From a User Model for Query Sessions to Session Rank Biased Precision (sRBP)

Aldo Lipani
University College London
London, United Kingdom
aldo.lipani@ucl.ac.uk

Ben Carterette
Spotify
New York, NY, United States
carteret@acm.org

Emine Yilmaz
University College London
London, United Kingdom
emine.yilmaz@ucl.ac.uk

ABSTRACT

To satisfy their information needs, users usually carry out searches on retrieval systems by continuously trading off between the examination of search results retrieved by under-specified queries and the refinement of these queries through reformulation. In Information Retrieval (IR), a series of query reformulations is known as a query-session. Research in IR evaluation has traditionally been focused on the development of measures for the ad hoc task, for which a retrieval system aims to retrieve the best documents for a single query. Thus, most IR evaluation measures, with a few exceptions, are not suitable to evaluate retrieval scenarios that call for multiple refinements over a query-session. In this paper, by formally modeling a user's expected behaviour over query-sessions, we derive a session-based evaluation measure, which results in a generalization of the evaluation measure Rank Biased Precision (RBP). We demonstrate the quality of this new session-based evaluation measure, named Session RBP (sRBP), by evaluating its user model against the observed user behaviour over the query-sessions of the 2014 TREC Session track.

KEYWORDS

session search, retrieval evaluation, user model, sRBP

ACM Reference Format:
Aldo Lipani, Ben Carterette, and Emine Yilmaz. . From a User Model for Query Sessions to Session Rank Biased Precision (sRBP). In . ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

In order to improve search we need to evaluate it. Research in Information Retrieval (IR) evaluation plays a critical role in the improvement of search engines [13]. Traditionally, IR evaluation has been focused on the evaluation of search engines capable of serving the best results given a *single query*. However, this kind of evaluation does not reflect a more realistic search scenario where users are allowed to reformulate their queries.

Users usually start their search with under-specified queries that are later reformulated based on the new information acquired during the search [14]. Reformulations can be numerous and endure until the users have either satisfied their information need, or given up with the search due to frustration. Carterette et al. [1], for the Text REtrieval Conference (TREC) Session track, define such reformulations as a *query-session*.

However, the evaluation measures used in the TREC Session track are still traditional evaluation measures developed for a single-query search. In this track, search engines are given as input a recording of a user session, which consists of an initial query and subsequent reformulations with their search results and clicks. The task of the engines is to provide a search result to the last reformulation of each session for which no search result and clicks are provided. Thus evaluating over the full session is rank-equivalent to evaluating only the last reformulation.

In this paper we introduce a novel session evaluation metric. We start by developing a user model for query sessions. From this user model we then develop a query session-based evaluation measure. This evaluation measure results in a generalization of RBP [11]. Hence, we name it session RBP (sRBP). sRBP extends RBP by introducing a new parameter, named balancer (b). This parameter quantifies the users proclivity to persist in their search by reformulating the query rather than examining more documents from the current search result. Their persistency, like for RBP, is controlled by the other parameter p. With our experiments we aim to answer the following research questions:

RQ1. How does the user model at the base of session-based evaluation measures predicts the expected user behaviour on sessions? How do they compare to single-query measures?
RQ2. Are the single-query measures, DCG and RBP, similar to the session-based measures, sDCG and sRBP, in evaluating search engines?
RQ3. In the same context of RQ2, is sDCG similar to sRBP?

We evaluate these research questions using the 2014 TREC Session track query-session dataset. The contributions of this paper are: (i) a novel user model for query-sessions, (ii) a theoretically grounded derivation of the evaluation measure sRBP (and with obvious simplifications also of RBP), (iii) the derived evaluation measure, (iv) a thorough comparison of the single-query measures (DCG and RBP) and session-based measures (sDCG and sRBP).

The remainder of this paper is structured as follows. In Section 2 we present related work. In Section 3 we introduce the notation used throughout the paper. The user model is presented in Section 4. In Section 5 we derive the evaluation measure sRBP. In Section 6 we present the experiments and results. We conclude in Section 7.

, ,
.

2 RELATED WORK

The main limitation of single-query search evaluation measures is that they force us, when considering sessions, to either (i) evaluate only the last reformulation, or (ii) aggregate evaluations of the query and all reformulations together. However, both approaches fail in capturing the session length, which like the length of search results, affects user satisfaction; both approaches consider a session with n reformulations equal to a session with $n+1$ reformulations. This makes this evaluation biased towards longer sessions because engines by gaining more information provide better results [12].

In the TREC Session track, search engines were given as input session information plus an additional reformulation whose search result needed to be provided by the engines. In this case, these two approaches make sense only when comparing sessions to themselves. When comparing search engines, we can analytically show that these approaches produce rank equivalent results (their scores differ only by a constant). In fact, only the first approach was used in the TREC Session track; the second approach would have provided the same ranking of search engines.

