Understandability Biased Evaluation for Information Retrieval

Guido Zuccon
g.zuccon@qut.edu.au

Queensland University of Technology (QUT), Australia

Abstract. Although relevance is known to be a multidimensional concept, information retrieval measures mainly consider one dimension of relevance: topicality. In this paper we propose a method to integrate multiple dimensions of relevance in the evaluation of information retrieval systems. This is done within the gain-discount evaluation framework, which underlies measures like rank-biased precision (RBP), cumulative gain, and expected reciprocal rank. Albeit the proposal is general and applicable to any dimension of relevance, we study specific instantiations of the approach in the context of evaluating retrieval systems with respect to both the topicality and the understandability of retrieved documents. This leads to the formulation of understandability biased evaluation measures based on RBP. We study these measures using both simulated experiments and real human assessments. The findings show that considering both understandability and topicality in the evaluation of retrieval systems leads to claims about system effectiveness that differ from those obtained when considering topicality alone.

1 Introduction

Traditional information retrieval (IR) evaluation relies on the assessment of topical relevance: a document is topically relevant to a query if it is assessed to be on the topic expressed by the query. The Cranfield paradigm and its subsequent incarnations into many of the TREC, CLEF, NTCIR or FIRE evaluation campaigns have used this notion of relevance, as reflected by the collected relevance assessments and the retrieval systems evaluation measures, e.g., precision and average precision, recall, bpref, RBP, and graded measures such as discounted cumulative gain (DCG) and expected reciprocal rank (ERR).

Relevance is a complex concept and the nature of relevance has been widely studied [16]. A shared agreement has emerged that relevance is a multidimensional concept, with topicality being only one of the factors (or criteria) influencing the relevance of a document to a query [8,28]. Among others, core factors that influence relevance beyond topicality are: scope, novelty, reliability and understandability [28]. However, these factors are often not reflected in the evaluation framework used to measure the effectiveness of retrieval systems.

In this paper, we aim to develop a general evaluation framework for information retrieval that extends the existing one by considering the multidimensional

nature of relevance. This is achieved by considering the gain-discount framework synthesised by Carterette [4]; this framework encompasses the widely-used DCG, RBP and ERR measures. Specifically, we focus on a particular dimension of relevance, understandability, and devise a family of measures that evaluate IR systems by taking into account both topicality and understandability. While the developed framework is general and could be used to model other factors of relevance, there are a number of compelling motivations for focusing on an extension to understandability only:

- even if a document is topically relevant, it is of no use to a user if it cannot be understood at all;
- understandability is a key factor when assessing relevance in many domain-specific scenarios, e.g., consumer health search [1,12,13,26,27];
- resources exist that allow us to assess the impact of evaluating multidimensional relevance when considering understandability, both through simulations and explicit human assessments of understandability.

Specifically, we aim to answer the following research questions: (RQ1) How can relevance dimensions (and specifically understandability) be integrated within IR evaluation? (RQ2) What is the impact of understandability biased measures on the evaluation of IR systems?

2 Related Work

Research on document relevance has shown that users' relevance assessments are affected by a number of factors beyond topicality, although topicality has been found to be the essential relevance criteria. Chamber and Eisenberg have synthesised four families of approaches for modelling relevance, highlighting its multidimensional nature [24]. Cosijn and Ingwersen investigated manifestations of relevance such as algorithmic, topical, cognitive, situational and socio-cognitive, and identified relation, intention, context, inference and interaction as the key attributes of relevance [8]. Note that relevance manifestations and attributes in that work are different from what we refer to as factors of relevance in this paper. Similarly, the dimensions described by Saracevic [23], which are related to those of Cosijn and Ingwersen mentioned above, differ in nature from the factors or dimensions of relevance we consider in this paper.

The actual factors that influence relevance vary across studies. Rees and Schulz [20] and Cuadra and Katter [9] identified 40 and 38 factors respectively. Xu and Chen proposed and validated a five-factor model of relevance which consists of novelty, reliability, understandability, scope, along with topicality [28]. Zhang et al. have further validated such model [33]. Their empirical findings highlight the importance of understandability, reliability and novelty along with topicality in the relevance judgements they collected. Barry also explored factors of relevance beyond topicality [2]; of relevance to this work is that these user experiments highlighted that criteria pertaining to user's experience and background, including the ability to understand the retrieved information, influence relevance assessments. Mizzaro offered a comprehensive account of previous work attempting to define and research relevance [16].

