Inf Retrieval (2010) 13:46–69
DOI 10.1007/s10791-009-9099-7

Click-based evidence for decaying weight distributions in search effectiveness metrics

Yuye Zhang · Laurence A. F. Park · Alistair Moffat

Received: 13 November 2008 / Accepted: 10 June 2009 / Published online: 30 June 2009

Abstract Search effectiveness metrics are used to evaluate the quality of the answer lists returned by search services, usually based on a set of relevance judgments. One plausible way of calculating an effectiveness score for a system run is to compute the inner-product of the run's relevance vector and a "utility" vector, where the *i*th element in the utility vector represents the relative benefit obtained by the user of the system if they encounter a relevant document at depth *i* in the ranking. This paper uses such a framework to examine the user behavior patterns—and hence utility weightings—that can be inferred from a web query log. We describe a process for extrapolating user observations from query log clickthroughs, and employ this user model to measure the quality of effectiveness weighting distributions. Our results show that for measures with static distributions (that is, utility weighting schemes for which the weight vector is independent of the relevance vector), the geometric weighting model employed in the rank-biased precision effectiveness metric offers the closest fit to the user observation model. In addition, using past TREC data as to indicate likelihood of relevance, we also show that the distributions employed in the BPref and MRR metrics are the best fit out of the measures for which static distributions do not exist.

Keywords Effectiveness metric · Query log · Clickthrough · Rank-biased precision · Average precision · Reciprocal rank · BPref

Y. Zhang (✉) · L. A. F. Park · A. Moffat
Department of Computer Science and Software Engineering, The University of Melbourne, Melbourne, VIC, Australia
e-mail: zhangy@csse.unimelb.edu.au

L. A. F. Park
e-mail: lapark@csse.unimelb.edu.au

A. Moffat
e-mail: alistair@csse.unimelb.edu.au

1 Introduction

Search effectiveness metrics are used to evaluate the quality of the answer lists returned by search services, usually based on a set of standard queries and a partial or complete set of relevance judgments (Voorhees and Harman 2000). One plausible way of calculating an effectiveness score for a system run is to compute the inner-product of the run's relevance vector and a vector of weights, where the weights represent the relative benefit obtained by the user of the system if they encounter a relevant document at that depth in the ranking. For example, the simple metric *precision at depth five*, P@5, can be thought of as being the inner-product of a binary relevance vector, and the weight vector [0.2, 0.2, 0.2, 0.2, 0.2, 0.0,…]. Use of P@5 as an effectiveness metric is implicitly an argument that the user will derive equal utility from seeing a relevant document in any ranked position from one to five, and no utility whatsoever from relevant documents that appear at ranks below five.

More complex weighting schemes have also evolved. Many of these make use of infinite decaying weight distributions, so that every document in the ranked list has some—albeit, vanishingly small—influence on the effectiveness score assigned to that run. For example, the rank-biased precision (RBP) metric of Moffat and Zobel (2008) makes use of a geometric weight vector controlled by a parameter p; and the discounted cumulative gain (DCG) metric of Järvelin and Kekäläinen (2002) uses a vector of weights based on an inverse log function, controlled by the logarithm base b.

Our work in this paper is motivated by a desire to compare the behavior of different effectiveness evaluation metrics. By factoring out the relevance vector input required for various metrics, we are able to represent the intrinsic weighting models of those metrics as distributions, and then compare them with each other. Furthermore, because the utility of a system is based on the relevant documents seen by the users of the system, if we were to obtain a model of the manner in which users view retrieved documents, it is possible to quantify the relationship between the search effectiveness observed by users, and the search effectiveness score assigned by an evaluation metric.

In particular, we examine the correlation between effectiveness evaluation metrics and user behavior patterns in the web search context, with evidence for the relationship being inferred from a detailed analysis of a search query and clickthrough log. Our results show that for measures with static weight distributions (that is, for which the weighting vector is fixed across all queries and systems), the geometric weighting model employed in rank-biased precision offers the closest fit to the user observation model. In addition, using past TREC data as to indicate likelihood of relevance, we also show that the distributions employed in the BPref and MRR metrics are the best fit out of the measures for which static distributions do not exist. Furthermore, when used as a stand alone evaluation metric, the observation model shows a high degree of correlation with other metrics, most notably rank-biased precision, for which a system-order Kendall's τ value of greater than 0.98 is attained over the systems that took part in the TREC9 Web Track.

In summary, the key contributions made are:

- A framework describing static weighting models of inner-product evaluation metrics and their relationship to user observations in effectiveness evaluation;
- A method for deriving weighting models for complex evaluation metrics that require global relevance information and/or relevance positions, using past TREC runs;
- A process for creating a user observation model from query log clickthroughs; and
- An analysis of fit values between various established evaluation metrics and the observation model derived from a recent Microsoft Search query log.

The paper is structured as follows: Sect. 2 introduces some commonly employed measures in information retrieval effectiveness evaluation, and groups them based on the key characteristics. Section 3 then describes the generalized form of those measures via weight distributions, as well as a method for using past TREC runs to estimate the weighting models of complex measures built on empirical relevance information. The process of obtaining a model of the manner in which users view the results ranking based on the clickthrough data is presented in Sect. 4, together with methods for processing the observation model to obtain a more realistic representation. Section 5 shows the results of experiments that examine the fit of weighting distributions of evaluation metrics to an observation model created from a recent Microsoft Search (MSN) query log. Recent work in the area is outlined in Sect. 6, and compared to our study. Finally, Sect. 7 summarizes our results, and gives directions for possible future work in this area.

2 Measuring effectiveness in web search

A range of metrics are used to quantify the effectiveness of the document rankings generated by document retrieval systems. *Average precision* is often employed, as are P@k for some value k and *reciprocal rank*, the inverse of the depth in the ranking of the first relevant document. For experimental purposes, we divide these and other measures into two broad categories: *recall-based measures*, which require knowledge of the total number of relevant documents in the collection for each query; and *precision-based measures*, which do not require this value.

2.1 Recall-based measures

Knowledge of the number of relevant documents, denoted R, that are present (on a per query basis) in the collection allows the use of effectiveness metrics that encompass completeness. *Recall* itself is defined as the fraction of all relevant documents that have been retrieved. More complex metrics that incorporate R also include this notion of completeness. For example, precision can be biased to include the notion of recall by measuring it over a range of recall percentiles, or at depth R in the case of R-precision.

Average precision (AP) is computed by summing the observed precision at each relevant document in the ranking, and dividing by the total number of relevant documents for the query:

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

where r_i is the (binary) relevance of the ith document and $\mathcal{R} = \{r_i\}$ is the corresponding *relevance vector*. Average precision scores have the potential to shift both up and down as further documents get judged, particularly if the value of R is varied as a result of the additional judgments (Moffat and Zobel 2008). Nevertheless, AP tends to be reasonably well behaved in standard test environments, and mean average precision (MAP, the average of the AP scores across a set of topics) is probably the most widely reported IR effectiveness metric in current evaluations.

To address the issues that arise from incomplete relevance judgments, the BPref (or *binary preference*) measure rewards rankings in which known-relevant documents are

ranked highly, and punishes highly ranked known-irrelevant documents (Buckley and Voorhees 2004):

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

where $\mathcal{R}'$ is a modified results vector from which all unjudged documents have been removed, N is the number of known-irrelevant documents, and k is a tuning constant introduced to counter erratic outputs when R is small ($k = 10$ is commonly employed). The BPref metric is more stable than AP when there are unjudged documents, and is employed in large-scale experiments in which relevant documents are relatively less common. Sakai (2007) comments on the usefulness of BPref for effectiveness evaluation, and notes several situations in which it performs suboptimally in comparison to other measures.

The Q-measure is a further variant of AP (Sakai 2004):

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

For queries with a small number of relevant documents, the Q-measure assigns higher precision contributions when a relevant document is encountered at larger depths. Its behavior when dealing with large R values is similar to AP. The Q-measure is also designed specifically to handle graded relevance judgments in applicable test collections such as NTCIR.