Driven by the lack of evaluation measures able to evaluate query-session searches, Järvelin et al. [8] developed the evaluation measure session Discounted Cumulative Gain (sDCG), which is a generalization of the DCG evaluation measure [7]. sDCG evaluates a query session by weighting each subsequent reformulation with a progressively smaller weight:

$$\text{sDCG}(r,q) = \sum_{m=1}^{M}\sum_{n=1}^{N} \frac{1}{(1+\log_{bq}(m))\log_b(n+1)} j(r_{m,n},q),$$

where M is the length of the query-sessions, N is the length of search results, r is the list of results, $j(r_{m,n},q)$ returns the relevance value for the topic q associated to the document $r_{m,n}$ ranked at position n for the query m, and bq and b are free parameters of the evaluation measure. These two parameters can model various user behaviours: a small value for bq (or b) models impatient users, who most likely will not reformulate their queries (or not examine in depth the search result), while a larger value models patient users, who most likely will make several reformulations (or examine in depth the search result). However, as we will show in this paper the sDCG discount function does not characterize well the behaviour observed on Session track users.

Beside developing session-based evaluation measures, another branch of research focuses into developing simulation-based approaches. Kanoulas et al. [9] simulate all the possible browsing paths to transform query-sessions into classical search results in order to enable the use of standard evaluation measures. With a focus on novelty, Yang and Lad [15] simulate all the possible reformulations to compute an expected utility of a query-session. Contrary to these works, in this paper we take an analytical approach, which does not rely on simulations.

To develop our own user model, we took inspiration from the probabilistic framework developed for user models for predicting user clicks in single-query search [4–6, 10]. These models are rooted in the *cascade model* proposed by Craswell et al. [3], which says that users examine documents from top to bottom until the first click, after which they never go back. In this paper we extend the cascade model to query-sessions. However, in contrast to the *click chain model* [5] and the *complex searcher model* [10] we do not consider clicks but examinations. Where an examination can be interpreted classically as the examination of a retrieved document or, as in the case of the TREC Session track test collection, as the examination of its snippet. This allowed us to simplify the derivation of the evaluation measure. We leave to future work the extension of the user model to consider also clicks.

3 NOTATION

The following is a set of symbols, function, and random variables used in this paper:

Symbols	
M	The length of the query-session.
N	The length of search results.
m	The current reformulation.
n	The current document in the search result.
$\mathcal{Q}$	A collection of query-sessions.
$\mathcal{R}$	A set of runs.
q	A query-session.
r	A run $r \in \mathcal{R}$.
r_n	The document retrieved at rank n.
$r_{m,n}$	The document at reformulation m and rank n.
Function	
$f(r,q)$	An evaluation measure for r and q.
$d(r_{m,n})$	Returns the discount value for $r_{m,n}$ defined for f.
$j(r_{m,n},q)$	Returns the degree of relevance of d for q.
Random Variables	
Q	System is Queried = $\{q,\overline{q}\}$.
E	Title and Snippet are Examined = $\{e,\overline{e}\}$.
L	User Leaves Topic or Search = $\{\ell,\ddot{\ell},\overline{\ell}\}$.
J	Document Relevance = $\{j,\bar{j}\}$.

4 USER MODEL

Users start searching by querying the search system. The system responds with a search result for this initial query. Users then examine the first document in the search result, by which we mean either the examination of the title and snippet of the document or its full content via clicking. After this examination, users face three options: (1) continue examining the next document in the search result; (2) continue re-querying the system with a reformulated query; or (3) leave the search, which can occur for two reasons: users have satisfied their information need or they have given up with the search due to frustration. A graphical representation of this user model is shown in the flow-chart in Figure 1.

We formalize this user model by associating to every user decision point a discrete random variable:

- $Q = \{q,\overline{q}\}$ for *querying* or not querying the system;
- $E = \{e,\overline{e}\}$ for *examining* or not examining a ranked document;
- $L = \{\ell,\ddot{\ell},\overline{\ell}\}$ for *leaving* search, continuing with a reformulation, or continuing to browse the ranked results, respectively;
- $J = \{j,\bar{j}\}$, which is the observed *relevance* of an examined document.

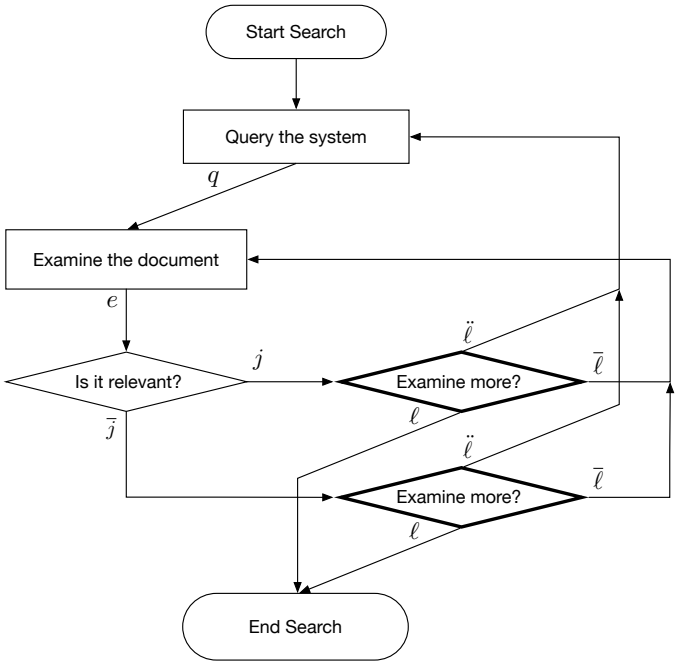

Figure 1: Flow-chart of the proposed user model.