While dimensions of relevance are often ignored in the evaluation of IR systems, notable exceptions do exists. The evaluation of systems that promote the novelty and diversity of the retrieved information, for example, required the development of measures that account for both the topicality and novelty dimensions. This need has been satisfied by fragmenting the information need into subtopics, or nuggets, and evaluating the systems against relevance assessments performed explicitly for each of the subtopics of the query. This approach has lead to the formulation of measures such as subtopic recall and precision [31], α-nDCG [6], and D#-measures [22], among others [5]. Nevertheless, the formulation of novelty and diversity measures differ from that of the measures proposed in this paper because we combine the gains achieved from different dimensions of relevance, rather than summing gains contributed by the different subtopics.

In this paper, the integration of understandability within IR evaluation proposed is cast within the gain-discount framework [4]. This framework generalises the common structure of many evaluation measures, which often involve a sum over the product of a gain function, mapping relevance assessments to gain values, and a discount function, that serves to modulate the gain by a discount based on the rank position at which the gain is achieved.

Previous work on quantifying the importance of understandability when evaluating IR systems has also used the gain-discount framework and simulations akin to those of Section 5, although at a smaller scale [34]. That work motivated us to further develop multidimensional based evaluation of IR systems and specifically to further investigate evaluation measures that account for both relevance and understandability assessments.

3 Gain-Discount Framework

In the gain-discount framework [4] the effectiveness of a system, conveyed by a ranked list of documents, is measured by the evaluation measure M, defined as:

$$M = \frac{1}{\mathcal{N}} \sum_{k=1}^{K} d(k) g(d@k) \tag{1}$$

where $g(d@k)$ and $d(k)$ are respectively the gain function computed for the (relevance of the) document at rank k and the discount function computed for the rank k, K is the depth of assessment at which the measure is evaluated, and $1/\mathcal{N}$ is a (optional) normalisation factor, which serves to bound the value of the sum into the range [0,1] (see also [25]).

Without loss of generality, we can express the gain provided by a document at rank k as a function of its probability of relevance; for simplicity we shall write $g(d@k) = f(P(R|d@k))$, where $P(R|d@k)$ is the probability of relevance given the document at k. A similar form has been used for the definition of the gain function for time-biased evaluation measures [25]. Measures like RBP, nDCG and ERR can still be modelled in this context, where their differences with respect to $g(d@k)$ reflect on different $f(.)$ functions being applied to the estimations of $P(R|d@k)$.

Different measures within the gain-discount framework use different functions for computing gains and discounts. Often in RBP the gain function is binary-

valued[1] (i.e., $g(d@k) = 1$ if the document at k is relevant, $g(d@k) = 0$ otherwise); while for nDCG $g(d@k) = 2^{P(R|d@k)} - 1$ and for ERR[2] $g(d@k) = (2^{P(R|d@k)} - 1)/2^{max(P(R|d))}$. The discount function in RBP is modelled by $d(k) = \rho^{k-1}$, where $\rho \in [0, 1]$ reflects user behaviour[3]; while in nDCG the discount function is given by $d(k) = 1/(\log_2(1 + k))$ and in ERR by $d(k) = 1/k$.

When only the topical dimension of relevance is modelled, as is in most retrieval systems evaluations, then $P(R|d@k) = P(T|d@k)$, i.e., the probability that the document at k is topically relevant (to a query). This probability is 1 for relevant and 0 for non-relevant documents, when considering binary relevance; it can be seen as the values of the corresponding relevance levels when applied to graded relevance.

4 Integrating Understandability

To integrate different dimensions of relevance in evaluation measures, we model the probability of relevance $P(R|d@k)$ as the joint distribution over all considered dimensions $P(D_1, \cdots, D_n|d@k)$, where each D_i represents a dimension of relevance, e.g., topicality, understandability, reliability, etc.