The single biggest drawback of recall-based measures is the need to determine R. For all but trivial document collections it is infeasible to carry out exhaustive relevance judgments, and R is usually approximated using pooling—the set of top ranked documents from runs submitted by participating systems are judged, and all documents that were not ranked highly by any of the participating systems are left unjudged and (for the purposes of calculating AP and other recall-based measures) deemed to be irrelevant. As more documents are introduced into the pool, it is to be expected that a more accurate approximation of R is obtained, but experiments have shown that the fidelity of the approximation is query dependent (Zobel 1998). Furthermore, although recall-based tasks may be appropriate in some IR domains, the nature of web search lends itself more to precision-based tasks.

Despite these issues, recall-based measures are used extensively in evaluation (including for web search). They are reliable discriminators between different systems in large experiments, and provide a solid basis for quantitative analysis of effectiveness performance.

2.2 Precision-based measures

Precision-based measures calculate effectiveness scores solely as a function of the relevant documents encountered as the document ranking is traversed, and do not make use of R. Precision itself—the fraction of retrieved documents that are relevant—is normally measured at fixed depths that are significant for some reason, such as result page boundaries, or at numerically significant values such as recall percentiles. In addition, a range of other precision-based methods have been proposed, and are summarized in this section.

The discounted cumulative gain (DCG) metric allocates decaying weight contributions to relevant documents using a modified log-harmonic series (Järvelin and Kekäläinen 2002):

$$\mathrm{DCG}(\mathcal{R}, b) = \sum_{i=1}^{b} r_i + \sum_{i=b+1}^{|\mathcal{R}|} \frac{r_i}{\log_b i},$$

where b is a discounting factor, and is typically taken to be 2, and r_i can be real-valued, $0 \le r_i \le 1$, if graded relevance judgments are being used. Because there is no use made of R, DCG is purely precision-based. A drawback of DCG is that the log-harmonic sequence is divergent, and if DCG scores are required to fit the range zero to one, a non-constant scaling factor is required, computed as a function of $|\mathcal{R}|$, the length of the ranked document list.

Another way of bounding the range of DCG is provided by *normalized DCG* (NDCG) (Järvelin and Kekäläinen 2002), which is computed by dividing the DCG score by the maximum DCG score that could have been obtained at that point in the ranking, assuming that all of the documents relevant to the query appear at the top of the ranking (and in decreasing relevance order, should a multi-valued relevance scale be employed). Unlike DCG, or the scaled-to-one equivalent of DCG, NDCG is a recall-based measure because it requires knowledge of the relevant documents in order to compute the effectiveness score.

Rank-biased precision (RBP) (Moffat and Zobel 2008) is a related effectiveness metric based on a geometric weight sequence, and a corresponding user model in which the user is presumed to move sequentially through the ranked list of documents, stepping from one document to the next with probability p, or finishing the search at that point with probability $(1 - p)$. Unlike DCG which is divergent, the geometric sequence is convergent, and the RBP score of a ranking is always a number in the range [0, 1]:

$$\mathrm{RBP}(\mathcal{R}, p) = (1 - p) \sum_{i=1}^{|\mathcal{R}|} r_i p^{i-1},$$

where p is the persistence parameter, and is expressed as a value between 0 and 1. Low persistence values place greater emphasis on the relevance of documents near the top of the ranking, with the corresponding interpretation being that low-p (impatient) users are less likely to move further down the ranking than are high-p users, and so will judge the quality of the ranking according to what is encountered near the top of it. A p value of 0.8, equivalent to a user examining on average approximately five results, is a realistic value for web search (Park and Zhang 2007).

Rank-biased precision has the property that every rank position adds a predetermined amount to the evaluation score, and hence that as a run is explored to increasing depth the confirmed RBP score is non-decreasing. In addition, the geometric sequence (used for the RBP weights) is convergent. Hence, it is possible to calculate the set of score contributions of all unjudged documents through to depth infinity, and determine a *residual*, or error bound, which then provides an upper bound on the score assigned to that ranking. This can be done regardless of how many unjudged documents there are in the ranking, or where they occur. In the context of pooled evaluation, the residual error bounds can be used to guide the design of the experiment (Moffat and Zobel 2008), and also to determine which documents can most productively be judged if the experimental resource is limited (Moffat et al. 2007).

Sum of precisions (SP) is, like DCG, an unbounded metric that is not reliant on knowledge of R:

$$\mathrm{SP}(\mathcal{R}) = \sum_{i=1}^{|\mathcal{R}|} r_i \frac{\sum_{j=1}^{i} r_j}{i}.$$

Recent work has proposed a method for standardizing SP (as opposed to normalizing by R to get AP), using the mean and standard deviation calculated from a set of contributing systems (Webber et al. 2008). Based on the topic means and deviations, adjusted z scores are computed for the contributing systems, with an across-systems mean of 0.5 on every topic, and a uniform standard deviation. Standardized SP (sSP) is thus a metric that depends on a set of systems being jointly available at the time the judgments are being undertaken, as is the case for pooled relevance judgments; but not that those pooled judgments then be used as the basis for determining a reliable estimate of the value of R for each query. Experiments show that relatively few systems are needed to obtain reliable standardization factors, and that the standardization technique can be applied to other metrics (Webber et al. 2008).

Also worth noting is that reciprocal rank (and mean reciprocal rank across a set of topics, MRR) can also be included as a precision-based metric—RR is the value of P@$f_{s,t}$ where $f_{s,t}$ is the rank position of the first relevant document identified by system s on topic t.

We have commented several times that precision-based measures do not explicitly incorporate R in their formulation. The drawback of this independence is that the maximum realizable score for some topic calculated using them is unknown to the user. For example, P@10, the fraction of the top-ten documents that are relevant, cannot exceed $\min\{R/10, 1.0\}$ for a topic with R relevant documents. If is not known whether $R \geq 10$, the realizable limit for P@10 is also unknown. Similarly, the infinite sum used in the RBP metric means that an RBP score of 1.0 is impossible to attain unless $R \geq 1$, and $p = 1.0$. When R is unknown, the realizable maximum RBP score is also unknown. That is, while the value of R is not required as scores are computed for precision-based metrics, knowledge of R will be required if those scores are then to be normalized to fill the realizable range prior to them being compared.

2.3 User observation measures

In terms of numbers of users or queries issued, web search is the single largest application of information retrieval, and the queries and browsing actions of millions of users can be analyzed to provide considerable insight into their searching habits. For example, web queries are shorter in length than other types of queries; web users engage in relatively few query interactions before completing their search task; and web users are reluctant to browse past the first page of results (Jansen and Spink 2004).

Clickthroughs from user-viewed result pages can be used as an indicator of perceived interest in relation to result snippets, and as a source of implicit relevance feedback (Joachims 2002). In particular, eye-tracking experiments show that (Joachims et al. 2005, p. 157):

- "on average users tend to read the results from top to bottom";
- users "…view substantially more abstracts above than below the click";
- "a sharp drop occurs [in views] after link 10, as ten results are displayed per page"; and
- "…users click substantially more often on the first than on the second link, while they view the corresponding abstract with almost equal frequency."

Additionally, by switching the positions of results in the ranked list, Joachims et al. discovered that relevance does indeed affect users' clicking patterns, although this was subject to biases stemming from the overall quality of results and the preconceived quality of the search engine.

A framework for utilizing clickthroughs as relative relevance judgments for learning retrieval functions is presented by Agichtein et al. (2006). Clickthroughs can also be utilized as features in determining user intent (Lee et al. 2005; Teevan et al. 2008), and, more recently, as a query-independent measure of page importance in the web graph (Liu et al. 2008).

3 Decaying weight models for effectiveness

We now consider a more general way of defining effectiveness measures, and show how a wide range of current mechanisms can be described using a common framework.

3.1 Weighted relevance rankings

Effectiveness evaluation metrics seek to measure the utility of a particular ranking in a quantifiable and objective manner. In some situations, this implies a combination of precision-based and recall-based evaluation. On the other hand, there are also many situations in which the metric should quantify the user experience in some natural manner, based solely on the set of documents (or document summaries) that is presented to the user as the result of the search. The latter approach is particularly important in web searching, in which the pool of documents in the system is effectively infinite, and the perceived quality of the user experience is largely influenced by the documents presented in the first page or two of results. If the user finds the information they are after in a highly-ranked document they are probably satisfied, and the fact that the same information is available in the further 20 (or 200, or 2,000) relevant documents not shown in the first page of results has little effect on their perception of the quality of the retrieval system.