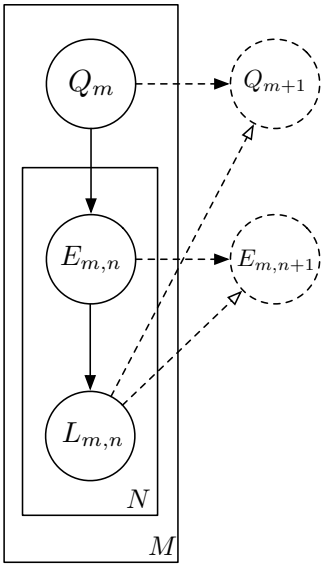

Figure 2: Graphical model of the proposed user model.

Each one of these random variables is indexed with one or two indices: m identifies the query or the result produced by this query in the query-sessions, and n identifies the rank of the document at which the document has been retrieved by the query m. M and N are the lengths of the query-session and search result ranking respectively.

The graphical model in Figure 1 defines (a) the dependence structure among these random variables, and (b) how these variables interact at each examination step. The former formally translates into the following:

$$p(Q_m, E_{m,n}, L_{m,n}) = p(Q_m)p(E_{m,n}|Q_{m,n})p(L_{m,n}|E_{m,n}), \quad (1)$$

where we define that the probability of leaving depends on the examination of the document, and the probability of examining a document in the search ranking depends on the outcome of the querying of the system. The latter is presented in the following paragraphs.

We define the probability of querying the system $p(Q_m)$ in function of the random variables associated to the previous query $(m-1)$, in the following recursive way:

$$p(Q_1 = q) = 1$$
$$p(Q_m = q) = \sum_{n=1}^{N} p(Q_{m-1} = q, L_{m-1,n} = \ddot{\ell}). \quad (2)$$

The first equation says that the probability of issuing the first query is certain. The second equation says that the probability of querying the system at step m is equal to the sum over the documents of the search result for the query $m-1$ of the joint probability between having queried the system and having left the search result in order to query the system with a reformulated query.

We define the probability of examining a document $P(E_{m,n})$ in function of the random variables associated to the previous document $(n-1)$, in the following recursive way:

$$p(E_{m,1} = e|Q_m = q) = 1$$
$$p(E_{m,n} = e|Q_m = q) = p(E_{m,n-1} = e, L_{m,n-1} = \overline{\ell}|Q_m = q). \quad (3)$$

The first equation says that the probability of examining the first document is certain. The second equation says that the probability of examining a document at rank n is equal to the joint probability between having examined a document at rank $n-1$ and then continued search by not leaving the search result.

The next two propositions will be useful in the next section for developing an evaluation measure. The first proposition is:

$$P(E_{m,n} = e|Q_m = \overline{q}) = 0. \quad (4)$$

This is the probability of examining a document if the user has not issued any query, which should be 0. The second proposition is:

$$p(L_{m,n} = \ell|E_{m,n} = \overline{e}) = 1. \quad (5)$$

This is the probability of leaving search if the user has not examined the current document, which should be 1.

5 EVALUATION MEASURES

Most utility-based measures compute the effectiveness of a search engine as the product between a discount function d and a relevance function j as follows [2]:

$$f(r, q) = \sum_{n=1}^{N} d(r_n) \cdot j(r_n, q), \quad (6)$$

where r_n is the document retrieved at rank n on a search result r returned by the search engine given as input a query q and a collection of documents. However, f is not suitable to evaluate

query-sessions because it is limited to a single-query, while sessions may happen over a number of queries.

To overcome this limitation, we generalize Eq. (6) as follows:

$$f(r,q) = \sum_{m=1}^{M}\sum_{n=1}^{N} d(r_{m,n}) \cdot j(r_{m,n}, q_m), \tag{7}$$

where $r_{m,n}$ is the document retrieved at rank n on the search result r returned by the search engine given as input a query q_m and a collection of documents. Where q_m is one of the reformulation of the query-session q. This generalization expands the set of actions that can be taken into account by the discount function d. In Eq. (6) d can only model actions a user can take within a single search result, now in Eq. (7) d extends this set of actions by including actions a user can take over a query-session.

5.1 User Model Driven Derivation of a Discount Function

Based on the user model developed in Section 4 we define the discount function for a document equal to its probability of being examined. Formally, this can be written as:

$$d(r_{m,n}) = p(E_{m,n} = e), \tag{8}$$

where $r_{m,n}$ is the document retrieved by the search engine for the query q_m at the the rank n.

Based on Eq. 1, Eq. 8 can be expanded as:

$$\begin{aligned}
p(E_{m,n} = e) &= \sum_{Q_m}\sum_{L_{m,n}} p(Q_m, E_{m,n} = e, L_{m,n}) \\
&= \sum_{Q_m}\sum_{L_{m,n}} p(Q_m)p(E_{m,n} = e|Q_m)p(L_{m,n}|E_{m,n} = e) \\
&= \sum_{Q_m} p(Q_m)p(E_{m,n} = e|Q_m) \underbrace{\sum_{L_{m,n}} p(L_{m,n}|E_{m,n} = e)}_{1} \\
&= p(Q_m = q)p(E_{m,n} = e|Q_m = q),
\end{aligned}$$

where the last equality is simplified by using Eq. 4.