To compute the joint probability we assume that dimensions are compositional events and their probabilities independent, i.e., $P(D_1, \cdots, D_n|d@k) = \prod_{i=1}^{n} P(D_i|d@k)$. These are strong assumptions and are not always true. Eisenberg and Barry [10] highlighted that user judgements of document relevance are affected by order relationships, and proposals to model these dynamics have recently emerged, for example see Bruza et al. [3]. Nevertheless, Zhang et al. used crowdsourcing to prime a psychometric framework for multidimensional relevance modelling, where the relevance dimensions are assumed compositional and independent [33]. While the above assumptions are unrealistic and somewhat limitative, note that similar assumptions are common in information retrieval. For example, the Probability Ranking Principle assumes that relevance assessments are independent [21].

Following the assumptions above, the gain function with respect to different dimensions of relevance can be expressed in the gain-discount framework as:

$$g(d@k) = f(P(R|d@k)) = f\big(P(D_1, \cdots, D_n|d@k)\big) = f\Big(\prod_{i=1}^{n} P(D_i|d@k)\Big)$$

Evaluation measures developed within this framework would differ by means of the instantiations of $f\big(P(D_1, \cdots, D_n|d@k)\big)$, other than by which dimensions are modelled.

4.1 Understandability Biased Evaluation

In the remaining of this paper we investigate measures that limit the modelling of multidimensional relevance to only topicality, characterised by $P(T|d@k)$, and

[1] Although there is no requirement for this to be the case and RBP can be used for graded relevance [17].

[2] Where $P(R|d@k)$ captures either binary ($P(R|d@k)$ either 0 or 1) or graded relevance and $max(P(R|d))$ is the highest relevance grade, e.g., 1 in case of binary relevance.

[3] High values representing persistent users, low values representing impatient users.

understandability, characterised by $P(U|d@k)$. In the following, $P(R|d@k)$ is thus modelled by the joint $P(T, U|d@k)$ that is in turn computed as the product $P(T|d@k)P(U|d@k)$ following the assumptions discussed above. This transforms the gain function into:

$$g(d@k) = f(P(R|d@k)) = f\big(P(T|d@k)P(U|d@k)\big) \tag{2}$$

For simplicity, we further assume that $f(.)$ satisfies the distributive property, such that $f\big(P(T|d@k)P(U|d@k)\big) = f\big(P(T|d@k)\big) \cdot f\big(P(U|d@k)\big)$; this is often the case for estimations of gain functions used in IR. For example, if the gain function used in RBP is applied as $f(.)$ to both topicality and understandability, then the equality above would be satisfied.

Next, we consider specific instantiations of a well-known IR measure, RBP, to the case of multidimensional relevance, and specifically when considering both topicality and understandability. Because topicality is a factor that is traditionally used to instantiate measures, we name the newly proposed measures as understandability biased, to highlight the fact that they model understandability, in addition to topicality, for evaluating the effectiveness of the systems. Nevertheless, the same approach can be applied to other dimensions of relevance.

Rank-biased precision (RBP) [17] is a well understood measure of retrieval effectiveness which fits within the gain-discount framework. In RBP, gain is measured by a function $r(d@k)$ which is 1 if $d@k$ is relevant and 0 otherwise; discount is measured by a geometric function of the rank, i.e., $d(k) = \rho^{k-1}$, and $1 - \rho$ acts as a normalisation component. Formally, RBP is defined as:

$$RBP = (1 - \rho) \sum_{k=1}^{K} \rho^{k-1} r(d@k) \tag{3}$$

Within the view presented in Section 3, $r(d@k)$ is an initialisation of $f(P(T|d@k))$, where $f(.)$ is the identity function and $r(d@k)$ estimates $P(T|d@k)$. To integrate understandability, we assume that $f(P(R|d@k)) = f(P(T|d@k) \cdot P(U|d@k))$ in line with Section 3, thus obtaining the understandability-biased RBP:

$$uRBP = (1-\rho) \sum_{k=1}^{K} \rho^{k-1} P(T|d@k) \cdot P(U|d@k) = (1-\rho) \sum_{k=1}^{K} \rho^{k-1} r(d@k) \cdot u(d@k) \tag{4}$$

where $r(d@k)$ is the function that transforms relevance values into the corresponding gains and $u(d@k)$ is the function that transforms understandability values into the corresponding gains.