In such a metric, the benefit delivered by a relevant document needs to be tempered by knowledge of its position in the ranking. For example, the utility model implicit in DCG (taking $b = 2$) presumes that the first and second documents in the ranking are of equal importance to the user, that the third is approximately 2/3 as important as each of those two, that the fourth is half as important, that the one hundredth document is approximately 1/7 as important as the first two, and so on.

Generalizing the DCG and RBP approaches, we can thus postulate a family of effectiveness metrics based on a simple inner-product calculation of a relevance vector and a weighting vector:

$$\text{Effectiveness}(\mathcal{R}, \mathcal{W}) = \mathcal{R} \cdot \mathcal{W} = \sum_{i=1}^{|\mathcal{R}|} r_i w_i, \tag{1}$$

where

- $\mathcal{R} = \{r_1, r_2, \ldots\}$ is a (usually finite) relevance vector, as was presumed in the previous section; and
- $\mathcal{W} = \{w_1, w_2, \ldots\}$ is a (usually infinite) weighting vector in which w_i defines the relative importance (or utility) to the user of encountering a relevant document at position i in the ranking.

A large number of effectiveness metrics can be defined in this way, including P@k, DCG, and RBP. Common to all three is the presumption that $\mathcal{W}$ is non-increasing—that a relevant document at position i in the ranking does not convey more utility to the user than a document at some position $j < i$. In addition, note that the relevance values r_i are free to

take on values between zero and one; and that graded relevance evaluations are thus naturally supported.

If the further assumption that $\sum_{i=1}^{\infty} w_i = 1$ is added (which is the case for RBP, and also for DCG as specified in Eq. 2, below), then $\mathcal{W}$ can be interpreted as a probability distribution, and Eq. 1 is a statement that the effectiveness score for a ranking is the expectation that a randomly selected document (according to $\mathcal{W}$, a probability distribution over documents) is relevant (according to $\mathcal{R}$). Interpreting weighting vectors probabilistically in this manner allows us to compare individual representations of user behavior on a uniform scale, and also to examine the weight vectors themselves, rather than the effectiveness scores they give rise to.

From this point onward, we use the term *weighting model* to describe a vector $\mathcal{W}$ that defines an effectiveness metric in the way described by Eq. 1. In addition, when $\mathcal{W}$ sums to one, we will regard $\mathcal{W}$ as being a probability distribution over the documents in the ranking that gives the likelihood of each document being selected by a user.

3.2 Static weighting models

We now examine a range of common weighting models that might be used as the basis for effectiveness metrics.

3.2.1 Discrete uniform distribution

This distribution provides equal probability for a defined range and zero probability outside of that range:

$$\mathcal{W}_{U,k}(i) = \begin{cases} 1/k & \text{for } 1 \le i \le k \\ 0 & \text{otherwise.} \end{cases}$$

As an example, suppose that for some query a retrieval system has provided a ranked list of results in which relevant documents occur at ranks {2, 5, 6, 13, 20} (that is, $\mathcal{R} = \{0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1\}$). When $k = 10$, evaluation using the discrete uniform distribution as the weighting model gives rise to P@10, the fraction of relevant documents observed in the top ten:

$$\text{Effectiveness}(\mathcal{R}, \mathcal{W}_{U,10}) = \sum_{i=1}^{10} \frac{r_i}{10} = 0.3.$$

When $k = |\mathcal{R}| = 20$, $\text{Effectiveness}(\mathcal{R}, \mathcal{W}_{U,20})$ decreases to 0.25.

3.2.2 Zipf distribution

The Zipf distribution represents a general power law with a long tail, and is observed in many aspects of large document collections. If the range of values is limited to $k \le |\mathcal{R}|$, the scaled weight associated with rank i is given by:

$$\mathcal{W}_{Z,\beta,k}(i) = \begin{cases} i^{-\beta}/S & \text{for } 1 \le i \le k \\ 0 & \text{otherwise.} \end{cases}$$

where β characterizes the distribution, and $S = \sum_{i=1}^{k} i^{-\beta}$ is a scaling factor that makes the (truncated) distribution sum to one. For the same example ranking as was used earlier, a Zipf weighting model with $\beta = 1$ and $k = |\mathcal{R}| = 20$ gives:

$$\text{Effectiveness}(\mathcal{R}, \mathcal{W}_{Z,1,20}) = \frac{1}{3.598}\sum_{i=1}^{20}\frac{r_i}{i},$$

which is $(0.500 + 0.200 + 0.167 + 0.077 + 0.050)/3.598 = 0.276$. Note that in this example, if s were taken to be 100 instead of 20, the score would drop to 0.191, and an aggregate weight of 0.297 would be indeterminate (as a residual)—meaning that the final effectiveness score might be as large as 0.488, if all of the unjudged documents between rank 21 and rank 100 were in fact relevant.

3.2.3 Poisson distribution

The Poisson distribution is generally used to express the number of events occurring within a specific period of time, but its monotonicity when $0 \le \alpha \le 1$ means that it can also be used to represent the probability of observing the i th document in the ranking:

$$\mathcal{W}_{P,\alpha}(i) = \frac{\alpha^{(i-1)}e^{-\alpha}}{(i-1)!},$$

where α characterizes the distribution. When $\alpha = 1$ the Poisson distribution reduces to:

$$\mathcal{W}_{P,1}(i) = \frac{e^{-1}}{(i-1)!},$$

implying that $\mathcal{W}_{P,1}(1) = \mathcal{W}_{P,1}(2)$. It has been observed that the probability of examining the first and second result snippets is the same (Joachims et al. 2005; Turpin et al. 2006), and the Poisson weighting model might thus be useful when modelling those situations.

3.2.4 Geometric distribution

The geometric distribution models the probability of $i - 1$ successes before a failure on the ith attempt:

$$\mathcal{W}_{G,p}(i) = p^{i-1}(1-p),$$

where p is the probability of success. This is the distribution used in the RBP metric, and its behavior as a metric is documented by Moffat and Zobel (2008).

3.2.5 Log-harmonic distribution

The DCG metric utilizes a log-harmonic distribution, modified to assign equal probability to all ranks up to rank b:

$$\mathcal{W}_{L,b,k}(i) = \begin{cases} 1/S & \text{for } i \le b \\ 1/(S\log_b i) & b \le i \le k \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $S = b + \sum_{i=b+1}^{k}(1/\log_b i)$ is a fixed scaling factor that yields weights that sum to one, and does not affect the relative values of scores provided that all topics and systems are evaluated using the same run depth $k \le |\mathcal{R}|$. Note also that this linear scaling process is not the normalization approach used by Järvelin and Kekäläinen (2002) in the definition of

normalized discounted cumulative gain, and that DCG and NDCG do not necessarily generate the same system orderings, even when evaluated to the same depth.

3.3 Distributions for complex measures

Recall-based measures can be cast into the same framework if we allow the weight vector $\mathcal{W}$ to incorporate one or both of:

- The number of relevant documents for this query, R; and/or
- The positions of relevant documents within the ranking, as expressed by the vector $\mathcal{R}$.

With these additional factors allowed, the corresponding weighting vectors for AP, Q-measure, and SP are given by:

$$\mathcal{W}_{\mathrm{AP},\mathcal{R}}(i) = \frac{\sum_{j=1}^{i} r_j}{i \cdot R}, \tag{3}$$

$$\mathcal{W}_{\mathrm{Q-measure},\mathcal{R}}(i) = \frac{2\sum_{j=1}^{i} r_j}{(i + \min(i,R)) \cdot R},$$

and

$$\mathcal{W}_{\mathrm{SP},\mathcal{R}}(i) = \frac{\sum_{j=1}^{i} r_j}{i}.$$

The weighting model for BPref is slightly different, as it requires all unjudged documents to be removed from the relevance vector prior to calculation. Since unjudged documents make no contribution, all of R, N and k remain unaltered, and r_i is unaltered assuming an unjudged document has relevance 0. To make the weighting model consistent with our unmodified relevance vectors, we introduce a variable J_i to represent the number of judged documents up to rank i:

$$\mathcal{W}_{\mathrm{BPref},\mathcal{R}}(i) = \frac{1}{R} - \frac{\min\left(k + R, J_i - \sum_{j=1}^{i} r_j\right)}{\min(k + R, N) \cdot R}.$$

It is also clear that a weighting function $\mathcal{W}_{\mathrm{RR}}$ for the reciprocal rank measure can be specified as:

$$\mathcal{W}_{\mathrm{RR},\mathcal{R}}(i) = \begin{cases} 1/i & \text{if } i = \min\{j \mid r_j \in \mathcal{R}, r_j = 1\} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

All of these are rather contrived expressions. For example, the definition of $\mathcal{W}_{\mathrm{AP},\mathcal{R}}$ in Eq. 3 suggests that the probability of examining the i th document is zero if there are no relevant documents within the first i of the ranking; and also means that $\mathcal{W}_{\mathrm{AP},\mathcal{R}}(i)$ is not necessarily non-increasing in i. Nor, as defined, is $\mathcal{W}_{\mathrm{AP},\mathcal{R}}$ a probability distribution. Nevertheless, by casting AP into this framework, it makes it clear that it can be computed as the inner-product of a relevance vector and a weight vector, and sets the scene for $\mathcal{W}(i)$ to be estimated across a whole set of rankings, rather than as something that is specific to a single ranking.

