# Incorporating user behavior information in IR evaluation

Emine Yilmaz Milad Shokouhi Nick Craswell Stephen Robertson
Microsoft Research Cambridge
Cambridge, UK
{eminey, milads, nickcr, ser}@microsoft.com

## ABSTRACT

Many evaluation measures in Information Retrieval (IR) can be viewed as simple user models. Meanwhile, search logs provide us with information about how real users search. This paper describes our attempts to reconcile click log information with user-centric IR measures, bringing the measures into agreement with the logs. Studying the discount curve of NDCG and RBP leads us to extend them, incorporating the probability of click in their discount curves. We measure accuracy of user models by calculating 'session likelihood'. This leads us to propose a new IR evaluation measure, Expected Browsing Utility (EBU), based on a more sophisticated user model. EBU has better session likelihood than existing measures, therefore we argue it is a better user-centric IR measure.

## 1. INTRODUCTION

This paper is concerned with user-centric IR evaluation, where an evaluation measure should model the reaction of a real user to a list of results, evaluating the utility of the list of documents to the user. Web search experiments usually employ an IR measure that focuses only on top-ranked results, under the assumption that Web users deal 'shallowly' with the ranked list. This is probably correct, but we might ask: How can we be sure that Web search users are shallow, and how should we choose the degree of shallowness. In this paper, our solution is to make IR evaluation consistent with real user click behavior. We still evaluate based on relevance judgments on a list of search results, but the importance of each search result is brought in line with the probability of clicking that result.

In our experiments we use click logs of a search engine (`bing.com`) taken from January 2009, combined with relevance judgments for 2057 queries. For each judged query we extracted the top-10 results for up to 1000 real query instances, and the pattern of clicks in the form of 10 Booleans (so each result is either clicked or not clicked). More than 91% of all top-10 query-URL pairs were judged on the 5-level scale {Perfect, Excellent, Good, Fair, Bad}. Unjudged documents are assumed to be Bad. We divide the queries into two sets of equal size: training and test.

A key difference between user-centric IR evaluation measures, such as Normalized Discounted Cumulative Gain (NDCG) [2] and Rank Biased Precision (RBP) [3], is the choice of discount function. Many experiments with NDCG apply a discount at rank $r$ of $1/\log(r+1)$. Another metric, RBP, has a persistence parameter $p$ so that the probability of seeing position $r$ is $p^{r-1}$. Note, some evaluation measures such as Average Precision are not easily interpretable as a user model. Such measures are beyond the scope of this paper, since we focus on user-centric evaluation.

The next section considers the discount curves of NDCG and RBP, in contrast to real click behavior. Noting a discrepancy, we extend the two metrics based on information about the probability of click on each relevance label. Having done so, the discount curves are more in line with real user behavior. However, the curves do not incorporate information about the user's probability of returning to the results list, having clicked on a result. Therefore the next section introduces our new evaluation measure Expected Browsing Utility (EBU). Finally we introduce Session Likelihood, a test for whether an evaluation measure is in agreement with click logs. Under that test, EBU is most in line with real user behavior, therefore we argue it is a superior user-centric evaluation measure.

## 2. DISCOUNT FUNCTIONS AND CLICKS

One of the key factors for differentiating between the evaluation metrics is their *discount functions*. Most user-centric IR evaluation metrics in the literature can be written in the form of $\sum_{r=1}^{N} p(\text{user observes document at rank r}) \cdot gain(r)$ as the discount function is assumed to be modeling the probability that the user observes a document at a given rank. Therefore, the quality of a metric is directly dependent on how accurately the discount function estimates this probability. In the case of Web search, this probability value should ideally correspond to the probability that the user *clicks* on a document at rank $r$. Hence, one can compare the evaluation metrics based on how their discount function (their *assumed* probability of click) compare with the actual probability that the user clicks on a document. Discount functions that are more consistent with click patterns are more flexible in explaining – and evaluating – the users Web search behavior.

Next, we compare the user models associated with the underlying discount functions of RBP and NDCG. The top two plots in Figure 1 show the average probability of click (averaged over all sessions in the test data) per rank. We then compare this actual probability of click with the click probability *assumed* by different evaluation metrics. As mentioned above, this probability corresponds to the *discount* function used in the definition of the metrics. The upper


SIGIR'09, July 19-23, 2009, Boston, USA.