This general expression for uRBP can be further instantiated by making explicit how the respective gain functions are computed. For example, $r(d@k)$ could be computed in the same way the corresponding function is computed in RBP. Similarly, the function responsible for translating understandability estimations into gains, i.e., $u(d@k)$, can be instantiated such that it returns a value of 1 if the document is assessed understandable ($P(U|d@k) = 1$) and a value of 0 if it is assessed as not understandable ($P(U|d@k) = 0$).

Alternatives may include collecting graded assessments of the understandability of documents, and associating different gains to different levels of understanding, akin to the use of graded relevance in measures like nDCG. This approach provides a graded variant of understandability-biased RBP, which we indicate as uRBPgr. Specifically, in Section 6 for uRBPgr we instantiate $u(d@k)$ as the function that provides a gain of 1 if $d@k$ is very easy to understand, 0.8 if it is somewhat easy to understand, 0.4 if it somewhat hard to understand and a gain of 0 (no gain) if it is very difficult to understand. Thus, if a document is very difficult (easy) to understand, its contribution to the value of uRBPgr would be 0 (1), regardless of the relevance of the document itself – this is in line with uRBP. However, when documents are somewhat easy or difficult to understand, the corresponding gains are used to modulate (or scale) the gains derived from the relevance assessments, in practice reducing these gains because of a partial lack in understandability.

5 Simulating Understandability Biased Evaluation

In the previous sections we have introduced a general framework for including understandability along with topicality in the evaluation of IR systems, and we have proposed instantiations of the framework based on the rank-biased precision measure (answering RQ1). Next, we aim to answer our second research question (RQ2): what is the impact of accounting for understandability in the evaluation of IR systems. To answer this, we instruct two empirical analyses.

The first analysis (Section 5) relies on simulations, where the understandability of documents is estimated using computational models of readability, which thus serves as a proxy to assess understandability. User understandability requirements are estimated using two (simple) user models. For this analysis we only consider the binary uRBP measure for brevity.

The second analysis (Section 6) relies on human provided assessments of understandability of documents, and considers both binary and graded uRBP.

Both analyses use the CLEF eHealth collection assembled to evaluate consumer health search tasks [12,13,18]. The collection consists of more than one million health related webpages. For the simulations we use the query topics distributed in 2013 and 2014 (for which there is no explicit understandability assessment) in addition to those distributed in 2015. Instead, for the experiments of Section 6 we use the query topics distributed in 2015 only as these come with explicit understandability assessments. Queries in this collection relate to the task of finding information about specific health conditions, treatments or diagnosis. We have chosen to study the impact of understandability biased evaluation using this collection because real-world tasks within consumer health search often require that the retrieved information can be understood by cohorts of users with different experience and understanding of health information [1,12,26,27,35]. Indeed, health literacy (the knowledge and understanding of health information) has been shown as a critical factor influencing the value of information consumers acquire through search engines [11].

Along with the queries, we also obtain the runs that were originally submitted to the relevant tasks at CLEF 2013–2015 [12,13,18][4]. Both the simulations and the analysis with real user assessments focus on the changes in system rankings obtained when evaluating using standard RBP and its understandability variants (uRBP and uRBPgr). System rankings are compared using Kendall rank correlation (τ) and AP correlation (τ_{AP}) [30], which assigns higher weights to changes that affect top performing systems.

In all our experiments the RBP parameter ρ which models user behaviour (RBP persistence parameter) was set to 0.8 for all variants of this measure, following the findings of Zhang et al. [32].

5.1 Readability as Estimation of Understandability

To computationally simulate the impact of understandability on the evaluation of search engines, we use readability as a proxy for understandability and we integrate this in the evaluation process, along with standard relevance assessments. Readability, although not providing a comprehensive account of understandability, is one of the aspects that influence the understanding of text [28].

To estimate readability (and thus understandability), we employ established general readability measures as those used in previous work that studied the readability of health information (including that returned by search engines [1,26,27]), e.g., SMOG, FOG and Flesch-Kincaid reading indexes. These measures consider the surface level of language in documents, i.e., the wording and syntax of sentences. Thus, long sentences, words containing many syllables and unpopular words, are each indicators of difficult language to read [15]. In this paper, we use the FOG measure to estimate the readability of a text; FOG is computed as

$$FOG(d) = 0.4 * (avgslen(d) + phw(d)) \tag{5}$$

where $avgslen(d)$ is the average length of sentences in a document d and $phw(d)$ is the percentage of hard words (i.e., words with more than two syllables) in d.