3.4 Generating background values

If we are given a set of TREC-style rankings, generated by a set of s IR systems against a set of t topics, and also have relevance judgments for those topics, then a total of $s \cdot t$ relevance

vectors can be computed. Those $s \cdot t$ runs can be used as training information to estimate (for that set of systems, and that set of queries) a value for $\mathcal{W}_{\rm AP}(i)$ as it is defined in Eq. 3, by averaging across the systems (using the correct value of R for each topic and the $\sum(r_j/i)$ values through the rankings for that topic), and then averaging across the topics. That is, if training data is available, it is possible to estimate the distribution $\mathcal{W}_{\rm AP}$ independently of the relevance vector for any particular system/topic combination, by computing:

$$\mathcal{W}_{\rm AP}(i) = \frac{1}{s \cdot t \cdot i} \sum_{j=1}^{s} \sum_{\ell=1}^{t} \sum_{m=1}^{i} \frac{r_{j,\ell,m}}{R_\ell},$$

where R_ℓ is the number of relevant documents associated with the ℓth topic, and $r_{j,\ell,m}$ is the relevance (or not) of the mth document in the ranking generated by system j on topic ℓ.

An inferred weight distribution for reciprocal rank can also be constructed based on Eq. 4, by aggregating across the set of system-topic runs the rank locations $f_{j,\ell}$ of the first relevant document identified by system j on topic ℓ:

$$\mathcal{W}_{\rm RR}(i) = \frac{1}{s \cdot t \cdot i} \cdot \sum_{j=1}^{s} \sum_{\ell=1}^{t} |\{f_{j,\ell} \mid f_{j,\ell} = i\}|.$$

It is similarly possible to calculate empirical weighting vectors for BPref, Q-measure, and SP. All of these inferred distributions are explored in the experiments reported below. Like all TREC outcomes, the weighting vectors are not portable from one experimental regime to another, and should only be used in the context of the sets of systems and topics from which they were generated. Nevertheless, the existence of empirical $\mathcal{W}_{\rm AP}$ and $\mathcal{W}_{\rm RR}$ distributions allows the AP and RR measures to be directly compared with methods such as RBP and DCG, which use static weighting vectors, within a particular TREC dataset. We will return to this point later.

4 Predicting web user observations

We have presented a range of weighting models corresponding to established evaluation metrics. Logic or rhetoric might then be used to argue that one weighting scheme is more plausible than another. Instead of those approaches, we turn to experiment, and ask, based on the evidence available in a large search engine clickthrough log, whether the pattern of user clickthroughs observed in the log provides support for any particular weighting model.

4.1 User gap distributions

It has already been noted that a weighting model can be interpreted as being a probability distribution over documents, in which case the effectiveness score is the expected relevance. An obvious question to ask is then, what documents do users examine when looking at ranked lists of documents? And, if we can create an *observation model* that establishes the probability that a user looks at a particular item in a ranking, what—if any—correspondence is there between the observation model and the various weighting models that provide the basis for different effectiveness metrics?

One simple way of forming an observation model is to note the rank positions of users' clickthroughs, data that is readily available in the web search context. In selecting a snippet to click on, the user has certainly observed that document. The user will also have observed

other snippets that were not selected for clickthrough, and these non-clicked snippets should also be included in the user observation model. The problem is that, except with the use of specialized laboratory equipment, there is no record of these passive interactions with snippets.

We thus make a number of critical assumptions about user behavior that allow the density of a users' clickthrough distribution to be analyzed to infer information about snippets that were observed, but not clicked, based on the findings by Joachims et al. (2005), discussed in Sect. 2:

- Users observe the results snippet listing in rank (presentation) order;
- A clickthrough at rank n implies that the user certainly observed all ranks 1, …, n;
- Users are more inclined to stop observing at result page boundaries than they are within a result page; and
- Users may also observe one or more snippets past the last one they clicked.

In particular, we use the *density* of the clickthrough distribution for each user in order to estimate the number of documents that were observed beyond the last clicked one and to allow accurate estimates of the observation model. We express the *click gap distribution* for a user as $P(\text{gap} = i \mid u, q)$, the probability that user u observes exactly i consecutive snippets in the result ranking for query q without clicking on any of them. The key assumptions then mean that, aggregated over all of the queries they issued, this user's probability of observing a document j ranks beyond the last observed clickthrough (or of observing the document in rank position j if there are no clickthroughs) can be expressed as

$$P(\text{gap} \geq j \mid u, q) = \sum_{i=j}^{\infty} P(\text{gap} = i \mid u, q).$$

This cumulative distribution can be averaged out over all observed queries for user u,

$$P(\text{gap} \geq j \mid u) = \text{mean}_q P(\text{gap} \geq j \mid u, q),$$

and then appended to the known last clickthrough position for each query issued by this user, to obtain the probability that they observed a result at any given rank:

$$P(\text{observed} = i \mid u, q) = \begin{cases} 1 & \text{for} \quad i \leq \text{LC}(u, q) \\ P(\text{gap} \geq (i - \text{LC}(u, q)) \mid u) & \text{otherwise,} \end{cases}$$

where $\text{LC}(u, q)$ is the rank of the last clickthrough in query q as issued by user u, and is zero if there were no clickthroughs for query q by this user.

The probability $P(\text{observed} = i \mid u, q)$ can then be normalized to form the *observation model* of user u on this query, and if the user issued multiple queries, the individual observation models for their queries can be combined to compute a overall observation model $P(\text{observed} = i \mid u)$ for the user. The observation models could then be averaged across all sampled users to obtain a pooled distribution, and then compared with effectiveness weighting distributions. However, we insert two further steps before finalizing the user observation models, to account for potentially biased samples, and to obtain a better representation of user behavior. The next two subsections discuss these adjustments.

4.2 Smoothing observation predictions

Despite their large volume, most query logs contain relatively few samples for each user, and the observed gaps for any single user form a sparse distribution. To adjust for this

problem, we incorporate a smoothing step, to balance the observed gap distribution for an individual user against the clickthroughs issued by all users:

$$P_{\text{smooth}}(\text{gap} \geq i \mid u) = \alpha_u P(\text{gap} \geq i \mid u) + (1 - \alpha_u) P(\text{gap} \geq i \mid U),$$

where $P(\text{gap} \geq i \mid U)$ is the click gap distribution computed across all users, and α_u is the smoothing parameter for user u.

Although it is possible select the value of α_u manually (Jelinek-Mercer smoothing), we opted to use Dirichlet smoothing, which selects α_u based on the number of samples available for each user:

$$\alpha_u = \frac{\text{CT}(u)}{\text{CT}(u) + \mu},$$

where $\text{CT}(u)$ is the number of clickthroughs observed for user u, and $\mu \in \mathbb{R}^+$ is a constant. Less smoothing is applied if a large number of samples are available.

4.3 Page boundary handling

Another important factor that needs to be compensated for is that the predicted observations will often span page boundaries, which typically occur every ten results for web search engines. Because the observation model for a user aims to represent the manner in which that user views results in a homogeneous ranked list rather than a segmented one, page boundary effects need to be identified in the query log data, and allowed for in the subsequent analysis.