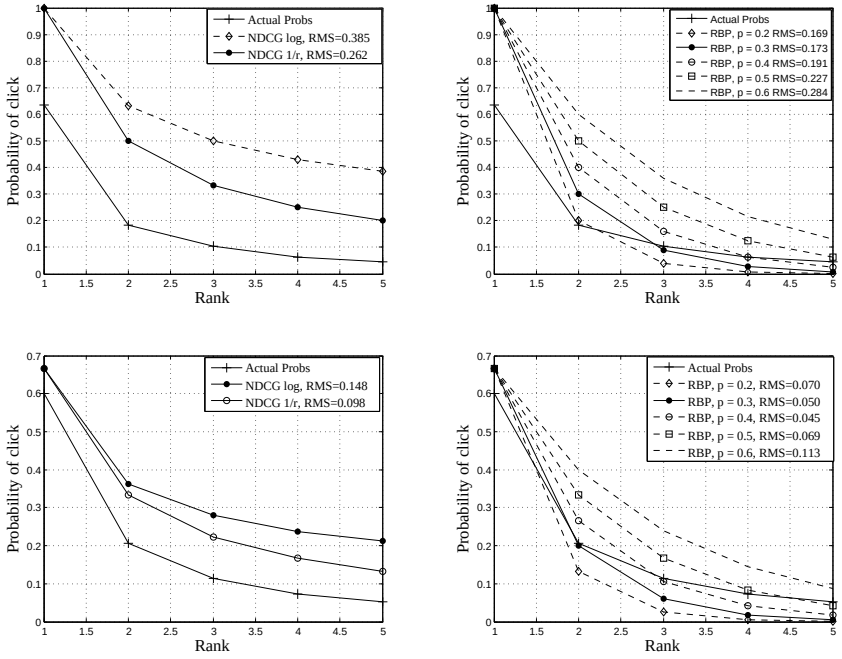


**Figure 1: P(click) vs. rank for different metrics.**

left and right plots compare the discount function of NDCG (with the commonly used $1/\log_( r+1)$ and $1/r$ discounts) and RBP (with $p \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$) with the actual click probability, respectively. For comparison purposes, the plots report the Root Mean Squared ($RMS$) error between the probability of click assumed by a metric and the actual probability of click. It can be seen that the probability of click assumed by these two metrics is quite different than the actual click probability.

As the discount functions in NDCG and RBP are not derived from search logs, it is not surprising to see that they are not successful in predicting clicks. In the following section, we show how extending such metrics by incorporating the quality of snippets can significantly improve the discount functions for predicting the probabilities of clicks.

## 3. MODELING THE IMPACT OF SNIPPETS

One reason for the discrepancy between the described discount functions and the click patterns is that these metrics do not account for the fact that the users only click on *some* documents depending on the relevance of the summary (snippets). Both RBP and NDCG assume that the user *always* clicks on the document at the first rank, whereas the actual probability of click calculated from our training search logs shows that the probability that the user clicks on the first ranked document is only slightly higher than 0.6.

To address this issue, we enhance the NDCG and RBP user models by incorporating the snippet quality factor and considering its impact on the probability of clicks. We hypothesize that the probability that the user clicks on a document (i.e., the quality of the summary) is a direct function of the relevance of the associated document. Table 1 supports our claim by showing $p(C|summary) \sim p(C|relevance)$ obtained using the training dataset.[1]It can be seen that the probability that the user clicks on a document tends to increase as the level of relevance of the document increases. Note that this behavior is slightly different for *Bad* and *Fair* documents, in which case there is a slight difference in the click probability. This is caused by the fact that (1) the documents judged as *Fair* tend to be slightly relevant to the user information need; hence, they are *effectively Bad* to the user, and (2) the unjudged documents are treated as *Bad* in our computations.

**Table 1: Probability of click given the relevance**

| Relevance | $P(click\|relevance)$ |
|---|---|
| Bad | 0.5101 |
| Fair | 0.5042 |
| Good | 0.5343 |
| Excellent | 0.6530 |
| Perfect | 0.8371 |

Motivated by these observations, we extend NDCG and RBP to incorporate the *summary* quality into their discount functions as follows: If the discount function of the metric dictates that the user visits a document at rank $r$ with probability $p(d_r)$, then the probability that the user *clicks* on the document at rank $r$ can be computed as $p(d_r) \cdot p(C|summary_r)$ (where the click probabilities are shown in Table 1). The bottom two plots in Figure 1 show how the extended versions of metrics then compare with the actual click probability. It can be seen that the extended versions

[1]For simplicity, we assume that the quality of summaries and the relevance of documents are strongly correlated. That is, relevant summaries for relevant documents and vice versa.