While often used in studies to evaluate the readability of health information, doubts have been casted on the suitability of these measures, especially to the specific health context. For example, Yan et al. [29] claimed that the highest readability difficulties are experienced at word level rather than at sentence level. Alternative approaches that measure language readability beyond the surface characteristics of language have been proposed, e.g., language models [7] and supervised support vector machine classifiers [14]. Nevertheless, these measures appear to be adequate for the purpose of the analysis reported here (study the impact of understandability on evaluation).

5.2 Modelling $P(U|d@k)$

Equation 5 provides document readability scores; we then transform readability scores into the probability of a document being understandable (i.e., $P(U|d@k)$) by means of user models. In this case, user models encode different ways in which

[4] Obtained from the CLEF eHealth repository, https://github.com/CLEFeHealth.

users (and their capacity to understand retrieved information) are affected by different document readability levels.

Specifically, we consider two user models. In the first user model (characterised by the probability estimations $P_1(U|d@k)$), a user has a readability threshold th and every document that has a readability score below th is considered certainly understandable, i.e., $P_1(U|d@k) = 1$; while documents with readability above th are considered not understandable, i.e. $P_1(U|d@k) = 0$. This Heaviside step function is centred in th and its use to model $P(U|d@k)$ is akin to the gain function in RBP (also a step function). Thus, uRBP for this first user model can be rewritten as:

$$uRBP_1 = (1-\rho)\sum_{k=1}^{K}\rho^{k-1}r(k)u_1(k) \tag{6}$$

where, for simplicity of notation, $u_1(k)$ indicates the value of $P_1(U|d@k)$ and $r(k)$ is the (topical) relevance assessment of document k (alternatively, the value of $P(T|d@k)$); thus $g(k) = f(P(T|d@k)P_1(U|d@k)) = P(T|d@k)P_1(U|d@k) = r(k)u_1(k)$.

In the second user model, the probability estimation $P_2(U|d@k)$ is similar to the previous step function, but it is smoothed in the surroundings of the threshed value. This provides a more realistic transition between understandable and non-understandable information. This behaviour is achieved by the following estimation:

$$P_2(U|d@k) \propto \frac{1}{2} - \frac{\arctan\Big(FOG(d@k) - th\Big)}{\pi} \tag{7}$$

where arctan is the arctangent trigonometric function and $FOG(d@k)$ is the FOG readability score of the document at rank k. (Other readability scores could be used instead of FOG.) Equation 7 is not a probability distribution per se, but one such distribution can be obtained by normalising Equation 7 by its integral between $[\min\big(FOG(d@k)\big), \max\big(FOG(d@k)\big)]$. However Equation 7 is rank equivalent to such distribution, not changing the effect on uRBP. These settings lead to the formulation of a second simulated variant of uRBP, $uRBP_2$, which is based on this second user model and is obtained by substituting $u_2(k) = P_2(U|d@k)$ to $u_1(k)$ in Equation 6.

5.3 Analysis of the Simulations

In the simulations we consider three thresholds to characterise the user models with respect to the FOG readability values: $th = 10, 15, 20$. In general, documents with a FOG score below 10 should be near-universally understandable, while documents with FOG scores above 15 and 20 increasingly restrict the audience able to understand the text. We performed simple cleansing of the HTML pages, although a more conscious pre-processing may be more appropriate [19].

In the following we report the results observed using the CLEF eHealth 2013 topics and assessments. Figure 1 reports RBP vs. uRBP for the 2013 systems. Table 1 reports the values of Kendall rank correlation (τ) and AP correlation (τ_{AP}) between system rankings obtained with RBP and uRBP.