Now, we need to derive $P(Q_m = q)$ and $P(E_{m,n} = e|Q_m = q)$. We start from the latter. In order to estimate $P(E_{m,n} = e|Q_m = q)$, according to Eq. 3 we need to quantify $p(E_{m,n} = e, L_{m,n} = \bar{\ell}|Q_m = q)$. Based on Eq. 1 we obtain:

$$\begin{aligned}
&p(E_{m,n} = e, L_{m,n} = \bar{\ell}|Q_m = q) = \\
&\qquad = p(E_{m,n} = c|Q_m = q) \underbrace{p(L_{m,n} = \bar{\ell}|E_{m,n} = e)}_{\alpha_{m,n}},
\end{aligned}$$

where, for the sake of clarity, we substitute the probability of not leaving search given the document for the query m at rank n has been examined with $\alpha_{m,n}$. Substituting this equation to Eq. 3 we can simplify it as following:

$$p(E_{m,n} = e|Q_m = q) = \prod_{n'=1}^{n-1} \alpha_{m,n'}. \tag{9}$$

This is possible thanks to a property of the product operator that returns 1 when its upper bound is lower than its lower bound.

In order to estimate $P(Q_m = q)$, according to Eq. 2 we need to quantify $\sum_{n=1}^{N} p(Q_m = q, L_{m,n} = \ddot{\ell})$. Based on Eq. 1 we obtain:

$$\begin{aligned}
&\sum_{n=1}^{N} p(Q_m = q, L_{m,n} = \ddot{\ell}) \\
&\quad = \sum_{n=1}^{N}\sum_{E_{m,n}} p(Q_{m,n} = q, E_{m,n}, L_{m,n} = \ddot{\ell}) \\
&\quad = p(Q_m = 1)\sum_{n=1}^{N}\sum_{E_{m,n}} p(E_{m,n}|Q_m = q)p(L_{m,n} = \ddot{\ell}|E_{m,n}) \\
&\quad = p(Q_m = 1)\sum_{n=1}^{N} p(E_{m,n} = e|Q_m = q) \underbrace{p(L_{m,n} = \ddot{\ell}|E_{m,n} = e)}_{\beta_{m,n}},
\end{aligned}$$

where the last equality is simplified by using Eq. 5. For the sake of clarity, we substitute the probability of continuing search by reformulating a query given that the document for the query m at rank n has been examined with $\beta_{m,n}$. Substituting this equation to Eq. 2 we can simplify it as following:

$$p(Q_m = q) = \prod_{m'=1}^{m-1}\sum_{n=1}^{N}\prod_{n'=1}^{n-1} \alpha_{m',n'}\beta_{m',n}. \tag{10}$$

Combining Eq. 9 and Eq. 10 we now can compute the discount function as:

$$d(r_{m,n}) = \prod_{m'=1}^{m-1}\sum_{n'=1}^{N}\prod_{n''=1}^{n'-1} \alpha_{m',n''}\beta_{m',n'}\prod_{n'=1}^{n-1}\alpha_{m,n'}. \tag{11}$$

Two probabilities need to be estimated in order to calculate this discount function. These probabilities are the ones substituted by $\alpha_{m,n}$ and $\beta_{m,n}$.

5.2 Session Rank-Biased Precision

In order to estimate the two probabilities $\alpha_{m,n}$ and $\beta_{m,n}$ we make the simple assumption that these probabilities do not change over m and n, they are constant. If these probabilities do not change ($\alpha_{m,n} = \alpha$ and $\beta_{m,n} = \beta$), we can simplify Eq. 11 as:

$$d(r_{m,n}) = \prod_{m'=1}^{m-1}\sum_{n'=1}^{N} \alpha^{n'-1}\beta\alpha^{n-1}.$$

Moreover, we can observe that by taking the limit of N to infinity, the discount function simplifies as following:

$$d(r_{m,n}) = \left(\frac{\beta}{1-\alpha}\right)^{m-1}\alpha^{n-1}.$$

We can then understand α and β as the probabilities of, after having examined a document, continuing searching by not leaving the current ranking and leaving the current ranking, respectively. The sum of these probabilities plus the probability of leaving search ($p(L_{m,n} = l|E_{m,n} = e)$), which we substitute with γ has, by definition, to sum to one:

$$\alpha + \beta + \gamma = 1.$$

This means that the range of parameter values these probabilities can take is restricted by this last equation. To avoid this problem, and give a more human-friendly interpretation of these parameters we apply the following substitutions:

$$\alpha = bp, \quad \beta = (1-b)p, \quad \gamma = 1-p,$$

where we name $b \in [0,1]$ as the *balance* parameter, which balances between reformulating queries and examining more documents

in the search result, and; we name $p \in [0, 1]$ as the *persistence* parameter because it is similar to the persistence parameter of RBP [11], which defines the persistence of users in continuing search.

Applying these substitutions to the discount function we obtain:

$$d(r_{m,n}) = \left(\frac{p - bp}{1 - bp}\right)^{m-1} (bp)^{n-1}. \quad (12)$$

It turns out that the sum of the discount values so defined over a query-session of infinite length is equal to $1 - p$. This value can be used as a normalization factor, similarly to how RBP is normalized.

