectiveness estimates, and the discovery of further relevant documents in a ranking usually reduces measured effectiveness. These issues are not addressed by recent AP-based metrics such as those of Buckley and Voorhees [2004] or Sakai [2004].

Underlying these issues are two problems with recall. One, which is widely known, is that in current systems complete relevance judgments are impractical and thus recall tends to be overestimated. Figure 1 shows this problem using the standard Venn diagram approach. The TREC methodology is discussed in more detail below. Another problem with recall, which has not received such wide attention, is that it does not correspond to a likely model of user behavior.

In this article, we introduce a new metric, *rank-biased precision* (RBP), that avoids many of the failings of average precision. The basis of RBP is that it measures the rate at which utility is gained by a user working at a given degree of persistence; by adjusting persistence, a parameter that represents an aspect of user behavior, RBP has the advantage of capturing the critical facets of AP, RR, and P@d. An additional benefit of RBP compared to AP is that it allows accurate quantification of experimental errors when only partial relevance judgments are available, which is useful when large-scale experiments are being carried out. Rank-biased precision also has some similarities with the



(a) The Cranfield arrangement

(b) The TREC arrangement

Fig. 1. Recall and precision: (a) the Cranfield arrangement, with all documents in the collection categorized for relevance, and in which precision is $|A|/(|A| + |B|)$, and recall is $|A|/(|A| + |C|)$; and (b) the TREC arrangement, in which only a subset of the documents are categorized for relevance, and the size of the sets A' and C' is not known, but for calculation purposes it is assumed that $|A'| = |C'| = 0$ and that $|A| + |B| + |B'| = d$, the number of documents retrieved.

discounted cumulative gain (DCG) metric of Järvelin and Kekäläinen [2002], and is compared to that measure below.

To understand some of the issues in existing measures, we first consider recall and precision and the way in which relevance judgments are collected. We next review the measures in common use in experimental work, as well as other measures that have been proposed in the literature. We then describe rank-biased precision, and experimentally examine its behavior.

2. RECALL AND PRECISION

Recall and precision have been in use for more than four decades [van Rijsbergen 1979, Chapter 7]. Recall is the proportion of the relevant documents that have been retrieved, while precision is the proportion of retrieved documents that are relevant. In the context of a ranked list of documents, recall and precision can be measured at each document retrieved, at each relevant document retrieved, or at recall percentiles in the ranking. Alternatively, if a single value is required, recall and precision can be measured at any fixed point in the ranking. Recall and precision tend to be in tension; when recall is high, precision is usually low, and vice versa. Textbooks describe these two metrics in detail, and the associated concept of *recall-precision curves*; see, for example, van Rijsbergen [1979, Chapter 7] and Witten et al. [1999, Chapter 4].

Relevance is a human concept that requires human judgment, and a well-known problem with current experiments is that exhaustive relevance judgments stop being practical once collections exceed a few tens of megabytes. "Trying to get an indication of which proportion of the existing relevant information items was retrieved by a system . . . is a hopeless undertaking" [Frei and Schäuble 1991, page 154]. Practical experiments such as TREC[1] make use of an approximation to the true number of relevant documents.

[1] See `trec.nist.gov`.

Participating TREC systems generate query rankings containing 1000 documents, from which a set of relevant documents is identified via a pooling approach [Harman 1995; Zobel 1998], with (typically) the top 100 documents from each system for each query assessed for relevance. That is, in TREC the number R of relevant documents in a collection for some query is deemed to be the number of relevant documents that appear in the top 100 in the ranking of some participating system when executing that query. Figure 1(b) shows the TREC arrangement of documents retrieved in a query's result list by some system; documents that have been judged, after a pool is derived based on prefix rankings contributed by multiple systems; and documents that are relevant for that query. For a particular query and system, all eight zones in Figure 1(b) can be nonempty.

A key assumption of the pooling approach is that, by having a large enough number of participants, the great majority of the relevant documents are identified and thus, for each query, that a good first-order approximation to the number R of documents relevant to that query can be determined. Postexperiment analysis has indeed shown that, for some queries, it is highly likely that all the relevant documents have been discovered, but that, for others, there are clearly many more that have not been identified [Zobel 1998]. For these, and other reasons, Saracevic [1995] referred to recall as a "metaphysical measure: how does one know what is missed when one does not know that it is missed?"

What is arguably a more profound problem relates to a very simple question: what is it that recall is measuring? The purpose of a metric is to evaluate whether a system is successful completing some task, or rather, a computable abstraction of a task. We measure the CPU time consumed by an algorithm because we claim that, to be an interesting property that has consequences for how the algorithm will be used, we don't usually measure (say) the number of bytes in the compiled instruction sequence, because that tells us nothing of great interest. In IR, we need "criteria representing the objectives of the system," and that they should be concerned with the question of "how to provide a prospective user with useful information" [Saracevic 1995]. For example, measuring precision to depth d is valid because a user who is given six answers in the top 10 documents is probably better off than a user who is given three. That is, precision can be interpreted as a measure of *user satisfaction* when the user's actions are modeled in a certain plausible way.

It might similarly be argued that a user who is given all the answers to a query is better off than a user who is given only some, and thus that calculation of recall is of value. However, in our view this line of reasoning does not stand scrutiny, because unless a user has "perfect" knowledge of the documents included in some retrieval system, they cannot know that they have seen all the answers (or half the answers, or 27% of the answers). As an extreme example, a system comparison based on recall asserts that, if a collection has only one answer, the user is 100% satisfied once they have seen it, but, if there is another relevant document that they neither know about nor view, they are only half satisfied. Many authors have also written of similar concerns. For example, Buckley and Voorhees [2005, page 61] indicated that "one of the current debates in IR is whether recall is important outside a few specific applications such as patent searching." Also worth noting is that, in these high-recall applications, users typically pose multiple queries, and seek to address their information need in a variety of ways, including via browsing and citation following. In such cases, the recall arising from the whole sequence of interactions with the collection is what is of interest, regardless of how satisfied (or not) the user feels after each individual search. That is, while missing a legal precedent or a medical experiment with every query in an information-seeking session might be disastrous, the recall of one query considered in isolation is not particularly informative.

Another perspective on this issue is that, crudely, queries can be said to be of two kinds: "find a document" and "find a lot of documents." For the first kind, any relevant document is satisfactory, and the query is resolved when the answer is found (and perhaps confirmed), no matter how many other relevant documents there are—consider the query "boiling point of lead," for example. In this kind of search the concept of recall is not applicable. For the second kind of query, for large and uncurated collections such as the Web, the user almost certainly does not know whether all relevant documents have been found until they have phrased a variety of queries and undertaken a multipronged exploration of the data, meaning that recall is unrelated to their satisfaction level in regard to any single query.

In a particularly pertinent comment, Cooper [1973, page 95] wrote the following:

> The involvement of unexamined documents in a performance formula has long been taken for granted as a perfectly natural thing, but if one stops to ponder the situation, it begins to appear most peculiar. . . . Surely a document which the system user has not been shown in any form, to which he has not devoted the slightest particle of time or attention during his use of the system output, and of whose very existence he is unaware, does that user neither harm nor good in his search.

On the next page, Cooper [1973] went on to say the following:

> Instead of attempting to estimate recall in spite of all the difficulties, what should have been done was to find a way to overcome the deficiencies of the precision measure without bringing a second measure into the picture.

Although written well over 30 years ago, Cooper's words remain applicable today.

Precision is a rather more straightforward metric than recall. It clearly does provide a measure of user satisfaction, particularly when evaluated at a query-independent value such as $d = 10$ or $d = 20$. On the other hand, if precision is evaluated at a single point that is in some way determined by a numeric recall level, then the criticisms above still apply.

Su [1994] observed and interviewed 40 users who had employed professional intermediaries to assist with Boolean querying in online library services.