Note that the drop in observation probability only applies for predicted observations past the last recorded clickthrough. For example, if the last recorded clickthrough for a user/query combination is at rank 15, the inferred probability of observing ranks 1–15 is still one. Similarly, the estimated probability of observing ranks 16–20 is unchanged relative to the gap-based estimation alone. But the probability of the user observing ranks 21–30 needs to be discounted, to acknowledge that users are likely to discontinue observing at page boundaries; and ranks 31–40 need to be doubly discounted.

The rate of "observation leakage" at page boundaries can be estimated from query and clickthrough logs. Let ℓ_p represent the number of issued queries for which the last clickthrough $\text{LC}(u, q)$ was recorded in page $p \in \{1, 2 \ldots\}$, taken across all users; and define $\text{Page}(i)$ as the page on which the snippet at rank i occurs. When there are n results per page, $\text{Page}(i) = \lceil i/n \rceil$; in the context of web retrieval, $n = 10$ is a suitable value for a wide variety of web search engines.

The observation model then becomes:

$$P_{\text{boundary}}(\text{observed} = i \mid u, q) = \begin{cases} 1 & \text{for } i \leq \text{LC}(u, q) \\ P_{\text{smooth}}(\text{gap} \geq (i - \text{LC}(u, q)) \mid u) \cdot B(i) & \text{otherwise,} \end{cases}$$

where

$$b(i) = \sum_{p=i}^{\infty} \ell_p,$$

and

$$B(i) = \frac{b(\text{Page}(i))}{b(\text{Page}(\text{LC}(u,q)))}.$$

As required, observation probabilities for snippets up until the last clickthrough, as well as snippets past the last clickthrough but still occurring on the same results page, remain unaltered compared to the original observation model. All other observations receive a reduction for each page boundary crossed.

4.4 An example

We now illustrate the process for calculating observation models with an example. Suppose that a user has clicked on ranks 1, 5 and 6 for query A, with gaps of 1, 4, and 1; and then clicked on ranks 2, 4, and 10 for query B, with gaps of 2, 2, and 6. The second column in Table 1 shows the unmodified click gap distribution for this example user. The third column in the table shows the click gap distribution for the example user, formed by cumulatively summing the previous probabilities. The small number of clickthroughs for the user gives rise to zero samples for gaps of 3, 5, and all lengths greater than 6.

We now incorporate Dirichlet smoothing with $\mu = 2$, and hence $\alpha_u = 6/(6 + 2) = 0.75$, so that 25% percent of $P_{\text{smooth}}(g \geq i \mid u)$ is contributed from global observation gap probabilities. The last column in Table 1 shows the results of applying smoothing, resulting in a more rounded click gap distribution when using the hypothetical global click gap distribution shown in column four. Using this and the last clickthrough information, we can now obtain $P(\text{observed} = i \mid q)$ for each query issued by the user, and apply page boundary reductions to the applicable observations past the last clickthrough. For this example, assuming there were 1,000 clickthroughs finishing on the first page of results out of a global 1,860, predicted observations crossing the first page boundary would have their probabilities multiplied by 860/1,860.

Since the last clickthrough for query A is at rank 6 and $P_{\text{smooth}}(\text{gap} \geq i \mid u)$ is non-zero for gaps of up to 7, ranks 7 through to 13 obtain predicted observation probabilities as part of $P(\text{observed} = i \mid u, q)$. Because we defined the page boundaries to occur every 10 results, ranks 11–13 have page boundary reductions applied. Query B is handled similarly. Finally, after obtaining $P_{\text{boundary}}(\text{observed} = i \mid u, q)$ for queries A and B, we average to obtain $P(\text{observed} = i \mid u)$. This distribution is shown in Fig. 1, which also includes the observation models for both queries after page boundary adjustment.

5 Observation modelling with query logs

To obtain an observation model based on real world search behavior, we used an MSN query log, containing approximately 5 million user search sessions and 12 million clickthroughs, collected over a one month period (May 2006) from US users of Microsoft's MSN web search service[1] Because this data is anonymized in the sense of there being short-term query identifiers, but not longer-term user identifiers, we have no choice but to assume in our calculations that each session corresponds to a unique user. Using the computed observation model derived from the data, we are able to compare effectiveness metric weighting models and actual user behavior.

[1] http://search.msn.com, now replaced by bing.

Table 1 Click gap distribution for the example user, with an assumed global gap distribution, and smoothed values using Dirichlet smoothing with $\mu = 2$ (and hence $\alpha_u = 0.75$)

i	$P(\text{gap} = i \mid u)$	$P(\text{gap} \geq i \mid u)$	$P(\text{gap} \geq i \mid U)$	$P_{\text{smooth}}(\text{gap} \geq i \mid u)$
1	2/6	6/6	10/10	1.000
2	2/6	4/6	9/10	0.725
3	0/6	2/6	8/10	0.450
4	1/6	2/6	5/10	0.375
5	0/6	1/6	3/10	0.200
6	1/6	1/6	2/10	0.175
7	0/6	0/6	1/10	0.025

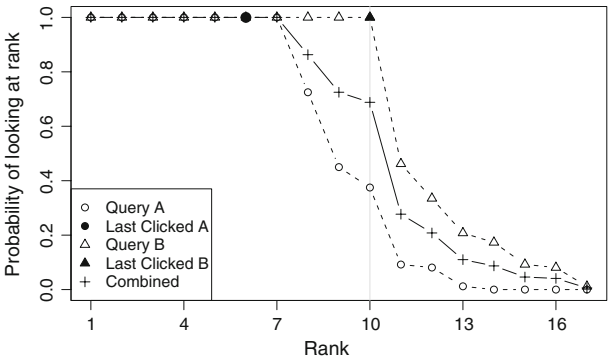

Fig. 1 Calculation of overall observation model $P(\text{overlap} = i \mid u)$ for the example user shown in Table 1, using the mean from the query A and B models, each with prior smoothing and page boundary reductions applied

5.1 Click gap distribution

From the user sessions and clickthrough data, we generated a click gap distribution for each query in the log, and then aggregated them to form a global cumulative distribution $P(\text{gap} \geq i \mid U)$, as shown in Fig. 2. Small click gaps have a very high probability—fully 65% of the click gaps observed correspond to a gap of one, and another 14% to a gap of two—and this means that the cumulative distribution is similarly focussed on low ranks. Note that $P(\text{gap} \geq 1 \mid U) = 1.0$ by definition, which implies that in our model users always

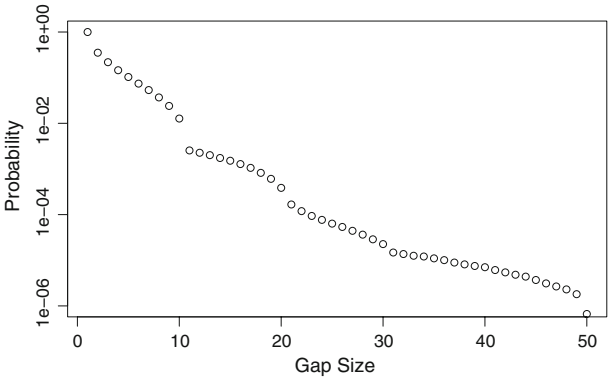

Fig. 2 Global click gap distribution $P(\text{gap} \geq i \mid U)$ across all users in the MSN query log. Given that a measured click gap exists, it must be at least one, hence the 1.0 value for $i = 1$. In turn, this implies that users always look at least one result past their last clickthrough (except when they click on the last snippet in the page)

observe at least one more document after their last clickthrough, except when they have clicked on the last snippet of a results page. In particular, a user that issues no clickthroughs is still assumed to have always observed the highest ranked snippet.

Using the calculated cumulative global click gap distribution, observation models were calculated for each user, and the set of resultant observation models averaged to obtain a single curve. Figure 3 shows the original clickthrough distribution for the MSN log along with observation models computed using Dirichlet smoothing with $\mu \in \{0, 5\}$. In constructing this graph, each set of probabilities was normalized to a total of 1.0, so that what is plotted for each of the three data sets can be thought of as being on the same scale, and is now of a form that can be later compared with the effectiveness metric distributions discussed in Sect. 3. Because the observation models have more mass at higher ranks, the normalized click probability at rank one is greater than the normalized observation probabilities.