Higher values of th produce higher correlations between systems rankings obtained with RBP and uRBP; this is regardless of the user model used in uRBP (Table 1). This is expected as the higher the threshold, the more documents will have $P(U|d@k) = 1$ (or ≈ 1 for $uRBP_2$): in this case uRBP degenerates to RBP. Overall, $uRBP_2$ is correlated with RBP more than $uRBP_1$ is to RBP. This is because of the smoothing effect provided by the arctan function. This function in fact increases the number of documents for which $P(U|d@k)$ is not zero, despite their readability score being above th. This in turn narrows the scope for ranking differences between systems effectiveness. These observations are confirmed in Figure 1, where only few changes in the rank of systems are shown for $th = 20$ ($\times$ in Figure 1), with more changes found for $th = 10$ ($\circ$) and $th = 15$ ($+$).

The simulations reported in Figure 1 demonstrate the impact of understandability in the evaluation of systems for the considered task. The system ranked highest according to RBP (`MEDINFO.1.3.noadd`) is second to a number of systems according to uRBP if user understandability of up to FOG level 15 is wanted. Similarly, the highest $uRBP_1$ for $th = 10$ is achieved by `UTHealth_CCB.1.3. noadd`, which is ranked 28th according to RBP, and for $th = 15$ by `teamAEHRC.6.3`, which is ranked 19th according to RBP and achieves the highest $uRBP_2$ for $th = 10, 15$.

We repeated the same simulations for the 2014 and 2015 tasks. While we omit to report all results here due to space constraints, we do report in Table 2 the results of the simulations for the first user model tested on the 2015 task, so that these values can be directly compared to those obtained using the real assessments (Section 6). The trends observed in these results are similar to those re-

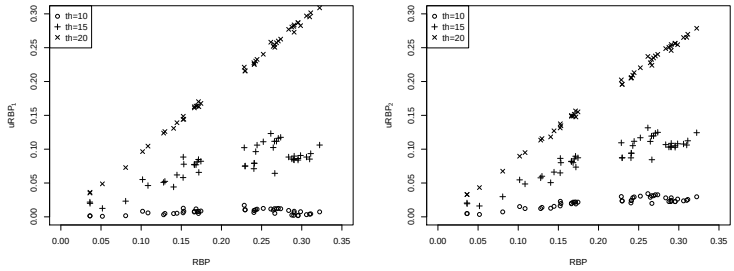

Fig. 1. RBP vs. **uRBP** for CLEF eHealth 2013 systems (left: $\mathbf{uRBP_1}$; right: $\mathbf{uRBP_2}$) at varying values of readability threshold ($\mathbf{th = 10, 15, 20}$).

	$th = 10$	$th = 15$	$th = 20$
RBP vs.	$\tau = .1277$	$\tau = .5603$	$\tau = .9574$
$uRBP_1$	$\tau_{AP} = -.0255$	$\tau_{AP} = .2746$	$\tau_{AP} = .9261$
RBP vs.	$\tau = .5887$	$\tau = .6791$	$\tau = .9574$
$uRBP_2$	$\tau_{AP} = .2877$	$\tau_{AP} = .4102$	$\tau_{AP} = .9407$

Table 1. Correlation (τ and $\tau_{\mathbf{AP}}$) between system rankings obtained with **RBP** and $\mathbf{uRBP_1}$ or $\mathbf{uRBP_2}$ for different values of readability threshold on CLEF eHealth 2013.

	$th = 10$	$th = 15$	$th = 20$
RBP vs.	$\tau = .5931$	$\tau = .8898$	$\tau = .9986$
$uRBP_1$	$\tau_{AP} = .5744$	$\tau_{AP} = .8777$	$\tau_{AP} = .9990$

Table 2. Correlation (τ and $\tau_{\mathbf{AP}}$) between system rankings obtained with **RBP** and $\mathbf{uRBP_1}$ for different values of the readability threshold on CLEF eHealth 2015.

ported for the 2013 data (and for the 2014 data), i.e., the higher the threshold th, and the larger the correlation between RBP and uRBP becomes. However larger absolute correlations values between RBP and $uRBP_1$ are found when using the 2015 data, if compared to the correlations reported in Table 1 for the 2013 task. The full set of results, including high resolutions plots, is made available at `http://github.com/ielab/ecir2016-UnderstandabilityBiasedEvaluation`.

6 Evaluation with Real Judgements

Next, we study the impact of understandability in the evaluation of IR systems by considering judgements about document understandability and topicality provided by human assessors. To this aim, we consider topics and systems from CLEF eHealth 2015 Task 2 [18], in which both topicality and understandability assessments (binary and graded) were collected. We can thus compute uRBP according to its two instantiations in Section 4.1 and compare their system rankings with those of RBP.