Based on these observations, substituting the discount function to Eq. 7 we define the new evaluation measure sRBP as follows:

$$\text{sRBP}(r, q) = (1 - p) \sum_{m=1}^{M} \left(\frac{p - bp}{1 - bp}\right)^{m-1} \sum_{n=1}^{N} (bp)^{n-1} \cdot j(r_{n,m}, q).$$

When $b = 1$, sRBP simplifies to RBP for the first query, and ignores the rest of reformulations in the query-session[1]. When $b = 0$, sRBP will score only the first document of every reformulation.

We can now show that sRBP is a generalization of RBP. If we set $b = 1$ and consider a query-session with only a single query, sRBP is equal to:

$$\text{sRBP}_{b=1}(r, q) \overset{\text{if } |q|=M=1}{=} (1 - p) \sum_{n=1}^{N} p^{n-1} \cdot j(r_n, q) = \text{RBP}(r, q).$$

This equality not only demonstrates that this user model is consistent with the user model at the base of RBP but also provides an additional intuition on RBP since this derivation is grounded on probability theory.

6 EXPERIMENTS

This experimental section aims to answer the three research questions presented in the introduction. The software used in this paper is available on the website of the first author.

6.1 Material

We used the 2014 Session TREC track test collection [1]. This dataset contains 1257 query-sessions, including queries and reformulations, ranked results from a search engine, and clicks; Out of these 1257 query-sessions, only 101 of them have been judged for relevance, using a pool of retrieved documents from 73 participating teams. The judgment process produced 16,949 judged documents. These 101 judged query-sessions are developed over 52 unique topics.

6.2 Experimental Setup

To evaluate the quality of sRBP with respect to sDCG in predicting the expected user behaviour (RQ1), we compare their user models with the user behaviour observed over the 1257 query-sessions. In order to observe the examination of a document we assume that if a document at rank n has been clicked, then all the documents with rank lower or equal than n have been examined. Using these query-sessions we compute the probability of a user to examine a document at rank n for a query m. We do this for every m and n. In Table 1 we observe these probabilities.

The observed user behaviour is compared against the user model of the two session-based evaluation measures, sDCG and sRBP. To compute the probability of a user to examine a document at rank n we use the discount functions of the two evaluation measures. Following the discount function for sDCG:

$$d_{\text{sDCG}}(r_{m,n}) = \frac{1}{(1 + \log_{bq}(m)) \log_b(n + 1)}.$$

The discount function for sRBP is in Eq. 12. To compute the probability of a user to examined a document at rank n for a query m we compute the discount function, which is then normalized by its sum in order to generate a probability distribution over the queries (m) and documents (n).

To compare the user models and the observed user behaviour we use the 3 following measures of error: Total Squared Error (TSE), Total Absolute Error (TAE), and Kullback-Leibler Divergence (KLD).

All our evaluation measures have parameters. To find the best parameter values we perform a grid search on the TSE measure with grid dimension of 0.01. For sDCG, we search the parameter values in their recommended ranges [8], $1 < bq \leq 1000$ and $1 < b \leq 20$. For sRBP, we search the parameter values in their ranges, $0 \leq p \leq 1$ and $0 \leq b \leq 1$.

In order to compare the quality of the single-query measures, RBP and DCG, against the two session-based measures we perform a similar experiment as above. However, in this case we consider every query and reformulation in the query-sessions as individual queries, assuming each is an independent event. The learning strategy, range of the parameters, and measure of error we used during learning are the same as above. This experiment, in addition to showing how the single-query measures compare against the observed user-behaviour, will also show how the learned parameters change when we evaluate engines without query-session information.

To analyze if single-query measures provide a different perspective with respect to session-based measures (RQ2), we perform a correlation analysis between the measures DCG and sDCG, and RBP and sRBP. We use the Kendall's tau correlation coefficient. This analysis is performed using the parameters learned in the previous experiment.

We do this first over the 101 judged query-sessions, and then over the combination the 73 search results and the 101 judged query-sessions. In the former case, when comparing query-sessions, for the single-query measures we two approaches, as also mentioned in the introduction: (i) we evaluate only the last reformulation, and (ii) we evaluate all queries and reformulations. When presenting results, we refer to these two approaches as "RBP (i)" and "RBP (ii)" or "DCG (i)" and "DCG (ii)".

Finally, we compare sRBP against sDCG (RQ3) by performing the same correlation analysis as done for RQ2. This analysis will inform us about how similar is the information provided by sRBP with respect to sDCG.

6.3 Results

6.3.1 RQ1: Model accuracy. We first address the question of whether our user model provides an accurate prediction of actual user behavior. Table 1 shows observed user behaviour measured on query-sessions, while Tables 2 and 3 show the predictions made by the sRBP and sDCG user models. Comparing the model predictions (in Table 2 and Table 3) by eye, we can clearly see that sRBP better

[1] This is the case since $0^0 = 1$ while $0^n = 0$

Table 1: Observed user behaviour on the Session Track 2014 query-sessions. Every cell contains the probability of a user examining a document retrieved at the n-th rank (rows) in the search result produced by the m-th reformulation (columns).