Since clickthroughs in the top ten ranked results account for over 99% of all clickthroughs, there is a significant reduction in observation probability associated with crossing the first page boundary, shown in the first row of Table 2. Subsequent page boundaries incur smaller penalties, and once a user has crossed the second page boundary, they are quite likely to continue into the fourth and then fifth results pages. There are no samples in the log beyond the sixth page boundary, and we assign a constant reducing factor of $b(p) = 10^{-8}$ for those boundaries.

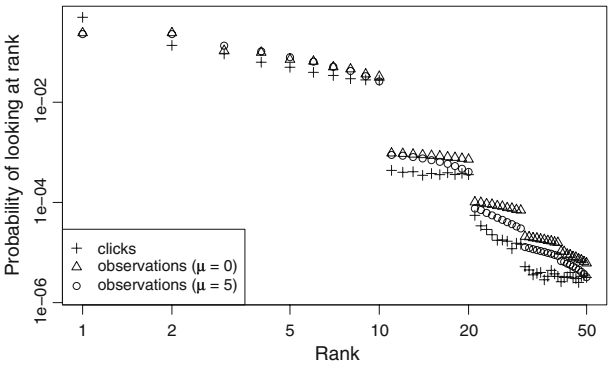

Fig. 3 Generated observation models with Dirichlet smoothing compared to raw clickthrough distribution for MSN dataset. Clickthroughs are strongly biased towards the first results in the ranking, and have sharper drop-offs when moving to later ranked results within each page compared to generated models. Smoothing appears to have little impact on the overall observation model

Table 2 Page boundary reduction values $b(p + 1)/b(p)$, which indicate the observed conditional probability (as evidenced by clickthrough activity) of the user stepping from page p to page $p + 1$ in the Microsoft query log

p	l_p	$b(p)$	$b(p + 1)/b(p)$
1	8,793,770	8,831,275	0.0042
2	35,014	37,505	0.0664
3	2,001	2,491	0.1967
4	224	490	0.5429
5	265	266	0.0038
6	1	1	0.0000

Clickthroughs within the top ten results account for over 99% of all clicks

The calculated observation models depicted in Fig. 3 show a smooth decline within each of the results pages, with sharp drops at the page boundaries. The large number of queries with a single rank-one clickthrough, combined with the cumulative method for calculating click gap distributions, means that ranks one and two end up with equal observation probability. This coincides with findings from eyetracking experiments (Joachims et al. 2005), as was discussed in Sect. 2. Page boundaries have been preserved in the modelling process, and the overall observation model is monotonically decreasing. Except at rank one, the normalized observation probability is higher than the normalized click probability, as expected.

Use of $\mu = 5$ in the Dirichlet smoothing gives rise to smaller normalized probabilities than $\mu = 0$, except at ranks one and two, and a more pronounced decline within each page. The majority of the results below make use of $\mu = 5$.

5.2 Parameter estimation for static distributions

Each of the static weighting models described in Sect. 3 was compared to the observation models, to determine in each case a "best" parameter for the static distribution, plus a goodness-of-fit coefficient to indicate how well the distributions match each other. The results of this evaluation are shown in Table 3. The Kullback–Leibler (KL) divergence is a non-commutative measurement of the extent to which an approximate probability distribution (the four different effectiveness models in the rows of the table) departs from a reference distribution (the three different observation models in the columns), and represents the additional cost of entropy coding the reference models using the probabilities given by the approximate distribution rather than by their own probability distribution, measured in "nats per symbol", where one nat is the equivalent of approximately 1.44 bits. When the two distributions are identical, the KL divergence is zero. To create Table 3, a search over the parameter space of each candidate approximate distribution was used to determine the parameter value that gave the lowest KL divergence score. Note that the log-harmonic distribution derived from the DCG metric is unlike the other distributions, in that the parameter must be an integer. Figure 4 depicts the fitted distributions, along with the observation model formed using Dirichlet smoothing $\mu = 5$.

The KL best fits and fitted parameters are similar across all three observation models shown, and the choice of μ plays a minor role only. The geometric distribution clearly offers the best KL fit for all instances of μ, with the lowest KL value of

Table 3 Kullback–Leibler divergence scores and fitted parameter values for weighting models of static distributions, where a smaller KL score indicates a better fit to the reference distribution listed as the column heading

Distribution	$\mu = 0$		$\mu = 1$		$\mu = 5$	
	KL div.	Param.	KL div.	Param.	KL div.	Param.
Geometric	**0.051**	0.731	**0.043**	0.729	**0.040**	0.727
Poisson	0.355	3.719	0.303	3.690	0.268	3.661
Zipf	0.272	1.455	0.281	1.451	0.289	1.450
Log-harmonic	0.941	2	0.946	2	0.953	2

When $\mu = 0$ the observation model is formed without performing Dirichlet smoothing. The parameter in the log-harmonic distribution is restricted to be an integer. KL divergence of 0.1 or less (in bold) represent close agreement between the two distributions

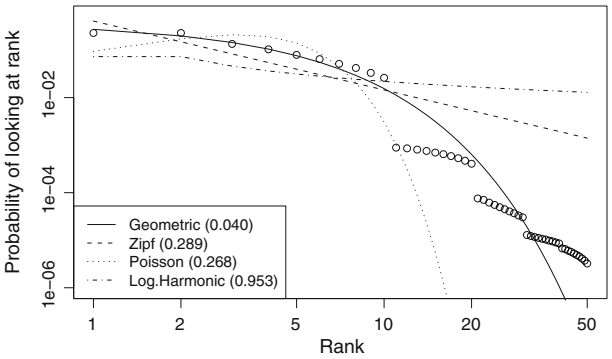

Fig. 4 Predicted observation model for the MSN query/click log (shown as circles, calculated with $\mu = 5$) and KL-best probability distributions according to different effectiveness metric weighting schemes. The parenthetical numbers are the KL divergence scores, and the parameter values for each of the distributions are as listed in the last column of Table 3. The geometric weighting scheme yields the closest fit to the observation model, with parameter $p = 0.73$

approximately 0.04, representing a high level of agreement between the reference (observation) and approximate (effectiveness metric-based) probability distributions. The Zipfian and Poisson distributions also perform moderately well, with divergences of around 0.3. On the other hand, the log-harmonic distribution (with evaluation depth $k = 50$) gives rise, even at best, to low quality approximations, with KL divergences close to 1. This distribution greatly underestimates the observation probabilities for highly ranked documents, and similarly overestimates the observation probabilities for ranks greater than ten.

The geometric distribution gives rise to a best-fit p value of approximately 0.73, implying that the average user observes $1/0.73 = 3.7$ documents. This is close to the value of $p = 0.8$ suggested by previous work with the same query/click log (Park and Zhang 2007).

5.3 Judgment-based weightings for complex distributions

We employed three TREC tracks to generate weighting models for the complex distributions, using the approach described in Sect. 4: the TREC9 Web Track (Hawking 2000), the TREC2001 Web Track (Craswell and Hawking 2001), and the TREC2004 Terabyte Track (Clarke et al. 2004). The two Web Tracks reflect the emphasis of this paper on web querying, and the fact that the observation models were generated using a web query log. We also included the Terabyte Track in these experiments to confirm that using web track data to generate weighting models for a web-based observation model should provide a better fit than using some other track.

Table 4 shows KL divergences when the empirical effectiveness weighting models derived from the three sets of TREC relevance judgments were compared with the observation distribution (using $\mu = 5$) calculated using the methodology of Sec. 4. The KL divergence scores listed in Table 4 can be directly compared with the $\mu = 5$ column in Table 3. With the exception of BPref, these measures do not involve further parameters. In preliminary experiments not described here, BPref was trialled with $k = 1$ and $k = 100$ as well as using the recommended value of $k = 10$, with $k = 1$ giving similar results to $k = 10$, and $k = 100$ being markedly inferior.