Su compared each user's subjective judgment of overall search success against a range of other measures, including precision, search time, search cost, the user's satisfaction with the completeness of search results, and the user's confidence in completeness of the search results. Su found that precision was not a strong indicator of subjective success and that the two "completeness" indicators were more closely aligned with overall success, concluding that "recall is more important than precision to users" [Su 1994, page 213]. However, Su was not able to directly assess recall, because, as she observed in her article, "recall . . . requires the knowledge of all the relevant documents in the databases(s) in relation to the users' needs or problems" [Su 1994, page 208]. Moreover, the experimental environment was of academic searchers who were given an unranked set of Boolean-matching documents with considerable effort going into the construction of each query—quite different in terms of both approach and audience to typical current search environments.

Other issues in retrieval experiments include the reliability of relevance assessment and the validity of binary relevance judgments; see, for example, Harter [1996], Mizzaro [1997], Borlund and Ingwersen [1998], Järvelin and Kekäläinen [2002], and Kekäläinen [2005]. These issues are significant, but they are separate from the topic of this article, and binary relevance judgments do provide a basis for assessing the utility of retrieval systems [Allan et al. 2005]. Likewise, there are many different ways of evaluating an IR system; see, for example, Kagolovsky and Moehr [2003]. These too are beyond the scope of this article. More broadly, there are many separate elements and decisions that must be considered in the design of a robust retrieval experiments [Tague-Sutcliffe 1992]; our focus here is on the rather narrower problem of providing a score that represents retrieval effectiveness once a document ranking has been obtained.

3. COMPOSITE MEASURES OF EFFECTIVENESS

To produce a statistic describing a retrieval system, many ways of combining recall and precision into a single number have been described. This section reviews some of those methods.

3.1 Average Precision

The measure of effectiveness most commonly used in experiments in recent years has been average precision, AP. There has been little discussion of AP in the literature; Buckley and Voorhees [2005] explained that AP, also known as *noninterpolated average precision*, was introduced after the first year of TREC to address deficiencies in previous measures. Like its predecessors such as 11-point average precision, AP combines recall and precision to give a single-value measure of a system.

Average precision is calculated by taking the set of ranks at which the relevant documents occur, calculating the precision at those depths in the ranking, and then averaging the set of precision values so obtained. For example, consider the ranking

`$$---$----$-----$---`,

in which (reading left to right) there are five relevant documents, indicated by the $ characters, interspersed among a further 15 nonrelevant documents. If it is assumed that there are a total of $R = 5$ documents that are relevant, then AP at depth $d = 20$ is calculated as

$$\frac{1}{5} \times \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{6} + \frac{4}{11} + \frac{5}{17} \right) = 0.6315.$$

If instead there were $R = 6$ relevant documents, the ranking would need to be extended until that sixth document was located. If that was not possible—for example, because it was a TREC-style pooled experiment and the ranking had already been computed before the number of relevant documents was known—the standard assumption is that the missing relevant documents have corresponding precision scores of zero. In this case the leading factor on the computation of AP would need to be 1/6, and the calculated value would be 0.5263. If there were $R = 7$ relevant documents, the AP score would be 0.4511, and so on. The average precision across a series of queries can be averaged to give *mean average precision* (usually referred to as *MAP*).

Formally, if we have a ranked relevance vector to depth d

$$\mathcal{R} = \langle r_i \mid i = 1, 2, \ldots, d \rangle,$$

where r_i indicates the relevance of the ith ranked document scored as either 0 (not relevant) or 1 (relevant), and if R is the number of relevant documents for this query, then AP is computed as

$$\mathrm{AP} = \frac{1}{R} \sum_{i=1}^{d} \left(\frac{r_i}{i} \cdot \sum_{j=1}^{i} r_j \right).$$

A consequence of this definition is that a system whose recall at the end of the ranking (in TREC experiments, at depth $d = 1000$) is x can thus at best hope to attain an AP of x. That is, recall bounds AP.

An alternative unnormalized formulation—useful if systems are to be compared on just a single query—is to remove the division by R, and simply sum the set of precision values:

$$\mathrm{SP} = \sum_{i=1}^{d} \left(\frac{r_i}{i} \cdot \sum_{j=1}^{i} r_j \right).$$

However, if more than one query is involved, this variation introduces serious scaling problems—it makes little sense to compute a mean unnormalized average precision when, taking two of the TREC-5 queries as a concrete example, there are (at least) 433 documents relevant to “volcanic and seismic activity levels,” and (possibly only) seven relevant to “DNA information about human ancestry.” In recent work, Webber et al. [2008] explored this point, and described a standardization approach that removes the bias caused by query variation. Their “standardized SP” metric is identical to “standardized AP,” but does not require knowledge of R, and has a range of other desirable properties.

Another tempting AP variant is to calculate the average over the relevant documents that are actually retrieved within the answer ranking through to depth d:

$$\mathrm{AP}^{*} = \frac{1}{\sum_{i=1}^{d} r_i} \sum_{i=1}^{d} \left(\frac{r_i}{i} \cdot \sum_{j=1}^{i} r_j \right).$$

But this leads to anomalous situations in which a ranking with an increased number of relevant documents can have a lower AP* score. For example, the ranking

`$---$----$-----$---`

was used above and gives rise to an AP* of 0.6315. Now consider the ranking

`$$---$----$-----$$$$`,

in which eight documents are relevant. One would intuitively expect average precision to be greater, since certainly P@20 has increased. In fact, the altered ranking has an AP* of 0.5324, with the decrease caused by the inclusion of three additional terms, each of which is smaller than the previous average.

A related problem is that AP is unstable in the presence of uncertainty. Consider the ranking

`$--$------??????????`,

in which question marks represent unjudged documents, and there are $R = 2$ relevant documents within the set of judged documents. Given these facts, AP is computed to be $(1/2) \cdot (1.0 + 0.5) = 0.75$. However, if any one of the 10 unjudged documents is in fact relevant, then the AP cannot be greater than 0.5909. Nor is the movement consistent: on the ranking

`-------$--??????????`

discovery of further relevant documents among the unjudged ones can cause AP to increase rather than decrease. This is a serious concern—on addition of more information AP can take any value at all between the limiting values of 0 and 1, irrespective of what mix of relevant and irrelevant documents appears in any finite prefix of the ranking. There is thus a perceived risk that the scores obtained in a retrieval experiment are a function as much of the resources spent carrying out the evaluation and judgments as they are of the system itself, and that "preliminary" results should be treated with caution.

The AP drift caused by unjudged documents is not merely an academic concern. Consider the data gathered by TREC in 1996 (TREC-5), a year in which 61 systems contributed to the Ad Hoc Retrieval Track, and processed 50 queries against 2 GB of newswire data.[2] The depth used for forming the pool of judgments was 100, and based on those judgments it is straightforward to also investigate the alternative outcomes that would have been observed were the judgments to be compiled over shallower pools. The performance of the 61 systems is plotted in Figure 2, comparing the calculated mean AP scores with

[2] See `trec.nist.gov` for details.

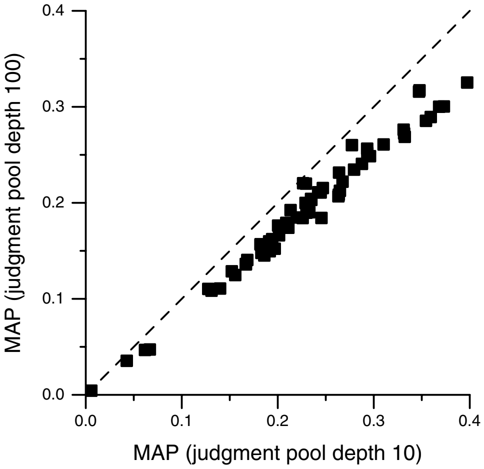


Fig. 2. Mean average precision of 61 TREC-5 systems, using relevance judgments compiled using two different pool depths. The dotted line is the identity relationship, with points below the line showing systems for which average precision decreased when additional documents were judged. The nonlinearity of the decrease shows that the ordering of systems is also affected.

pool depths of 10 and 100. Note how AP for an assessment pool of depth 10 is almost always an overestimate for the "correct" AP when calculated using an assessment pool of depth 100. Note also that the ordering of the systems changes as the pool depth is increased. We can only conclude that, were the pools to be extended to depth (say) 1000, further decreases in mean AP would be observed, and that there would be additional perturbations in the system ordering.