Figure 2 compares the evaluations of each CLEF system obtained with RBP and the two uRBP variants (binary, graded). In addition, the correlations between the measures are: RBP-uRBP, $\tau = 0.8666$, $\tau_{AP} = 0.8168$; RBP-uRBPgr, $\tau = 0.9077$, $\tau_{AP} = 0.8866$.

These results highlight that when human assessments of understandability are used, uRBP is generally strongly correlated with RBP. This is even more so for the graded uRBP variant because uRBPgr assigns a zero value of $P(U|d@k)$ to less documents than its binary counterpart, as documents that were assessed as somewhat hard to understand produce a small but not null gain (0.4) in uRBPgr, while they produce a zero gain in uRBP. When compared to the results of the simulations, the correlation trends between RBP and uRBP when real assessments are used is more in line with the findings obtained when the simulation used $th = 15$ than when other threshold values were used (Table 2).

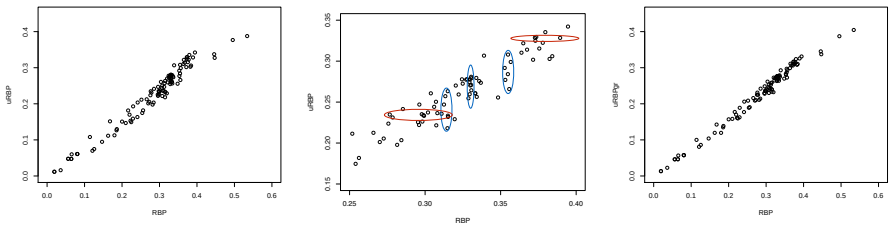

Fig. 2. RBP vs. **uRBP** for CLEF eHealth 2015 systems, with understandability judgements sourced from human assessors (binary uRBP left, uRBPgr (graded) right). Centre: a detail of the correlation between RBP vs. binary uRBP.

Nevertheless, despite being highly correlated, system rankings obtained with RBP and uRBP do differ. In particular, in our experiments differences seem concentrated when RBP ranges between 0.25 and 0.40, and uRBP (or uRBPgr) ranges between 0.15 and 0.35: this is depicted in the central plot of Figure 2 for the binary uRBP. Indeed, the analysis reveals that there is large variability in terms of uRBP for a number of systems, which instead appeared indistinguishable when evaluated using RBP: examples of such cases are highlighted in blue in the plot. These cases refer to situations in which there were a number of systems that returned a similar rank distribution of relevant documents (thus obtaining approximately the same RBP). However, these different systems retrieved different relevant documents and some of those documents are of no or limited understandability, and thus are discounted by uRBP. Similarly, in red we have highlighted examples where systems are evaluated as being equivalent in terms of uRBP, but are different in terms of RBP. This happens when the additional gains obtained by the systems that are superior in terms of RBP are due to documents that, despite being relevant, have been assessed as being of low or no understandability.

7 Conclusions

In this paper, we have proposed a method to integrate understandability in the gain-discount framework for evaluating IR systems. The approach is general and can be adapted to other dimensions of relevance. This is left for future work.

Using the proposed framework, we have devised understandability biased instantiations of rank-biased precision and studied their impact on the evaluation of IR systems. Other measures that are developed within the gain-discount framework can be extended following the proposed approach to consider relevance dimensions other than topicality, e.g., ERR and nDCG.

In our experiments, understandability assessments (or other estimations of the probability $P(U|d)$) were transformed into gains in a manner akin to how binary or graded relevance assessments are transformed into gains when computing gain-discount measures. Indeed, here topicality and understandability were given the same importance when determining the effectiveness of IR systems. However, different dimensions of relevance affect the perception of document relevance in different proportions. For example Xu and Chen [28] first, and Zhang et al. [33] later, have found that topicality is more influential than understandability. The weighting of different dimensions of relevance could be accomplished through a different $f(.)$ function for converting $P(T, U|d@k)$ into gain values. The exploration of this possibility and its implications for evaluation is left for future work.