n\m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.1598	0.0968	0.0698	0.0462	0.0286	0.0149	0.0086	0.0045	0.0020	0.0015	0.0008	0.0006	0.0005	0.0003	0.0001
2	0.0429	0.0290	0.0154	0.0096	0.0055	0.0026	0.0011	0.0006	0.0003	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001
3	0.0326	0.0218	0.0123	0.0080	0.0045	0.0019	0.0009	0.0006	0.0003	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001
4	0.0246	0.0179	0.0097	0.0063	0.0038	0.0015	0.0008	0.0005	0.0003	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
5	0.0194	0.0147	0.0086	0.0050	0.0033	0.0014	0.0006	0.0004	0.0003	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
6	0.0162	0.0120	0.0070	0.0042	0.0027	0.0011	0.0006	0.0003	0.0003	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
7	0.0135	0.0102	0.0061	0.0037	0.0027	0.0011	0.0006	0.0003	0.0003	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
8	0.0111	0.0086	0.0052	0.0033	0.0027	0.0010	0.0006	0.0003	0.0003	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
9	0.0088	0.0073	0.0046	0.0033	0.0026	0.0009	0.0006	0.0001	0.0003	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
10	0.0063	0.0057	0.0042	0.0031	0.0026	0.0009	0.0005	0.0001	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
61	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2: Normalized sRBP discount values (b = 0.64 and p = 0.86). The content of this table is meaningwise equal to Table 1.

n\m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.1504	0.1019	0.0690	0.0467	0.0316	0.0214	0.0145	0.0098	0.0066	0.0045	0.0030	0.0021	0.0014	0.0009	0.0006
2	0.0806	0.0545	0.0369	0.0250	0.0169	0.0115	0.0078	0.0053	0.0036	0.0024	0.0016	0.0011	0.0007	0.0005	0.0003
3	0.0431	0.0292	0.0198	0.0134	0.0091	0.0061	0.0042	0.0028	0.0019	0.0013	0.0009	0.0006	0.0004	0.0003	0.0002
4	0.0231	0.0156	0.0106	0.0072	0.0049	0.0033	0.0022	0.0015	0.0010	0.0007	0.0005	0.0003	0.0002	0.0001	0.0001
5	0.0124	0.0084	0.0057	0.0038	0.0026	0.0018	0.0012	0.0008	0.0005	0.0004	0.0003	0.0002	0.0001	0.0001	0.0001
6	0.0066	0.0045	0.0030	0.0021	0.0014	0.0009	0.0006	0.0004	0.0003	0.0002	0.0001	0.0001	0.0001	0.0000	0.0000
7	0.0035	0.0024	0.0016	0.0011	0.0007	0.0005	0.0003	0.0002	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
8	0.0019	0.0013	0.0009	0.0006	0.0004	0.0003	0.0002	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0010	0.0007	0.0005	0.0003	0.0002	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.0005	0.0004	0.0002	0.0002	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
61	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3: Normalized sDCG discount values (bq = 1.07 and b = 4.54). The content of this table is meaningwise equal to Table 1.

n\m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.0489	0.0032	0.0021	0.0017	0.0014	0.0013	0.0012	0.0011	0.0011	0.0010	0.0010	0.0009	0.0009	0.0009	0.0009
2	0.0309	0.0020	0.0013	0.0010	0.0009	0.0008	0.0008	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005
3	0.0245	0.0016	0.0010	0.0008	0.0007	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004	0.0004
4	0.0211	0.0014	0.0009	0.0007	0.0006	0.0006	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
5	0.0189	0.0012	0.0008	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003
6	0.0174	0.0011	0.0007	0.0006	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003	0.0003
7	0.0163	0.0011	0.0007	0.0006	0.0005	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
8	0.0154	0.0010	0.0007	0.0005	0.0005	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
9	0.0147	0.0010	0.0006	0.0005	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
10	0.0141	0.0009	0.0006	0.0005	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
61	0.0082	0.0005	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

characterizes the observed user behaviour. The measures of error comparing these user models are provided in Table 4, which shows that a visual inspection is correct: sRBP gives an order of magnitude better prediction as compared to sDCG.

Table 5 shows the errors measured on the transformed query-sessions – query and reformulation of each session are treated independently. This indicates that the user models behind sRBP and sDCG are no worse than their single-query variants.

Figure 3 shows the sensitivity of the sRBP parameters. In the first plot we depict the full TSE landscape. The two black lines identify the points of minimum for the two parameters b and p. In the second and third plots we show how the error changes when varying b (p) fixing p (b) to its optimal. The sRBP parameters sensitivity analysis shows that the TSE landscape has a convex shape for sRBP. This is not true for sDCG; its TSE landscape is concave. The convexity of the error landscape guarantees the stability of the