In addition to these relatively technical issues, average precision, like recall, is on uncertain foundations. Average precision can be said to represent an estimate of user satisfaction, but based on a complex abstraction that does not fit well with our usual understanding of how users interact with a retrieval system. Consider the necessary scenario: the user issues a query, obtains a ranked list of answers, and begins examining them. Every time a relevant document is encountered, the user pauses, asks "Over the documents I have seen so far, on average how satisfied am I?" and writes a number on a piece of paper. Finally, when the user has examined every document in the collection—because this is the only way to be sure that all of the relevant ones have been seen—the user computes the average of the values they have written.

Buckley and Voorhees [2005, page 59] also criticized AP, on the grounds that it "is an overall system evaluation measure, not an application measure," and that "there is no single user application that directly motivates MAP." We agree with this criticism, and posit that, in the absence of any task to which the measurements correspond, abstract measurements of a system are less interesting than those that are predicated on a plausible model of user behavior.

Average precision does have strengths. Perhaps the best evidence in its favor is its stability and robustness: AP-based differences between systems on one set of queries tend to be observed on other query sets, especially if the differences are statistically significant [Voorhees 2002; Sanderson and Zobel 2005; Buckley and Voorhees 2005].

3.2 Precision-at-Depth, R-Precision, and Reciprocal Rank

Several other measures are regularly employed by researchers. One of these is precision-at-depth, or P@d. This measure is free of several of the failings of AP, but it has the drawback of being insensitive to the rank positions of the relevant documents—the rankings "$$$$$-----" and "-----$$$$$" are identical in terms of precision at depth 10, but the first is almost certainly a better ranking than the second. A second problem with P@d is that interpretation of precision also needs to be tempered to a certain extent by knowledge of R [Buckley and Voorhees 2005]. In particular, when the number R of relevant documents in the collection is less than d, the number of documents retrieved, precision is restricted to $R/d < 1$. That is, while precision can be calculated at any depth d and knowledge of R is not required to do so, at depths $d \geq R$ precision is hobbled and cannot fully range over the interval 0.0 to 1.0. Because of this issue, P@10 tends to be a more reliable measurement than P@100, and P@1000 is of little interest.

The real issue is that, to compute P@d, a value of d must be selected, and because of the relationship between d and R it is hard to argue for any particular value of d. One way fixing d is to use R*-precision*, sometimes known as *missed-at equivalence*—the precision score at depth $d = R$. Scores are guaranteed to be able to fully range from 0.0 to 1.0, but this metric again requires that R, the number of relevant documents, be known for each query. And, as is the case for average precision, R-precision presents anomalies, such as that an increase in both R and the number of documents returned may lead to a reduction in measured effectiveness. In terms of user behavior, it seems implausible to suppose that a user would choose in advance to inspect exactly R documents for a given query, even if they could somehow be aware of what R was for their query.

Finally, another common measure is the *reciprocal rank* (RR) of the first relevant document. Reciprocal rank has the singular advantage of being completely independent of R, since only one relevant document needs to be located; other relevant documents are not considered at all. Reciprocal rank also has an attractive user model—a person who is only interested in the first relevant answer. But, because of the fact that the score for the first position is double that of the score for the second position, RR is not particularly stable when systems are being compared via an average, because good performance on a single query can compensate for poor performance on many others.

There are many variants of these simple measures, such as interpolated average precision, and 11-point and 3-point average precision. All of these metrics tend to correlate with each other, average precision and R-precision particularly so [Buckley and Voorhees 2005]. Nevertheless, we conclude that the standard methods for comparing document rankings have shortcomings that make them either difficult to interpret, or deficient in some other way. One simple problem is that these composite methods tend to be undefined if there are no relevant documents in the ranking. In the absence of any answers, effectiveness can easily be defined to be zero; nevertheless, it is irksome that an exception is required.

3.3 Other Measures of Effectiveness

In the three decades prior to the commencement of TREC in 1992, a variety of system evaluation measures had been in use. The TREC project not only introduced standardized large-scale test collections, but also standardized evaluation of retrieval systems, embodied in the TREC_EVAL evaluation software.[3]

In particular, the pre-1990 literature contains descriptions of several measures that since then have been largely neglected, or have been superseded by the measures in TREC_EVAL. Some of these were examined by van Rijsbergen [1979, Chapter 7], who noted that they are for the most part ad hoc in nature and "cannot be justified in any rational way" (p. 119). van Rijsbergen [1979] examined in detail several methods with a mathematical foundation, including measures and observations due to Swets, Brookes, Robertson, Teather, and Cooper, and the measures incorporated into the SMART retrieval system. All of these measures yield statistics that balance precision and recall.

A motivation in design of some of these measures is to explicitly weight for "the relative importance a user attaches to precision and recall" Shaw, Jr. [1986, 346], leading to measures based on combinations such as

$$\frac{1}{1/p + 1/r - 1},$$

where p and r are point measures of precision and recall at some depth d in the ranking [Shaw, Jr. 1986; Losee 2000]. We do not survey this early literature, but note that the importance of recall appears to be a near-universal assumption. In terms of practical experimentation, Keen [1992] described experience gleaned from work with the Cranfield test collection; and Cooper [1968] and Raghavan et al. [1989] defined a quantity they called the *expected search length*, being the expected number of documents retrieved before i relevant ones have been determined, including proper handling of cases in which there are ties in the ranking.

There have also been recent proposals that are of relevance to our work. One is the binary preference measure (BPref) proposed by Buckley and Voorhees [2004]. In this approach, the binary vector $\mathcal{R}$ of relevance values is modified to give a condensed vector $\mathcal{R}'$ of length d' by removing the documents for which there are no relevance judgments, and then computing

$$\text{BPref-}k = \frac{1}{R} \cdot \sum_{i=1}^{d'} \left\{ r_i' \cdot \left(1 - \frac{\min\left(R + k, i - \sum_{j=1}^{i} r_j'\right)}{\min(R + k, N)} \right) \right\},$$

where N is the number of documents known to be nonrelevant, and k is a tuning constant designed to avoid volatility when R is small. Buckley and Voorhees

[3] Available from trec.nist.gov.

[2004] showed that BPref is consistent with average precision when the judgments are complete, and has better behavior when the judgments are partial. However, it also shares some of the undesirable aspects of AP, including reliance on a knowledge of the number of relevant documents; a tendency to prefer a query with one answer over a query with many (that is, per-query scores are incomparable); lack of an obvious user model; and the fact that the calculated score can move upward or downward as more documents are judged. Sakai [2007] considered several further issues in connection with BPref.

Sakai [2004] proposed a Q-measure, first used at NTCIR in 2004, which for binary relevance can be expressed as

$$\text{Q-measure} = \frac{1}{R} \cdot \sum_{i=1}^{d} \left\{ r_i \cdot \left(\frac{2\sum_{j=1}^{i} r_j}{i + \min(i, R)} \right) \right\} .$$

This measure does not appear to have any advantages compared to AP or BPref.

Another recent proposal is the discounted cumulative gain (DCG) method of Järvelin and Kekäläinen [2002]. Discounted cumulative gain is monotonic in the number of relevant documents found, meaning that the score for a ranking which is a proper prefix of another serves as a lower bound on the score assigned to the longer ranking. However, DCG suffers from the problem of having no upper limit on the scores that can be assigned.