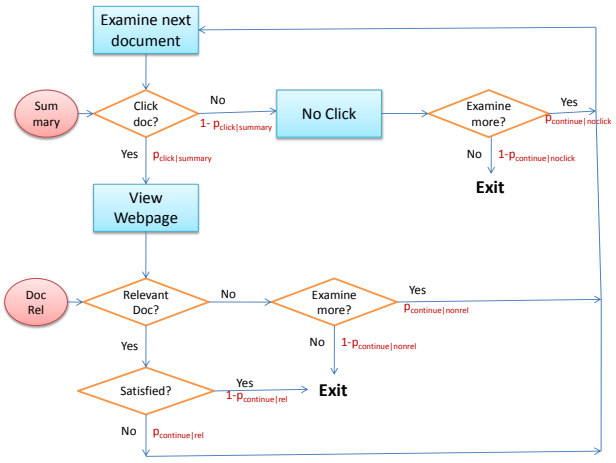


**Figure 2: The user browsing model associated with the new evaluation metric.**

**Table 2: Probability of continue given the relevance**

| Relevance | $P(cont\|relevance_r)$ |
|---|---|
| Bad | 0.5171 |
| Fair | 0.5727 |
| Good | 0.6018 |
| Excellent | 0.4082 |
| Perfect | 0.1903 |

of these metrics can approximate the actual probability of click substantially better than the standard versions.

We would like to note that Turpin et al. [4] recently also suggested that document summary information should be incorporated in evaluation retrieval evaluation, independent of our work. They showed that using the summary information in evaluation may alter the conclusions regarding the relative quality of search engines. However, their work mainly focus on average precision as the evaluation metric.

# 4. EXPECTED BROWSING UTILITY (EBU)

All the metrics described so far assume that the probability that the user will continue search at each rank is independent of (1) whether the user has clicked on a document or not, and (2) the relevance of the document seen by user. Intuitively, we expect the search behavior of users to change based on the relevance of the last visited document. That is, visiting a highly relevant document that perfectly satisfies the user's information need (e.g. a navigational answer) shall be strongly correlated with the probability of terminating the search session.

We confirmed our hypothesis by computing the probabilities of *continuing* the search session conditioned on the relevance of the last clicked document. The results generated from our training set are shown in Table 2. It can be seen that if the document is very relevant to the information need (e.g., *Perfect*), then the user is likely to stop browsing the results as he has found the information he was looking for. On the other hand, if the user clicks on a document that is not relevant to his information need (e.g., *Bad*), then he is again likely to stop browsing as he is frustrated with the result he has clicked on and thinks documents retrieved lower than that will probably be even less relevant.

Motivated by the probabilities of click and continue shown in Tables 1 and 2, we propose a novel user model in which: (1) When a user visits a document, the user may or may not click the document depending on the quality of the summary, and (2) The relevance of a document visited by a user directly affects whether the user continues the search or not.

Figure 2 shows the user model associated with our metric. The associated user model can be described as follows: The user starts examining the ranked list of documents from top to bottom. At each step, the user first just observes the *summary* (e.g., the snippet and the url) of the document. Based on the quality of the summary, with some probability $p(C|summary)$ the user clicks on the document. If the user does not click on the document, then with probability $p(cont|noclick)$ he/she continues examining the next document or terminates the search session with probability $1 - p(cont|noclick)$.

If the user clicks on the document, then he or she can assess the *relevance* of the document. If the document did not contain any relevant information, then the user continues examining with the probability $p(cont|nonrel)$ or stops with $1 - p(cont|nonrel)$ probability. If the clicked document was relevant, then the user continues examining with probability $p(cont|rel)$ (which depends on the relevance of the clicked document).

A similar user model has been suggested by Dupret et al. [1]. However, their work is mainly focused on predicting the future clicks, while our goal is to integrate the probabilities of clicks with evaluating the search results.

We use past click data together with relevance information to model the user search behavior. At each result position $r$, our model computes the expected probability of examining the document $p(E_r)$ as follows: We first assume that the user always examines the very first document, hence $p(E_1) = 1$. Now, suppose the user has examined the document at rank $r-1$ and we would like to compute $p(E_r)$. Given that the user has already examined the document at $r-1$, according to our model, with probability $p(C|summary_{r-1})$ the user clicks on the document at rank $r-1$, observes the relevance of the document at rank $r-1$ and continues browsing the ranked list with probability $p(cont|rel_{r-1})$. Alternatively, with probability $1 - p(C|summary_{r-1})$ the user does not click on the document at rank $r-1$ and continues browsing with probability $p(cont|noclick)$. Overall, the probability that the user will examine the document at rank $r$ can be written as:

$$\begin{aligned} p(E_r) &= p(E_{r-1}) \cdot [p(C|summary_{r-1}) \cdot p(cont|rel_{r-1}) \\ &+ (1 - p(C|summary_{r-1})) \cdot p(cont|noclick)] \end{aligned}$$

Given that the user has examined the document at rank $r$, the probability that the user clicks on this document is $p(C|summary_r)$. That is, the user clicks on a document at rank $r$ with probability $p(C_r) = p(E_r) \cdot p(C|summary_r)$.[1]

Therefore, in total, the **Expected Browsing Utility (EBU)** that the user receives from the output of the search engine is then $EBU = \sum_{r=1}^{N} p(C_r) \cdot rel_r$ (divided by the EBU value of an optimal list so that the metric is between 0 and 1), where $rel_r$ is the relevance of document at rank