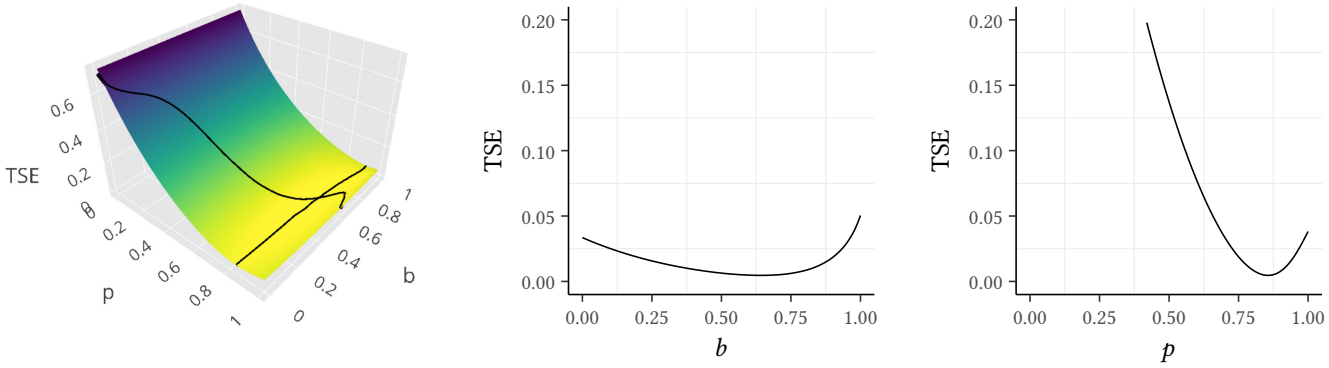

Figure 3: Sensitivity plots of sRBP parameters. The two lines in the first plot on the left delimit the points where the gradient is maximum. The two plots on the right are 2d-projections of the first plot on the left made by fixing a dimension to its best parameter value.

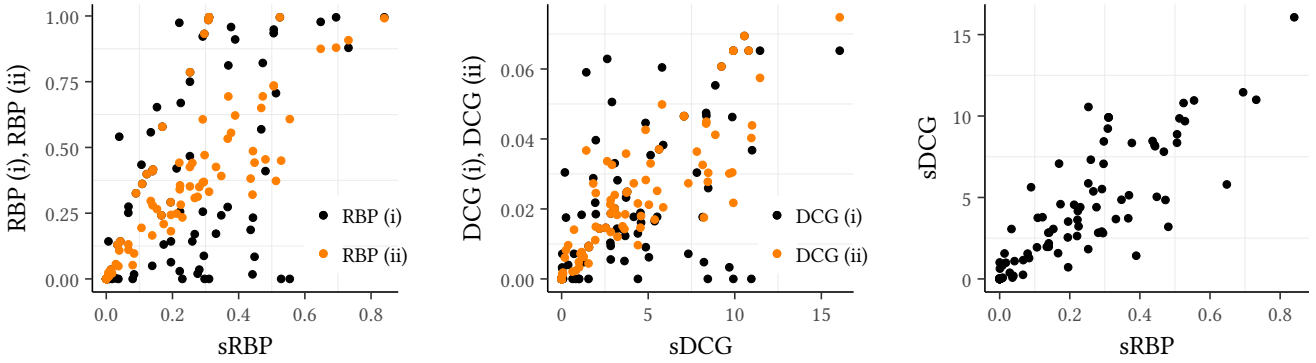

Figure 4: Scatter plots of evaluation measures over the 101 judged query-sessions. Every point is a query-session.

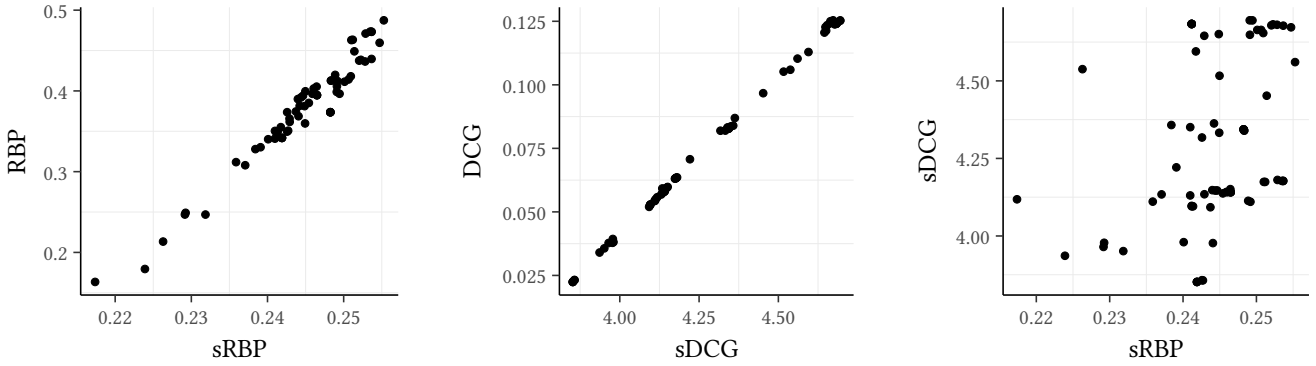

Figure 5: Scatter plots of evaluation measures over the 73 search results combined with the 101 judged query-sessions. Every point is a search result.

learned parameters. In particular, we notice that the best parameter value for p is in the optimal range as the one suggested by Moffat and Zobel [11] for a standard web-user (≈ 0.8). We conclude that session-based models better predict the expected user behaviour. In particular, among the evaluated models, sRBP performs the best.

6.3.2 RQ2: Single-query versus session measures. Next, we address the use of single-query measures versus full-session measures for the task of evaluating systems. In Figures 4 and 5, the first two plots on the left show the results obtained when comparing the evaluation single-query measures against the session-based mea-

Table 4: User models against the observed user behaviour on the Session Track 2014 query-sessions.

	Parameters	TSE	TAE	KLD
sRBP	$b = 0.64, p = 0.86$	**0.0046**	**0.4950**	**0.9475**
sDCG	$bq = 1.07, b = 4.54$	0.0362	1.3357	2.2710

Table 5: User models against observed user behaviour on the Session Track 2014 making every query and reformulations of query-sessions independent.