In a similar vein, Meng and Chen [2004] explored a metric they called *RankPower*, which also factors rank position into a precision-based score, but does so in a manner that is at odds with other metrics, in that strongly useful rankings generate low scores. There are a number of other anomalies with this metric that put it at odds with both DCG and RBP, including its treatment of rankings in which there are no relevant documents at all.

With a sufficient number of queries, binary relevance provides an accurate method of distinguishing between systems [Voorhees 2002]; but this does not mean that all relevant documents are in fact equal. There are many articles exploring graded relevance, which was used as early as the 1967 Cranfield-2 experiments [Voorhees 2002] and continues to be investigated; see, for example, the approaches of Borlund and Ingwersen [1998], Järvelin and Kekäläinen [2002], Della Mea and Mizzaro [2004], and [Kekäläinen 2005]. Our new mechanism can easily be applied to nonbinary judgments.

Finally in this section, note that all of the approaches to measurement discussed here relate to what Saracevic [1995] called *batch mode evaluation*, an abstraction that does not capture the richness of the ways in which users interact with practical search systems. However, the underlying similarity-measuring components of even complex systems are clearly worth measuring of themselves, and it is that measurement with which we are concerned in this article.

4. RANK-BIASED PRECISION

We now introduce the new metric for scoring rankings; describe some its properties, and compare it to the discounted cumulative gain mechanism of Järvelin and Kekäläinen [2002]. To motivate the discussion, Table I shows the

Table I.
Contribution made by each of five relevant documents toward the final AP for the ranking "$$---$----$-----$---". The AP of 0.6316 is the sum of the component contributions shown in the second to last row. The last row shows the component contributions expressed as a percentage. For example, the first document in the ranking contributes 0.3633/0.6316 = 58% of the final AP.

Precision	Relevant documents				
	1	2	6	11	17
$d = 1$	1/1				
$d = 2$	1/2	1/2			
$d = 6$	1/6	1/6	1/6		
$d = 11$	1/11	1/11	1/11	1/11	
$d = 17$	1/17	1/17	1/17	1/17	1/17
Total	1.8164	0.8164	0.3164	0.1497	0.0588
×1/5	0.3633	0.1633	0.0633	0.0299	0.0118
%	58	26	10	5	2

computation of average precision for the same ranking as was used as an example earlier, but with the final AP score attributed individually to the five relevant documents. For example, document number one contributes to all five of the precision scores that are averaged, and in doing so generates 0.3633 of the final AP of 0.6316. More than half of the final score is contributed by the first document in the ranking. Buckley and Voorhees [2005] noted the same issue of items dominating the scoring, and comment that this aspect of AP has been criticized by statisticians. Observing that each document in the ranking can be assigned a weight is one of the starting points of our proposed metric; the other is a model of user behavior.

4.1 Patient and Impatient Users

Consider the user of some retrieval system, sitting at a computer and issuing queries. Each query results in a ranked list of pretty much arbitrary length being returned to them. In our model of user behavior, we assume that a user always starts by examining the top-ranked document, then the second-ranked, then the third-ranked, and so on, until they stop looking. We further assume that, as the user looks at suggested answers, they are willing to pay $1 for each relevant answer provided by the system, but nothing for irrelevant answers. The $1 can be thought of as income to the search provider, in exchange for *utility gained by the searcher*. As the user progress down the ranked list, they are thus running up an account with the search provider, or, equivalently, increasing their total utility.

The user has no desire to examine every answer. Instead, our suggestion is that they progress from one document in the ranked list to the next with *persistence* (or probability) p, and, conversely, end their examination of the ranking at that point with probability $1 - p$. We assume that each termination decision is made independently of the current depth reached in the ranking, independently of previous decisions, and independently of whether or not the document just examined was relevant or not. (The implications of relaxing these

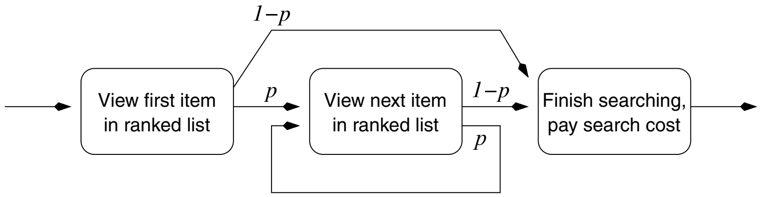

View first item in ranked list
View next item in ranked list
Finish searching, pay search cost
1−p
p
1−p
p

Fig. 3. The user model assumed by rank-biased precision.

assumptions are discussed below.) That is, we assume that the user always looks at the first document, looks at the second with probability p, at the third with probability p^2, and at the ith with probability p^{i-1}. Figure 3 shows this model as a state machine, where the labels on the edges represent the probability of changing state.

These assumptions imply that, on average,

$$\sum_{i=1}^{\infty} i \cdot p^{i-1} \cdot (1-p) = \frac{1}{1-p}$$

documents are examined during each search. If some query q has a relevance vector $\mathcal{R} = \langle r_i \mid i = 1, 2, \ldots, d \rangle$ as described earlier, then the total known expected utility derived by the user, and the income payable to the search service, are given by

$$\sum_{i=1}^{d} r_i \cdot p^{i-1}.$$

Dividing by the average number of items inspected yields an expected rate at which utility is transferred from the search provider to the user, and is the basis of our *rank-biased precision* metric:

$$\text{RBP} = (1-p) \cdot \sum_{i=1}^{d} r_i \cdot p^{i-1}.$$

This definition ensures that RBP takes on values greater than or equal to 0.0 and less than 1.0, since $\sum_{i=1}^{\infty} p^{i-1} = 1/(1-p)$.

The user model we propose is, we believe, a reasonable approximation of how people use answer lists, and similar behavior has been observed in user experiments. For example, Joachims et al. [2005] studied users in eye-tracking experiments while they were examining answer pages, and found that several suggested links would be scanned from the top of the page before a decision was made to click on one of the links to explore it, and that roughly half of users scanned only the first three suggested answers. Similarly, Hosanagar [2005] considerd the utility of returned documents relative to a cost model, and considerd how best to model the number of documents a user examines after executing a query. Järvelin and Kekäläinen [2002] also employed the notion of patient and impatient users, and proposed that an evaluation tuning knob be introduced that allows users' differing expectations and experiences to be quantified.

Table II.
Contribution made by each of five relevant documents toward the RBP score using three different p values, on the ranking "$$---$----$-----$---".

Document	$p = 0.50$	$p = 0.80$	$p = 0.95$
1	1.0000	1.0000	1.0000
2	0.5000	0.8000	0.9500
6	0.0313	0.3277	0.7738
11	0.0010	0.1074	0.5987
17	0.0000	0.0281	0.4401
Total	1.5322	2.2632	3.7626
$\times(1-p)$	0.7661	0.4526	0.1881

Nor is the notion of allowing for different types of user a new one. Nearly four decades ago, while discussing his expected search length measure and how to evaluate document rankings, Cooper [1968], page 37, described a range of possible user behaviors and information needs, and noted the following:

> This [discussion] suggests that the parameter of greatest interest in evaluating a system is . . . *expected search length per desired relevant document*—that is, the expected number of irrelevant documents to be screened out for each relevant document found.

By measuring the expected rate at which relevant documents are found, the RBP metric described here also captures elements of Cooper's intentions in this regard.

4.2 Rank-Biased Precision

As an example of the RBP computation, let us return to the same ranked list as was considered in Table I, in which relevant documents appear at ranks $1, 2, 6, 11$, and 17. Table II shows the RBP computation, for three different values of p. For this particular ranking the $p = 0.5$ score gives the largest RBP value, a consequence of the fact that the first two documents are both relevant. If the top-ranked document was not relevant, all of the scores would be lower than shown in the table, but the $p = 0.50$ scores would decrease by the greatest margin.