Acknowledgements. The author is thankful to Bevan Koopman, Leif Azzopardi, Joao Palotti, Peter Bruza, Alistair Moffat and Lorraine Goeuriot for their comments on the ideas proposed in this paper.

References

1. O. H. Ahmed, S. J. Sullivan, A. G. Schneiders, and P. R. McCrory. Concussion information online: evaluation of information quality, content and readability of concussion-related websites. *British journal of sports medicine*, 46(9):675–683, 2012.

2. C. L. Barry. User-defined relevance criteria: an exploratory study. *JASIS*, 45(3):149–159, 1994.
3. P. D. Bruza, G. Zuccon, and L. Sitbon. Modelling the information seeking user by the decision they make. In *Proc. of MUBE*, pages 5–6, 2013.
4. B. Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proc. of SIGIR*, pages 903–912, 2011.
5. C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proc. of WSDM*, pages 75–84, 2011.
6. C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. of SIGIR*, pages 659–666, 2008.
7. K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. *JASIST*, 56(13):1448–1462, 2005.
8. E. Cosijn and P. Ingwersen. Dimensions of relevance. *IP&M*, 36(4):533–550, 2000.
9. C. A. Cuadra and R. V. Katter. Opening the black box of'relevance'. *J. Doc.*, 23(4):291–303, 1967.
10. M. Eisenberg and C. Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *JASIS*, 39(5):293–300, 1988.
11. D. B. Friedman, L. Hoffman-Goetz, and J. F. Arocha. Health literacy and the world wide web: comparing the readability of leading incident cancers on the internet. *Informatics for Health and Social Care*, 31(1):67–87, 2006.
12. L. Goeuriot, G. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salanterä, H. Suominen, and G. Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In *Proc. of CLEF*, 2013.
13. L. Goeuriot, L. Kelly, W. Lee, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, and H. M. Gareth J.F. Jones. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In *Proc. of CLEF*, Sheffield, UK, 2014.
14. P. Larsson. *Classification into readability levels: implementation and evaluation*. PhD thesis, Uppsala University, 2006.
15. D. R. McCallum and J. L. Peterson. Computer-based readability indexes. In *Proc. ACM Conf.*, pages 44–48, 1982.
16. S. Mizzaro. Relevance: The whole history. *JASIS*, 48(9):810–832, 1997.
17. A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):2, 2008.
18. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. Jones, M. Lupu, and P. Pecina. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *Proc. of CLEF*, 2015.
19. J. Palotti, G. Zuccon, and A. Hanbury. The influence of pre-processing on the estimation of readability of web documents. In *Proc. of CIKM*, 2015.
20. A. M. Rees and D. G. Schultz. A field experimental approach to the study of relevance assessments in relation to document searching. Technical report, CWR University, 1967.
21. S. E. Robertson. The probability ranking principle in ir. *J. Doc.*, 33(4):294–304, 1977.
22. T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proc. of SIGIR*, pages 1043–1052, 2011.
23. T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. In *Proc. of ASIS*, volume 34, pages 313–327, 1997.
24. L. Schamber and M. Eisenberg. Relevance: The search for a definition. In *Proc. of ASIS*, 1988.
25. M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. of SIGIR*, pages 95–104, 2012.
26. T. M. Walsh and T. A. Volsko. Readability assessment of internet-based consumer health information. *Respiratory care*, 53(10):1310–1315, 2008.
27. R. C. Wiener and R. Wiener-Pla. Literacy, pregnancy and potential oral health changes: The internet and readability levels. *Maternal and child health journal*, pages 1–6, 2013.
28. Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *JASIST*, 57(7):961–973, 2006.
29. X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *Proc. of CIKM*, pages 540–549, 2006.
30. E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proc. of SIGIR*, pages 587–594, 2008.
31. C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proc. of SIGIR*, pages 10–17, 2003.
32. Y. Zhang, L. A. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69, 2010.
33. Y. Zhang, J. Zhang, M. Lease, and J. Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proc. of SIGIR*, pages 435–444, 2014.
34. G. Zuccon and B. Koopman. Integrating understandability in the evaluation of consumer health search engines. *Proc. of MedIR*, pages 32–35, 2014.
35. G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can. In *Proc. of ECIR*, pages 562–567, 2015.