Figure 5 presents a visual representation of the relevance-derived weighting models for the TREC9 Web Track, and compares them to the click-derived observation model

Table 4 Kullback–Leibler divergence scores for TREC-estimated effectiveness weighting distributions, compared to the click-derived observation model calculated using Dirichlet smoothing and $\mu = 5$

Distribution	KL divergence compared with observation model, $\mu = 5$		
	TREC9 Web	TREC2001 Web	TREC2004 TB
$\mathcal{W}_{\text{RR}}$	0.675	**0.722**	**0.968**
$\mathcal{W}_{\text{AP}}$	1.481	1.642	2.078
$\mathcal{W}_{\text{SP}}$	1.913	1.978	2.228
$\mathcal{W}_{\text{BPref,k=10}}$	**0.569**	0.805	1.725
$\mathcal{W}_{\text{Q-measure}}$	1.735	1.852	2.116

The best-fitting alternative in each column (of the listed options) is highlighted; note that Table 3 offers better-matched distributions

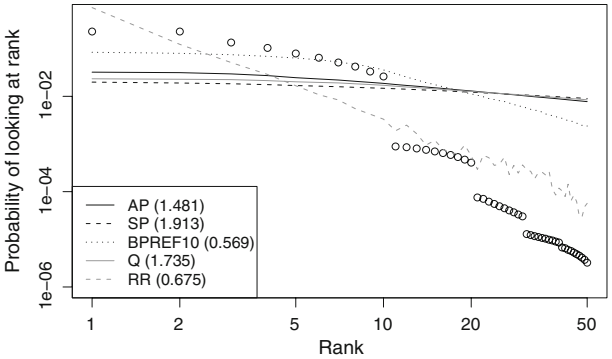

Fig. 5 Approximated weighting models for complex measures, using TREC9 Web Track topics and judgments to infer effectiveness weighting distributions for a range of complex effectiveness measures. The circles show the click-derived observation model weightings using Dirichlet smoothing with $\mu = 5$, as was also the case in Fig. 4. All of the relevance-based models shown in this graph underestimate the click-derived observational probabilities for ranks up to ten, and overestimate for lower ranks

computed using $\mu = 5$. It is clear from Table 4 and Fig. 5 that the static distributions shown in Fig. 4 provide higher fidelity approximations to the click-derived observation model than do the alternative weighting vectors derived empirically from AP, BPref, RR, and so on. Indeed, the complex distributions are all similar in nature, assigning comparable weightings. This characteristic is a consequence of the fact that all of the complex weighting models incorporate scoring methods that calculate precision at the rank positions at which relevant documents occur. Since the same query/run data is used for approximation in each method, they receive essentially the same inputs.

The nature of each experimental track also plays a part in determining the quality of fit. As is documented in Table 4, the TREC9 Web Track provided the best-fitting judgment-derived models, and the TREC 2004 Terabyte Track the worst, reflecting the fact that the observation model against which the various weighting schemes are being compared is derived from web data.

Finally in this section, having already determined that RBP with parameter $p = 0.73$ is a close fit to the observation model, we also identified the values of p that led to the closest match between RBP's geometric distribution, and each of $\mathcal{W}_{\text{AP}}$ and $\mathcal{W}_{\text{RR}}$, for each of the three TREC collections used. The results are shown in Table 5. When seeking to match

Table 5 Kullback–Leibler divergences, and best-fit parameter values, for RBP for each of two inferred probability distributions on the three TREC datasets used

Distribution	TREC9 Web		TREC2001 Web		TREC2004 TB	
	KL div.	p	KL div.	p	KL div.	p
$\mathcal{W}_{\mathrm{RR}}$	0.204	0.447	0.237	0.483	0.185	0.353
$\mathcal{W}_{\mathrm{AP}}$	0.150	0.951	0.129	0.949	0.231	0.957

Values of $p \approx 0.95$ are a good fit to $\mathcal{W}_{\mathrm{AP}}$, and values of $p \approx 0.5$ are a good fit to $\mathcal{W}_{\mathrm{RR}}$

RBP to the inferred AP weighting distribution, values of p of approximately 0.95 should be used. On the other hand, the inferred RR distribution is much more heavily weighted towards the early part of the ranking, and the best-fit values of p are around 0.5. In all cases, the low KL divergence values indicate a good quality of fit between the two probability distributions.

5.4 System rank correlations with the observation model

Our purpose throughout this paper has been to determine what evidence—if any—can be extracted from query/click log data in support of different retrieval effectiveness metrics, with an emphasis on those that can be expressed as an inner-product of a relevance vector and a unit-sum weighting vector. The litmus test of any effectiveness metric is how it scores systems, so that systems that are perceived (by users) to be "good" are scored more highly than systems that are perceived to be "not as good".

To complete our experiments, we thus applied all of the weighting models to compute overall system scores for the TREC9 Web Track data. The weightings included in this last round of experimentation included all of the static methods listed in Sect. 2; the judgment-derived distributions for AP, BPref, RR, and so on, as described in Sects. 3 and 4; the distribution of clickthroughs; and the click-derived observation model described in Sect. 3. Each of these normalized weight distributions can be used in the weighted-precision inner-product effectiveness metric defined in Eq. 1, and numeric values computed for TREC systems based on the TREC relevance judgments.

Then, once the set of TREC systems has been scored by a particular effectiveness metric, they can be ordered by score, and that metrics' system ordering compared with any other system ordering—including the ordering generated by any conventional metric such as MAP or MRR—to determine a correlation coefficient. Every pair of effectiveness metrics can be compared in this way. To compute the strength of each correlation, Kendall's τ (Kendall and Gibbons 1990) was used, yielding values between -1 (reverse ordering) and $+1$ (identical ordering).

A subset of the computed correlation values is shown in Table 6. The four columns of the table reflect the use of four empirical weighting distributions: two derived from the Microsoft query log, $\mathcal{W}_{\mathrm{click}}$ and $\mathcal{W}_{\mathrm{observation},\mu=5}$; and two derived from the aggregate of the TREC9 topics and systems (assuming knowledge of the relevance judgments), $\mathcal{W}_{\mathrm{AP}}$ and $\mathcal{W}_{\mathrm{RR}}$. The rows of the table reflect a range of standard effectiveness metrics, including RBP, which itself uses a decreasing weight vector in Eq. 1. With the exception of MRR, P@2 and RBP with $p = 0.5$, all of which are top-dominant metrics, the raw clickthrough data has lower correlation values than does the observation model, an outcome that validates the assumptions that led to the observation model. The click distribution and MRR-

Table 6 Kendall's τ correlation scores for overall system rankings for selected pairs of effectiveness weighting schemes, using the $t = 50$ topics and $s = 105$ systems of the TREC9 Web Track, with correlations greater than 0.9 highlighted in bold

Metric	From query log		From TREC judgments	
	Clicks	Obs., $\mu = 5$	$\mathcal{W}_{\text{AP}}$	$\mathcal{W}_{\text{RR}}$
MRR	**0.922**	0.853	0.710	**0.930**
P@2	**0.921**	0.893	0.734	**0.906**
P@10	0.851	**0.917**	0.820	0.794
RBP, $p = 0.5$	**0.963**	**0.908**	0.744	**0.930**
RBP, $p = 0.73$	**0.914**	**0.987**	0.795	0.859
RBP, $p = 0.95$	0.819	0.874	0.884	0.769
MAP	0.775	0.809	0.874	0.747
BPref, $k = 10$	0.775	0.815	0.836	0.745
DCG, $k = 1{,}000$	0.722	0.749	**0.932**	0.694
NDCG	0.761	0.782	0.898	0.736

Each of the rows represents a standard effectiveness metric, applied to the pool of systems to generate an ordering (from best to worst) of the systems. The four columns represents the same systems ordered by inner-product precision, based on an empirical weighting vector derived from either the MSN query log, or from the aggregate behavior of the TREC9 systems. Each entry in the table is a correlation score, showing the extent to which the row and column system orderings are similar, with 1.0 representing "identical" and 0 representing "no correlation". The observation model (using $\mu = 5$) shows a high correlation (>0.75) for all of the standard effectiveness measures included in the table, and extremely high correlation (>0.98) with rank-biased precision, $p = 0.73$

derived weightings are both heavily focussed on the first few positions in the ranking; whereas the other two approaches give weighting further into the ranking.

The observation model shows a relatively high degree of correlation with all the tested evaluation measures, with RBP being the standout, and P@10 also being strong. At the other end of the range, DCG generates the weakest correlation when both the both the observation model and the click distribution are used as effectiveness weighting vectors.