The use of different values of p reflects different ways in which ranked lists can be used. Values close to 1.0 are indicative of highly persistent users, who scrutinize many answers before ceasing their search. For example, at $p = 0.95$, there is a roughly 60% likelihood that a user will enter a second page of 10 results, and a 35% chance that they will go to a third page. Such users obtain a relatively low per-document utility from a search unless a high number of relevant documents are encountered, scattered through a long prefix of the ranking.

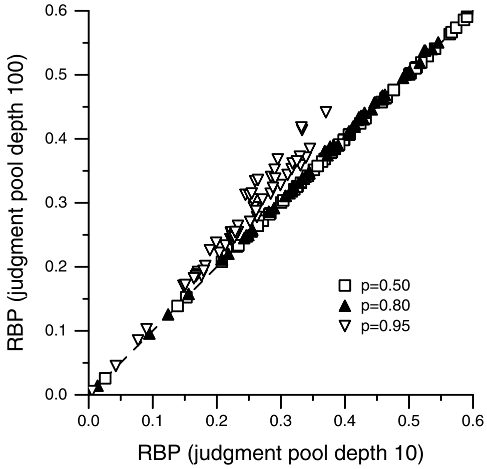


Fig. 4. Rank-biased precision of 61 TREC-5 systems, for three different values of p, using relevance judgments compiled using two different pool depths. Rank-biased precision at $p = 0.5$ and $p = 0.8$ is stable when the pool depth is increased from 10 documents per system to 100 documents. At $p = 0.95$ the RBP scores increase (and never decrease) when the pool depth is increased.

Compare the behavior of a persistent user to the one-in-a-thousand chance of a $p = 0.5$ user entering even the second page of 10 results. Users in the $p = 0.5$ category are highly impatient, but obtain high average per-document utility (that is, high RBP) whenever there is a relevant document in the first one or two rank positions. In the limit, use of $p = 0.0$ implies a user who is "feeling lucky" and is either satisfied or dissatisfied with the top-ranked document, and never looks any further. This latter mode corresponds exactly to evaluating the system using P@1.

Figure 4 shows the effect of calculating average RBP scores over the 61 systems that participated in TREC-5 in 1996, calculated using two different pool depths for the relevance assessments. Three different values of the parameter p were used, covering a range from relatively impatient users ($p = 0.5$) through to relatively patient users ($p = 0.95$). When $p = 0.5$ and $p = 0.8$, the system average scores calculated based on judgments extracted from a pool depth of 10 documents per run are almost identical to the scores generated when a pool depth of 100 is used. When $p = 0.95$, a pool depth of 10 is insufficient to give accurate RBP scores, and the correlation is weaker. Note, however, that adding further relevance judgments into the computation increases the system score, rather than decreasing it. That is, unlike the situation with AP that is depicted in Figure 2, system scores using rank-biased precision can always be regarded as lower bounds on the score that would be obtained were perfect relevance information to be available.

It was noted above that the interpretation of precision scores needs to be tempered by knowledge of R, the number of relevant documents. The same is also true of RBP, since a persistent user (with say $p = 0.95$) is guaranteed to obtain a low expected utility from a search with only a few relevant documents.

For example, when there are only five relevant documents and $p = 0.95$, it must be that RBP < 0.23. However, we resist the temptation to normalize the scores based on the maximum attainable score for each query, since to do so would defeat the purpose of introducing RBP. Instead, we observe that, however low the RBP score is for a particular query and p value, it still reflects the average rate at which utility is gained by that particular user. Impatient users will also obtain low RBP scores if none of the top few documents are relevant.

4.3 User Models

A key part of the RBP proposal is the user model, and the notion of scoring a ranking according to the average utility gained by the person using the ranking. Other user models also give rise to possible scoring regimes. For example, consider a simple user model in which documents are examined starting with the first, until a relevant one is found. The total utility gained from a ranking by such a user will always be $1, and they will have examined documents until the first relevant one. The average utility per document examined is thus exactly the score assigned by the reciprocal rank (RR) metric. That is, RR can also be thought of as a scoring regime with a tractable user model.

An obvious extension is then to consider other models. For example, a *sceptical* user might not stop when they see the first relevant document in the ranking, and instead continue until they have seen corroboration from a second relevant one. Such a user can score rankings by a RR2 metric, in which the reciprocal rank of the second relevant document is what matters. Another variation on RR is to use a "damping" factor, computing $1/(k + \min_i\{r_i = 1\})$ instead of $1/(\min_i\{r_i = 1\})$, where k is a constant. This metric corresponds to a *cautious* user, who stops examining documents only after they have looked at k past the first relevant one.

A further variant is to relax the assumption that p, the probability that the user advances to the next document in the ranking, is independent of whether or not the document just considered is relevant. An arrangement in which the conditional probability of advancing given a relevant document is p_1, and the conditional probability of advancing given an irrelevant document is p_2, would still allow the average utility per document inspected to be calculated, and would lead to another mechanism for scoring runs and thus comparing retrieval systems.

4.4 Bounding the Residual Error

A useful consequence of the proposed RBP metric is that it is possible to compute upper and lower bounds on effectiveness, even when the ranking and relevance judgments are partial rather than comprehensive. For example, consider the TREC environment, in which the top (say) 100 documents from multiple runs for each query are combined into a single pool and then judged, but systems are compared on the basis of 1000 answers for each query. By construction, all of the top 100 documents in each run of 1000 have been judged, and perhaps others beyond the top 100 too, because of the pooling. But the great majority of documents further down the rankings will be unjudged.

As noted earlier, the convention in TREC evaluation is that any unjudged documents are taken to be not relevant, and that only "lower bounds" on effectiveness should be computed. With this assumption, quoted effectiveness rates might be expected to be pessimistic, meaning that with a greater volume of judgments, measured effectiveness should increase. But Figure 2 clearly shows that, when average precision is used as the effectiveness metric, the default assumptions do not lead to a lower bound being calculated. That is, assuming that unjudged documents are irrelevant is not necessarily pessimistic in the context of AP.

In the RBP measure it is straightforward to accumulate an uncertainty value, or *residual*, that captures the unknown component of the effectiveness metric. The simplest case is when the ranking is calculated to a depth of d answers per query, and the contributions from depth $d + 1$ on are not available. Then the uncertainty in the RBP score is given by

$$(1-p) \cdot \sum_{i=d+1} p^{i-1} = (1-p) \cdot p^d \cdot \sum_{i=1} p^{i-1} = p^d.$$

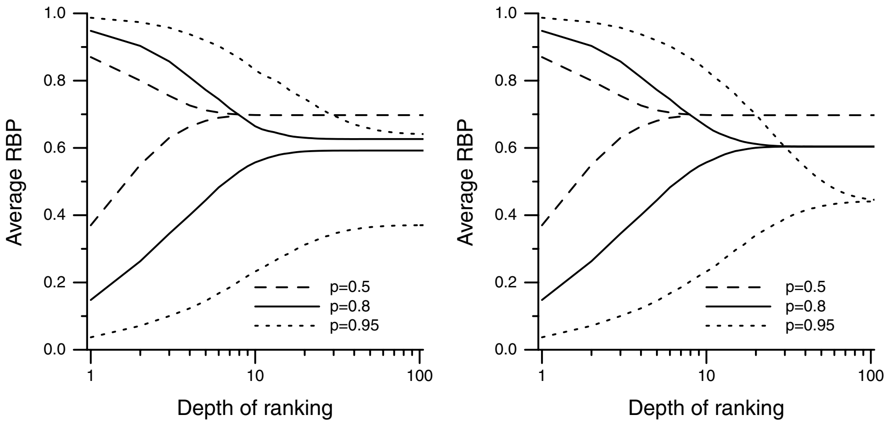
When the judgments are nonexhaustive, missing items should be added to the residual on an item-by-item basis, using the weight they would have had if they were relevant. For example, consider the ranking "`$$---$----$-??--?---`," where a "?" represents a missing judgment. The documents ranked in positions 13, 14, and 17 are unjudged, so the properties of the geometric distribution mean that the uncertainty is given by $p^{20} + (1-p)(p^{12} + p^{13} + p^{16})$. For $p = 0.5$, the RBP is bounded by 0.7661 and 0.7663; for $p = 0.8$, the RBP is bounded by 0.447 and 0.489; for $p = 0.95$, the RBP is bounded by 0.17 and 0.60. Each of these ranges encompasses the values given in Table II.