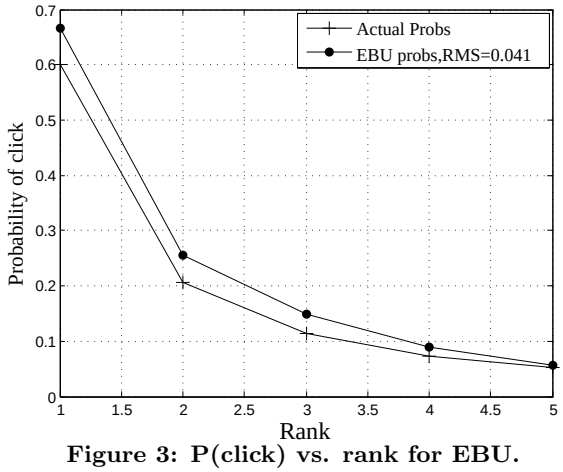


**Figure 3: P(click) vs. rank for EBU.**

$r$. In EBU, the importance of a document $d$ depends on (1) its relevance and (2) the probability that $d$ is clicked and viewed by the user.

Figure 3 shows the same curves using EBU as the metric (computed using probabilities from Table 1 and Table 2). Comparing the EBU curves with those in Figure 1, it can be seen that EBU is better than both versions of NDCG and RBP.

# 5. EVALUATING EVALUATION METRICS

In the above experiments we focused on the *average* click probability, i.e., the average probability that a user will click on a document at some rank $r$. Ideally, one would like to be able to infer the individual clicks per session. This way, the evaluation of user satisfaction per user session would be much accurate. Hence, in the second set of experiments, we compare the probability of click dictated by the discount function of a metric with the actual click observations per session.

For that, we use the click probability dictated by an evaluation metric as a generative model and then compute the probability that this distribution would generate the sessions that were observed in the test data (i.e., the *session likelihood*). Instead of computing the session likelihood, one can also compute the session log likelihood. Let $p(C_r|M)$ be the probability of click at rank $r$ dictated by the discount function of the metric $M$ and let the likelihood of a particular session $s$ given this metric be

$$P(s|M) = \prod_{\forall r, doc_r \in C_s} P(C_r|M) \cdot \prod_{\forall r, doc_r \in NC_s} (1 - P(C_r|M))$$

where $C_s$ and $NC_s$ correspond to the documents clicked and not clicked in session $s$, respectively and $doc_r$ refers to the document at rank $r$ in session $s$. The session log likelihood can then be written as:

$$\begin{aligned} log(P(sessions|M)) &= \log[\prod_{\forall s \in sessions} P(s|m)] \\ &= \sum_{\forall s \in sessions} \log(P(s|m)) \end{aligned}$$

The first column in Table 3 shows the session log likelihood for each metric. For comparison purposes, the second column in the table shows the average probability of observing the sessions in the test data. It can be seen that EBU can predict the behavior of an individual user (i.e., per session) much better than all the other metrics.

| | Session Log Likelihood | P(click per session) |
|---|---|---|
| RBP, p=0.2 | -2.3859 | 0.0920 |
| RBP, p=0.3 | -2.1510 | 0.1164 |
| RBP, p=0.4 | -2.0570 | 0.1278 |
| RBP, p=0.5 | -2.0732 | 0.1258 |
| RBP, p=0.6 | -2.2007 | 0.1107 |
| NDCG, log | -2.3064 | 0.0996 |
| NDCG, 1/r | -2.0435 | 0.1296 |
| EBU | -1.9371 | 0.1441 |

**Table 3: Likelihood of individual sessions given each evaluation metric.**

# 6. CONCLUSIONS

Most evaluation metrics in information retrieval aim at evaluating the satisfaction of the user given a ranked list of documents. Hence, these metrics are based on some underlying user models which are assumed to be modeling the way users search. However, most of these user models are based on unrealistic assumptions.

In this paper, we showed how click logs can be used to devise enhanced evaluation measures. We first extended two commonly used evaluation metrics, NDCG and RBP, to incorporate probability of click in their discount curves. We then introduced EBU, new evaluation metric that comes from a more sophisticated user model than the other metrics. Finally, using a novel evaluation methodology of evaluating evaluation measures (referred to as *session likelihood*), we compared these different metrics and showed that EBU is a better metric in terms of modeling user behavior.

# 7. REFERENCES

[1] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 331–338, Singapore, Singapore, 2008. ACM.

[2] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[3] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 375–382, Amsterdam, The Netherlands, 2007. ACM.

[4] A. Turpin, F. Scholer, K. Jarvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *SIGIR '09: Proceedings of the 32nd annual international ACM SIGIR conference on Research and development in informaion retrieval*, Boston, MA, USA, 2009. ACM. To Appear.