In part of the full pairwise correlation matrix not shown in Table 6, all of $\mathcal{W}_{\text{AP}}$, $\mathcal{W}_{\text{Bpref}}$, and $\mathcal{W}_{\text{Q-measure}}$ yield correlations of around 0.9 with each other, indicating that they order the TREC9 systems into quite similar arrangements.

6 Recent related work

Methods for quantifying retrieval effectiveness have been explored for many years. The problems we are particularly concerned with in this paper arise because of the large size of current test collections, and the impossibility of fully judging them, even against small topic sets. Zobel (1998) showed that even relatively deep pooling was unlikely to discover all relevant documents, and user studies such as that of Joachims et al. (2005) showed that users of web search systems were unlikely to pursue their interest deep in a ranking. These two observations have led to the recent interest in metrics that are heavily top-weighted, and in which what is being quantified is the "expected rate at which utility is transferred from the search provider to the user" (Moffat and Zobel 2008, p. 14). The issue then is to define a weighting scheme over the ranks at which relevant documents might be found;

DCG (Järvelin and Kekäläinen 2002) gives one method; rank-biased precision another (Moffat and Zobel 2008); and observation-based approaches, as considered here, a third.

In the same framework, Robertson (2008) recently described an interpretation of AP that fits this "expected utility" approach, building also on the *Expected Search Length* measure of Cooper (1968). Robertson observed that if a user is equally likely to abandon their search at any of the R relevant documents in the ranking, but to always continue their search after irrelevant documents are encountered, then AP is the expected value of precision, as observed over the universe of users. This expectation is directly comparable to the expectation generated by the various inner-product measures discussed in this paper; but employs a weighting vector that is dependent on the actual run being evaluated. Furthermore, this utility-based interpretation does not resolve the key issues that affect AP: that it is undefined when $R = 0$; and that it cannot be computed if R is unknown, or if the ranks of (any of) the relevant documents are unknown.

Robertson's normalized cumulative precision (NCP) can then be generalized by considering different probability distributions in regard to the stopping point, with one variant being a mechanism in which a truncated RBP-like geometric distribution is applied to the probability of terminating the search at relevant documents (Sakai and Robertson 2008). However, this approach again requires that users indefinitely scan rankings in which there are no answers, a significant shortcoming. Indeed, among methods where the ranking is allowed to influence the weighting vector, it seems more promising to consider arrangements "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" (Moffat and Zobel 2008, p. 17). Presuming that R is known, and truncating the weighting distribution after R relevant documents have been observed is also a choice that is debatable.

7 Conclusion

We have examined the relationship between traditional evaluation models and how users interact with results to web queries. Using the MSN query/clickthrough dataset, and manipulation of clickthrough data, we formed a user observation model describing the way in which web users examine documents in ranked answer listings, including making suitable allowance for boundaries between the pages of snippets presented to the user. Then, by comparing the observation model with the weighting models associated with evaluation metrics, we were able to establish how well those metrics correlated with the models of observed user behavior, and thus distinguish which metrics are indicative of user satisfaction.

One potential criticism of our methodology is that, while the sample used to generate the log-based observation model is non-trivial, there is no easy way of directly validating its accuracy without a comparable volume of user-based experimentation. For example, if many thousands of users could be monitored in eye tracking experiments, it would be possible to gauge the extent to which our derived observation model actually fits user behavior. But in the absence of a public opt-in approach in which users "volunteer their eyeballs" via on-computer cameras and ground-breaking tracking software, we are left to hypothesize, rather than actually demonstrate. Nevertheless, we believe that the hypothesis is a plausible one, and that it is not unreasonable to take the approach we have in this work, in order to construct a framework in which to compare the models associated with decaying-weight effectiveness metrics.

Of the tested evaluation measures, the geometric distribution employed in RBP gives a much better fit to the observation model than does any other approach. The best parameter was $p = 0.73$, a value that suggests that the average user can be expected to observe a total of 3.7 snippets in the results listing of a search engine. Suitably parameterized Poisson and Zipfian distributions also led to reasonable approximations of the observation model, but differed significantly in observational probabilities they predicted for documents later in the ranking. The log-harmonic distribution used in DCG was a poor fit to the observation model, and it greatly overestimates observation probabilities except near the head of the ranking.

Using TREC data and derived run statistics, we were also able to compute approximations for a range of recall-based measures such as AP, all of which utilize global relevance information and contextual relevance positioning. Of these measures, those derived from BPref and MRR were the best fits to the observation model, but were still poor compared to the static models. All of the recall-based measures tested share similar attributes, and tend to underestimate observation probabilities for documents at the head of the ranking, while overestimating the probabilities associated with the documents that appear later. Those evaluation measures should be used carefully when evaluating to significant result depths, and may not reflect the utility observed by typical web users.

Acknowledgments This work was supported by the Australian Research Council, Microsoft Research and NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program.

References

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 19–26). New York, NY: ACM. doi:10.1145/1148170.1148177.

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 25–32). New York, NY: ACM. doi:10.1145/1008992.100900.

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2004). Overview of the TREC 2004 terabyte track. In *Proceedings of the 2004 TREC text retrieval conference*. National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/TERA.OVERVIEW.pdf

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

Craswell, N., & Hawking, D. (2001). Overview of the TREC-2001 web track. In *Proceedings of the 2001 TREC text retrieval conference*. National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec10/papers/web2001.ps

Hawking, D. (2000). Overview of the TREC-9 web track. In *Proceedings of the 2000 TREC text retrieval conference*. National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec9/papers/web9.pdf

Jansen, B. J., & Spink, A. (2006). How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing and Management, 42*(1), 248–263. doi:10.1016/j.ipm.2004.10.00.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 133–142). New York, NY: ACM. doi:10.1145/775047.775067.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on*

research and development in information retrieval (pp. 154–161). New York, NY: ACM. doi:10.1145/1076034.1076063.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). New York: Oxford University Press.

Lee, U., Liu, Z., & Cho, J. (2005). Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on the World Wide Web* (pp. 391–400). New York, NY: ACM. doi:10.1145/1060745.1060804.

Liu, Y., Gao, B., Liu, T. Y., Zhang, Y., Ma, Z., He, S., et al. (2008). Browse rank: Letting web users vote for page importance. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 451–458). New York, NY: ACM. doi:10.1145/1390334.139041.

Moffat, A., Webber, W., & Zobel, J. (2007). Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 375–382). New York, NY: ACM. doi:10.1145/1277741.127780.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems, 27*(1–2), 1–27.

Park, L. A., & Zhang, Y. (2007). On the distribution of user persistence for rank-biased precision. In *Proceedings of the 12th Australasian document computing symposium* (pp. 17–24). Australia: School of Computer Science and Information Technology, RMIT University.

Robertson, S. (2008). A new interpretation of average precision. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 689–690). New York, NY: ACM. doi:10.1145/1390334.1390453.

Sakai, T. (2004). Ranking the NTCIR systems based on multigrade relevance. In *Proceedings of the Asian information retrieval symposium*. LNCS (Vol. 3411, pp. 251–262), Berlin, Heidelberg: Springer.

Sakai, T. (2007). Alternatives to BPref. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 71–78). New York, NY: ACM. doi:10.1145/1277741.1277756.

Sakai, T., & Robertson, S. (2008). Modelling a user population for designing information retrieval metrics. In *EVIA 2008 A satellite workshop of NTCIR-7: Proceedings of the second intenational workshop on evaluating information access* (pp. 30–41). Tokyo, Japan: National Institute of Informatics.

Teevan, J., Dumais, S. T., & Liebling, D. J. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 163–170). New York, NY: ACM. doi:10.1145/1390334.1390364.

Turpin, A., Scholer, F., Billerbeck, B., & Abel, L. A. (2006). Examining the pseudo-standard web search engine results page. In *Proceedings of the 11th Australasian document computing symposium* (pp. 9–16). Australia: Faculty of Information Technology, Queensland University of Technology.

Voorhees, E. M., & Harman, D. (2000). Overview of the ninth text retrieval conf. (TREC-9). In *Proceedings of the 2000 TREC text retrieval conference*. National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec9/papers/overview_9.pdf

Webber, W., Moffat, A., & 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* (pp. 51–58). New York, NY: ACM. doi:10.1145/1390334.1390346.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 307–314). New York, NY: ACM. doi:10.1145/290941.291014.