The residual calculation can be done in advance of any experimentation. For example, with $p = 0.8$ and a pooling depth of $d = 20$, the residual from all remaining terms in the geometric series is $0.8^{20} = 0.012$, which implies that calculated RBP figures should be quoted to at most two decimal places. Conversely, when four decimal digits of accuracy are required, the residual should be less than 0.0001, and the required depth to attain this is a function of the value of p being used:

$$p^d < 0.0001 \Rightarrow d > \frac{\ln 0.0001}{\ln p} \approx \frac{9.21}{1-p}.$$

When $p = 0.5$, $p = 0.8$, and $p = 0.95$, this expression suggests minimum evaluation depths of $d = 14$, $d = 42$, and $d = 180$, respectively. Another way of looking at this analysis is that, in a TREC-style pooled evaluation, a depth of $d = 100$ of guaranteed exhaustive judgments is sufficient to support four digits of accuracy in the computation of RBP only when $p \leq 0.91$. Use of larger values of p will require a greater pool depth if four digits of accuracy are to be presented.

The behavior of the lower and upper bounds on RBP for one TREC-5 run for three different values of p and two different judgment pool depths is illustrated in Figure 5. Within each of the two graphs in the figure, increasing the amount of information taken into account by increasing the depth d of the ranking allows



(a) Pool depth 10 (b) Pool depth 100

Fig. 5. Upper and lower bounds for RBP as p is varied and increasing numbers of documents are considered in the ranking for one of the submitted TREC-5 runs, for (a) a pool depth of 10; and (b) a pool depth of 100. Note how the upper and lower bounds stabilize as the depth d of the evaluation is increased, but, for larger values of p, do not converge if the pool depth on which the relevance judgments are based is too small.

increased accuracy in the estimated effectiveness values. Comparing the left-hand and right-hand graphs in Figure 5 shows that increasing the depth of the pool of relevance judgments allows convergence toward accurate scores, with (in the right-hand graph) the upper bound closing on the lower bound even when $p = 0.95$. The balance between p, the accuracy of the score, and the cost of relevance evaluations, is something that can be designed into retrieval experiments in a manner that is simply not possible with AP.

4.5 Choosing a Value for p

An obvious question is that of choosing a value for p. Ideally that choice would be made during the design phase of any experiment, as an estimate of the type of user characteristic being tested in the experiment, and as a parameter that helps determine how much the experiment will cost if it is to yield data of a specified accuracy. Alternatively, the choice of p can be made after the experiment has been carried out, in which case the accuracy of the resulting scores can be computed. A third option, for systems claimed to be “broad spectrum” and suitable for all types of users, would be to design the experiment using a high value of p, and then report RBP results for several different value of p.

Small values of p, less than around 0.5, place the bulk of their emphasis on the first few positions in the ranking, and provide less balance across the whole of a ranked list. However, this bias means that small values of p also allow cheaper evaluation, because fewer documents need to be judged to obtain a given level of accuracy in the scoring. As p gets larger, the emphasis on early rank positions is reduced, and an increasing fraction of the total weighting is available to later rank positions, modeling users who are more persistent, and

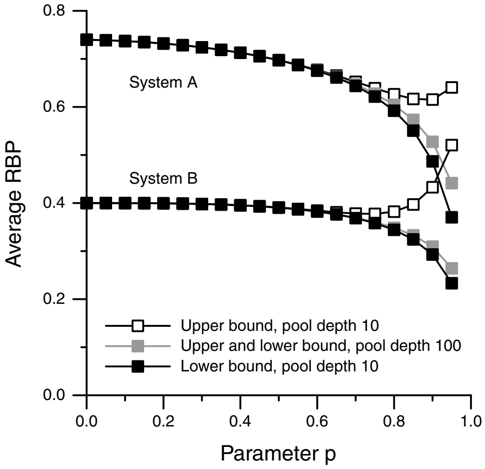


Fig. 6. Rank-biased precision scores for two of the TREC-5 runs, averaged across 50 queries. Upper and lower bounds for RBP at pool depth 10 are plotted as a function of p, together with the RBP value computed for that value of p when the pool depth is 100. When p is large, the error tolerance between the upper and the lower bounds is large for depth 10 evaluations. Increasing the pool depth to 100 gives convergence even at high values of p. The line denoted *System A* is the same system as illustrated in Figure 5.

likely to look at (in the Web search environment) the second or third page of Web results. The weightings are still monotonic, and, even with $p = 0.95$, the document in rank position 100 gets a weighting of just 0.6% of the document in rank position 1. But increasing p toward 1 also implies that an increasing amount of effort must be spent on relevance judgments, as otherwise the accumulated imprecision is too large.

Figure 6 shows this balance for two TREC-5 runs, with RBP averaged over the set of 50 applicable queries. As was also done for Figures 2, 4, and 5, the relevance judgments at a pool depth of 100 performed for TREC-5 were used to extract the set of relevance judgments that would have been formed if the pool depth was only 10. Three lines are plotted in the graph for each of the two systems: the upper and lower bounds on RBP with a pool depth of 10, and the (indistinguishable at the scale of the graph) upper and lower bounds on RBP when the pool depth is increased to 100. The pattern of the three curves shows typical behavior: with an assessment pool depth of only 10, values of p greater than around 0.7 lead to noticeable imprecision in the scores; but when the assessment pool depth is increased to 100, values of p as large as 0.95 can be handled with only small residual errors.

In summary, if reliable experiments with large p are required, the pool depth used to form the relevance judgments must be high. On the other hand, reliable scores can be generated using relatively shallow assessment pool depths when $p \leq 0.8$. Searching processes that are intended to be "high recall" should thus be assessed with a relatively high value for p, whereas Web-user search tasks can be assessed using a smaller value of p, and cheaper experiments.

4.6 Discounted Cumulative Gain

In work first presented at SIGIR in 2000, Järvelin and Kekäläinen [2002] described a metric they called *discounted cumulative gain* (DCG) that shares some of the features of RBP. For a relevance vector $\mathcal{R}$ of length d, DCG is defined as

$$\text{DCG-}b = \sum_{i=1}^{b} r_i + \sum_{i=b+1}^{d} \frac{r_i}{\log_b i},$$

where b is a persistence parameter akin to our p parameter, and relevance contributions are weighted more highly earlier in the ranking than they are in the later rank positions. Järvelin and Kekäläinen [2002] suggested the use of $b = 2$, and employed that value in their examples and experiments. The intention of DCG is that high-ranking relevant documents give more satisfaction than do low-ranking ones, the same notion as is built into RBP. However, where RBP discounts relevance via a geometric sequence, DCG does so using a log-harmonic one.

The change in discounting regime we propose in RBP is a critical one. Consider what happens with a relevance vector $\mathcal{R} = \langle 1, 1, 1, \ldots, 1 \rangle$, representing a ranking in which every retrieved item is relevant. With RBP, a score of close to 1.0 will be assigned, regardless of p, and regardless of the ranking depth, with the discrepancy between the actual score and 1.0 completely accounted for by the residual uncertainty. On the other hand, with DCG the maximum score grows without limit as the answer ranking is deepened. To limit the value of DCG to 1.0 for a given ranking depth, a scaling factor would be required, and would depend on d. For example, a scaled DCG score calculated for a ranking depth of $d = 100$ (for which the scaling constant would need to be 21.79 when $b = 2$) might decrease by a factor of almost 5 if the ranking depth d was increased to 1000 (for which the corresponding scaling constant would be 123.99).