	Parameters	TSE	TAE	KLD
RBP	$p = 0.59$	**0.0252**	0.4242	**0.6624**
DCG	$b = 1.29$	0.1521	1.2162	1.5035
sRBP	$b = 0.92, p = 0.64$	**0.0252**	**0.4238**	0.6679
sDCG	$bq = 1.01, b = 1.26$	0.1521	1.2162	1.5035

Table 6: Kendall's tau correlations over query-sessions.

	RBP (ii)	DCG (i)	DCG (ii)	sRBP	sDCG
RBP (i)	0.675	0.906	0.666	0.555	0.532
RBP (ii)	-	0.659	0.869	0.801	0.747
DCG (i)	-	-	0.702	0.532	0.560
DCG (ii)	-	-	-	0.746	0.780
sRBP	-	-	-	-	0.772

Table 7: Kendall's tau correlations over search results.

	DCG	sRBP	sDCG
RBP	0.293	0.843	0.270
DCG	-	0.315	0.950
sRBP	-	-	0.290

-sures on query-sessions and search results. The last plot on the right compare the two query-session based evaluation measures. In Tables 6 and 7 we find the Kendall's tau correlation coefficients over all possible pairs of evaluation measures on query-sessions and search results.

The evaluation of query-sessions varies a great deal between single-query measures and sessions-based measures (in Figure 4). The correlation between RBP (i) and sRBP, and DCG (i) and sDCG is low (0.555 and 0.560). However, when considering the second evaluation approach RBP (ii) and DCG (ii), the correlation increases (0.801 and 0.780). This suggests that the first part of the session provides a different information than only the last reformulation (as in (i)) and that the considering the full session (as in (ii)) correlated better with session-based measures.

When evaluating search results we observe that the correlation between RBP and sRBP, and between DCG and sDCG is higher (0.843 and 0.950). However, although DCG and sDCG correlate well, this is not true for RBP and sRBP. We conclude that query-session evaluation measures provide a different perspective when evaluating sessions or search results. In particular, we observe that difference between how sRBP would rank search results with respect to RBP is wider than the one we would get using sDCG with respect to DCG.

6.3.3 RQ3: sRBP versus sDCG. Finally, we compare the two session-based measures to each other. The last plots on the right in Figures 4 and 5 compare the two query-session based evaluation measures on query-sessions and search results.

sRBP and sDCG are very different in ranking sessions (Figure 4). Their correlation coefficients is 0.772, which is similar to the correlation between RBP (i) and DCG (i) and between RBP (ii) and DCG (ii) (0.659 and 0.869). This difference is particularly exacerbated when evaluating search results (Figure 5). Their correlation coefficient is in this case much lower (0.290). However, it is again consistent to the correlation between RBP and DCG (0.293). We conclude that sRBP and sDCG provide two different evaluation perspectives and that they are consistent with how RBP differ from DCG in single-query evaluation.

7 CONCLUSION

In this paper we have developed a user model for query-sessions under a well-defined probabilistic framework. This user model can be easily extended by making more realistic assumptions about the probabilities of leaving search, continuing examining the documents of the search result and continuing reformulating a query. Under simplifying assumptions – these probabilities are constant over time and independent by the relevance of the examined documents – we have on one hand, derived a new session-based evaluation measure, sRBP, on the other hand, demonstrated that this user model well approximates the expected behaviour as measured on 2014 TREC Session track query-sessions and justified its existence by showing that this evaluation measure provides a different perspective with respect to sDCG.

ACKNOWLEDGMENTS

This project was funded by the EPSRC Fellowship titled "Task Based Information Retrieval", grant reference number EP/P024289/1.

REFERENCES

[1] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. *Overview of the TREC 2014 session track.* Technical Report.
[2] Praveen Chandar and Ben Carterette. 2012. Using Preference Judgments for Novel Document Retrieval. In *Proc. of SIGIR.*
[3] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-bias Models. In *Proc. of WSDM.*
[4] Georges Dupret and Ciya Liao. 2010. A Model to Estimate Intrinsic Document Relevance from the Clickthrough Logs of a Web Search Engine. In *Proc. of WSDM.*
[5] Fan Guo, Chao Liu, Anitha Kannan, Tom Minka, Michael Taylor, Yi-Min Wang, and Christos Faloutsos. 2009. Click Chain Model in Web Search. In *Proc. of WWW.*
[6] Fan Guo, Chao Liu, and Yi Min Wang. 2009. Efficient Multiple-click Models in Web Search. In *Proc. of WSDM.*
[7] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002).
[8] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In *Proc. of ECIR.*
[9] Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. 2011. Evaluating Multi-query Sessions. In *Proc. of SIGIR.*
[10] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proc. of CIKM.*
[11] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (2008).
[12] Anne Schuth, Floor Sietsma, Shimon Whiteson, and Maarten De Rijke. 2014. Optimizing base rankers using clicks. In *Proc. of ECIR.*
[13] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval.* Vol. 1. MIT press Cambridge.
[14] Dietmar Wolfram, Amanda Spink, Bernard J. Jansen, and Tefko Saracevic. 2001. Vox populi: The public searching of the web. *JASIST* 52, 12 (2001).
[15] Yiming Yang and Abhimanyu Lad. 2009. Modeling Expected Utility of Multi-session Information Distillation. In *Proc. of ICTIR.*