Järvelin and Kekäläinen [2002] recognized the need for a normalized form of DCG, and took a different approach to the scaling problem. Rather than compute the normalization constant based on an "all relevant" ranking as hypothesized in the previous paragraph, they suggested that it should be computed relative to the DCG score of a "perfect" ranking at that depth, where a perfect ranking lists all (known) relevant documents first, followed by all nonrelevant documents. From our point of view, this approach is unsatisfactory, since, to calculate a *normalized discounted cumulative gain* (NDCG) score in this way, all relevant documents (and thus the value of R) must be identified. That is, NDCG has the same issues as AP and P@R. A similar assumption weakens the Q-measure of Sakai [2004], which was also proposed in both unnormalized and normalized forms.

4.7 Other Extensions

A further part of the rationale for DCG and NDCG (see also Sakai [2004] and Kekäläinen [2005]) is to provide for nonbinary relevance judgments, where documents are considered to be relevant to varying degrees, and the vector

$\mathcal{R}$ is constructed over a richer domain. In the experiments of Järvelin and Kekäläinen [2002], $r_i \in \{0, 1, 2, 3\}$, and the relevance judgments were four-way.

The same flexibility is readily accommodated in our framework, by scaling the r_i values to the unit range, and working with $r_i \in \{0.00, 0.33, 0.67, 1.00\}$ (or any other desired subset of the real numbers $0 \le r_i \le 1$). Rank-biased precision can then be used unchanged, with the RBP score reflecting the average per-answer rate at which the user gains utility from the ranking, assuming that a fractional relevance score reflects the fractional utility gained by the user when that document is presented to them.

Also worth noting is that the definition of RBP is readily modified to handle document rankings containing ties. For example, if the j documents in positions k to $k+j-1$ of the ranking are all deemed to be tied, then the score contribution of each is given by $(\sum_{i=k}^{k+j-1} p^{i-1})/j$, so that the total score weight attached to the group of documents is shared equally between them. Cooper [1968] and Raghavan et al. [1989] considered a similar solution in connection with the expected search length metric.

5. RBP VERSUS AP IN RETRIEVAL EXPERIMENTS

Rank-biased precision addresses many of the concerns that have been raised in connection with average precision. However, AP is widely regarded as a reliable way of comparing system retrieval performance, and has accumulated more than a decade of experimental confidence. It is thus natural to turn to experiment, to compare the usefulness of rank-biased precision and other metrics. We do this in two ways.

The first of these experiments makes direct use of the TREC evaluation methodology, which has as one of its objectives a desire to order retrieval systems, so that lessons can be inferred concerning techniques that work well and other techniques that do not. The question we ask is this: how different is the system ordering generated by RBP compared to the system orderings that result when other effectiveness metrics are used? Table III shows the results.

To build Table III, system orderings were generated using the submitted TREC-5 runs for each of a range of effectiveness metric computations and relevance assessment pool depths. Each system ordering contained 61 system run names, based on numeric average effectiveness scores, without regard to whether or not the ordering of adjacent items could be defended via a significance test. That is, each overall average system score was taken at face value, and used to assign that system a rank in a "performance league table."

The different system orderings were then pairwise compared using Kendall's τ, which calculates a numeric similarity score for a pair of ordered lists over a common domain. A subset of those results appears in Table III. A score of 1.0 indicates that the two lists of system names are in exactly the same order, while a score of -1.0 would indicate that one list is the reverse of the other. Four reference orderings, all calculated using an assessment pool depth of 100, are shown as the columns.

The preponderance of numbers greater than 0.8 in Table III shows that all of the listed effectiveness metrics are generating similar system orderings, and

Table III.
Kendall's τ correlation coefficients calculated from the system orderings generated by pairs of metrics using the 61 TREC-5 runs. A value of 1.0 indicates perfect agreement between the two metrics, in terms of the system ordering that they produce. The largest (nonself) value in each row is highlighted in boldface, with the top part of the table showing that RR is most like P@10; that P@10 is most like P@R; and that P@R is most like AP. The bottom group of rows shows the same correlation coefficients for RBP. When $p = 0.5$, RBP behaves most like RR. When RBP uses $p = 0.8$, the best agreement is with P@10. When RBP uses $p = 0.95$, there is good agreement with all of P@10, P@R, and AP.

Metric	Pool	Kendall's τ, pool depth 100			
	depth	RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**

are thus carrying out broadly the same task. Worth noting, however, is that RBP with $p = 0.5$ gives similar behavior to reciprocal rank; and that RBP with $p = 0.95$ compares well to all of P@10, P@R, and average precision (which is known to correlate well with P@R [Buckley and Voorhees 2005; Aslam et al. 2005]). Also worth comment is that RBP provides AP-similar system rankings even when the relevance assessment pool depth is just 10. That is, in terms of experimental cost, it may be preferable to use RBP with an assessment depth of 10 than it is to use AP with a depth of 100. Similar results (not shown here) were obtained when the same experiment was applied to the 127 system runs submitted to TREC-8 in 1999, and when Spearman correlation coefficients were computed rather than Kendall's τ.

To put Table III into perspective, we also computed the Kendall's τ correlation scores for the relationships plotted in Figures 2 and 4. In the case of the Figure 2 comparison between AP with a pool depth of 10 and AP with a pool depth of 100, the correlation score was 0.898. Figure 4 plots three sets of similar relationships; the Kendall's τ scores for $p = 0.5$, $p = 0.8$, and $p = 0.95$ were, respectively, 1.000, 0.986, and 0.891.

In the second investigation, we tested the consistency of a range of metrics, measured in terms of their ability to provide support for questions of the form "are these two systems significantly different?" In this experiment, the 61 TREC-5 systems were pairwise compared over the 50 queries, using query similarity scores computed using several different effectiveness metrics. For each combination of system pair and evaluation metric, two statistical tests, at two significance levels, were applied, and the number of pairwise system comparisons that generated "yes, they are significantly different" outcomes was counted.

Table IV.
The rate at which different effectiveness metrics allow significant distinctions to be made between retrieval methods. A total of 61 system runs were pairwise compared using the TREC-5 queries, making a total of $61 \times 60/2 = 1830$ system comparisons. The four columns show the number of those tests that were judged to be significant using the indicated statistical comparison. Of the traditional metrics, AP is the most consistent, in terms of allowing systems to be experimentally separated; of the RBP variants, that with $p = 0.95$ is the most consistent. The NDCG measure is a little better than both RBP and AP. In all cases the test undertaken was a two-tailed one, to answer the question "Are the two systems significantly different?"

Metric	Wilcoxon		*t* test	
	95%	99%	95%	99%
RR	1020	759	1000	752
P@10	1141	897	1150	915
P@*R*	1209	989	1142	931
AP	1259	1077	1164	969
RBP, $p = 0.5$	1067	834	1050	810
RBP, $p = 0.8$	1164	919	1166	917
RBP, $p = 0.95$	1231	1006	1209	987
NDCG	1291	1092	1269	1101

The results of the second experiment are shown in Table IV. The trend in each column of the table is clear—of the conventional metrics in the top part of the table, AP is more consistent than P@10 is more consistent than RR; and using the RBP approach, evaluation using $p = 0.95$ is more likely to yield system separation than is $p = 0.8$ or $p = 0.5$. The NDCG measure appears to be even more consistent in its behavior. The same pattern is observed over all four combinations of significance level and statistical test. We conclude that RBP, with an appropriate choice of p, is comparable to existing metrics in terms of its usefulness in supporting system comparisons.

6. CONCLUSION

We have defined a new measure of retrieval effectiveness, designed to address the shortcomings that can be observed in the measures that are currently in common use. Our rank-biased precision measure has the following attractive properties:

—It is derived from a straightforward state-based model of user behavior that has support in empirical user studies.

—It can be interpreted in an economic-modeling sense, as the average rate at which the user gains utility from performing their search, including in situations in which graded relevance judgments are being used.

—It measures only the behavior of the system as observed by the user, and it does not rely on unknowns such as collection size, or the number of documents relevant to each query.

- It provides a mechanism for allowing calculation of the required ranking depth if scores are to be presented to a certain level of accuracy.
- In the presence of uncertainty (partial rankings, or unjudged documents), an error bound can be precisely determined.
- It is well defined even when a query has no answers.
- Depending on the parameter p chosen, it gives overall system rankings similar to reciprocal rank, or similar to P@10, or similar to P@R and AP.
- It is as likely as comparable measures to lead to statistically significant system comparisons.

In addition, we have shown how the use of RBP at the "performing judgments" stage of an experiment can reduce the amount of effort needed (compared to pooling) when TREC-style system comparisons are being carried out using RBP as the evaluation metric [Moffat et al. 2007], noting that similar relationships based on MAP have also been proposed [Aslam et al. 2006; Carterette et al. 2006; Cormack and Lynam 2006; Büttcher et al. 2007].

Needing to be weighed in the balance against these benefits are the following:

- For practical purposes RBP values are always strictly less than 1, since only an infinitely long ranking of relevant documents can give rise to RBP = 1.
- Rank-biased precision scores reflect user satisfaction in absolute terms, rather than in "relative to the maximum possible for this query" terms. It may be that standardization [Webber et al. 2008] can be usefully applied to RBP, but we have yet to explore this possibility.
- In any evaluation, the person reporting the experiment must choose a value of p, and be willing to defend that choice to their target audience. If their chosen p is significantly higher than the p used by the person designing the experiment, then (a calculable level of) imprecision will result.

On balance, and taking both the drawbacks and benefits into account, we believe that rank-biased precision provides a useful tool that will be of benefit in all situations where recall, or precision, or some amalgam of them such as average precision, might currently be used.

ACKNOWLEDGMENTS

Jamie Callan, Bruce Croft, Mark Sanderson, Ellen Voorhees, and William Webber provided helpful assistance.

REFERENCES

ALLAN, J., CARTERETTE, B., AND LEWIS, J. 2005. When will information retrieval be "good enough"? See Marchionini et al. [2005], 433–440.

ASLAM, J. A., PAVLU, V., AND YILMAZ, E. 2006. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, Eds. ACM Press, New York, NY, 541–548.

ASLAM, J. A., YILMAZ, E., AND PAVLU, V. 2005. A geometric interpretation of r-precision and its correlation with average precision. See Marchionini et al. [2005], 573–574.

Borlund, P. and Ingwersen, P. 1998. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In *Proceedings of the Twenty-First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds. ACM Press, New York, NY, 324–331.

Buckley, C. and Voorhees, E. M. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, Eds. ACM Press, New York, NY, 25–32.

Buckley, C. and Voorhees, E. M. 2005. Retrieval system evaluation. In *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, Chapter 3, 53–75.

Büttcher, S., Clarke, C. L. A., Yeung, P. C. K., and Soboroff, I. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, Eds. ACM Press, New York, NY, 63–70.

Carterette, B., Allan, J., and Sitaraman, R. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, Eds. ACM Press, New York, NY, 268–275.

Cooper, W. S. 1968. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Amer. Document. 19*, 1 (Jan.), 30–41.

Cooper, W. S. 1973. On selecting a measure of retrieval effectiveness: Part I, the ‘subjective’ philosophy of evaluation. *J. Amer. Soc. Inform. Sci. 24*, 87–100.

Cormack, G. V. and Lynam, T. R. 2006. Statistical precision of information retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, Eds. ACM Press, New York, NY, 533–540.

Della Mea, V. and Mizzaro, S. 2004. Measuring retrieval effectiveness: a new proposal and a first experimental validation. *J. Amer. Soc. Inform. Sci. Tech. 55*, 6, 530–543.

Frei, H. P. and Schäuble, P. 1991. Determining the effectiveness of retrieval algorithms. *Inform. Process. Manage. 27*, 2/3, 153–164.

Harman, D. 1995. Overview of the second text retrieval conference (TREC-2). *Inform. Process. Manage. 31*, 3, 271–289.

Harter, S. P. 1996. Variations in relevance assessments and the measurement of retrieval effectiveness. *J. Amer. Soc. Inform. Sci. 47*, 1, 37–49.

Hosanagar, K. 2005. A utility theoretic approach to determining optimal wait times in distributed information retrieval. See Marchionini et al. [2005], 91–97.

Järvelin, K. and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inform. Syst. 20*, 4, 422–446.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. See Marchionini et al. [2005], 154–161.

Kagolovsky, Y. and Moehr, J. R. 2003. Current status of the evaluation of information retrieval. *J. Med. Syst. 27*, 5, 409–424.

Keen, E. M. 1992. Presenting results of experimental retrieval comparisons. *Inform. Process. Manage. 28*, 4, 491–502.

Kekäläinen, J. 2005. Binary and graded relevance in IR evaluations. *Inform. Process. Manage. 41*, 5, 1019–1034.

Losee, R. M. 2000. When information retrieval measures agree about the relative quality of document rankings. *J. Amer. Soc. Inform. Sci. 51*, 9, 834–840.

Marchionini, G., Moffat, A., Tate, J., Baeza-Yates, R., and Ziviani, N., Eds. 2005. *Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY.

Meng, X. and Chen, Z. 2004. On user-oriented measurements of effectiveness of Web information retrieval systems. In *Proceedings of the International Conference on Internet Computing*, H. R. Arabnia, O. Droegehorn, and S. Chatterjee, Eds. CSREA Press, Las Vegas, NV, 527–533.

Mizzaro, S. 1997. Relevance: The whole history. *J. Amer. Soc. Inform. Sci. 48*, 9, 810–832.

MOFFAT, A., WEBBER, W., AND ZOBEL, J. 2007. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, Eds. ACM Press, New York, NY, 375–382.

RAGHAVAN, V. V., JUNG, G. S., AND BOLLMANN, P. 1989. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inform. Syst. 7*, 3, 205–229.

SAKAI, T. 2004. Ranking the NTCIR systems based on multigrade relevance. In *Proceedings of the AIRS Asian Information Retrieval Symposium*. Lecture Notes in Computer Science, vol. 3411. Springer, Berlin, Germany, 251–262.

SAKAI, T. 2007. Alternatives to Bpref. In *Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, Eds. ACM Press, New York, NY, 71–78.

SANDERSON, M. AND ZOBEL, J. 2005. Information retrieval system evaluation: Effort, sensitivity, and reliability. See Marchionini et al. [2005], 162–169.

SARACEVIC, T. 1995. Evaluation of evaluation in information retrieval. In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, E. A. Fox, P. Ingwersen, and R. Fidel, Eds. ACM Press, New York, NY, 138–146.

SHAW, JR., W. M. 1986. On the foundation of evaluation. *J. Amer. Soc. Inform. Sci. 37*, 5, 346–348.

SU, L. T. 1994. The relevance of recall and precision in user evaluation. *J. Amer. Soc. Inform. Sci. 45*, 3 (Apr.), 207–217.

TAGUE-SUTCLIFFE, J. 1992. The pragmatics of information retrieval experimentation, revisited. *Inform. Process. Manage. 28*, 4, 467–490.

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval*, 2nd ed. Butterworths, London, U.K.

VOORHEES, E. M. 2002. The philosophy of information retrieval evaluation. In *Proceedings of the 2001 Cross Language Evaluation Forum Workshop*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, Eds. Lecture Notes in Computer Science, vol. 2406, Springer, Berlin, Germany, 355–370.

WEBBER, W., MOFFAT, A., AND ZOBEL, J. 2008. Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY. To appear.

WITTEN, I. H., MOFFAT, A., AND BELL, T. C. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd ed. Morgan Kaufmann, San Francisco, CA.

ZOBEL, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the Twenty-First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds. ACM Press, New York, NY, 307–314.

Received October 2005; revised September 2007; accepted April 2008