MM: A new Framework for Multidimensional Evaluation of Search Engines

Joao Palotti
Qatar Computing Research Institute
Doha, Qatar
jpalotti@hbku.edu.qa

Guido Zuccon
University of Queensland
St Lucia, Australia
g.zuccon@uq.edu.au

Allan Hanbury
TU Wien
Vienna, Austria
allan.hanbury@tuwien.ac.at

ABSTRACT

In this paper, we proposed a framework to evaluate information retrieval systems in presence of multidimensional relevance. This is an important problem in tasks such as consumer health search, where the understandability and trustworthiness of information greatly influence people's decisions based on the search engine results, but common topicality-only evaluation measures ignore these aspects. We used synthetic and real data to compare our proposed framework, named *MM*, to the understandability-biased information evaluation (UBIRE), an existing framework used in the context of consumer health search. We showed how the proposed approach diverges from the UBIRE framework, and how *MM* can be used to better understand the trade-offs between topical relevance and the other relevance dimensions.

ACM Reference Format:
Joao Palotti, Guido Zuccon, and Allan Hanbury. 2018. MM: A new Framework for Multidimensional Evaluation of Search Engines. In *2018 ACM Conference on Information and Knowledge Management (CIKM'18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269261

1 INTRODUCTION

Research has long established that the notion of relevance in information retrieval (IR) is multidimensional [1, 11]: the topicality of a document to a query or information need is central to the notion of relevance, but other factors (also called dimensions) that influence the relevance of a document do exist. These include novelty and diversity, timeliness, scope, understandability and trustworthiness, among others [10, 11]. In the context of consumer health search[1], in particular, the relevance dimensions of understandability and information trustworthiness are fundamental [4]. It means that health information is only valuable to users, allowing them to make appropriate health decision if it is understandable and correct. It is therefore important to take into account these relevance dimensions, along with topicality, when evaluating the effectiveness of search systems in the context of consumer health search tasks, and in general in other tasks with similar requirements.

An evaluation framework that integrates understandability into IR evaluation has been recently devised [12, 13] and it has been largely adopted to evaluate systems for consumer health search [8, 9, 14]. The framework, named *understandability-biased IR evaluation* (UBIRE), builds upon the gain-discount framework of evaluation measures used in IR (measures like normalised Discounted Cumulative Gain (nDCG), Expected Reciprocal Rank (ERR) and Rank Biased Precision metric (RBP) belong to this framework) [2]. UBIRE uses a discount based on the rank position at which documents are retrieved, and a gain function that integrates contributions from both topicality and understandability (see Section 2). The framework has been extended to integrate additional relevance dimensions such as trustworthiness [8]: since its extension is straightforward, without loss of generality, we refer to the UBIRE framework as the extended version capable of including in the gain function every dimension of relevance (provided certain assumptions are met).

A limitation of the approach used to model multidimensional relevance in UBIRE is that it is not trivial to identify how different dimensions of relevance affect the final evaluation score. This is because in UBIRE gains produced by documents for each of the considered dimensions of relevance are combined early on in the evaluation measure. This limitation makes the interpretation of evaluation results using UBIRE difficult as it is impossible to determine whether improvements (deteriorations) are due to more (less) understandable or more (less) topical documents being retrieved.

In this work, we propose an alternative to UBIRE, called the *MM* framework, which overcomes the interpretability limitation of UBIRE, while still enabling the combination of multidimensional relevance evidence when evaluating IR systems (Section 3). Using small synthetic data, we show the intuitive differences between UBIRE and *MM* and demonstrate how *MM* overcomes UBIRE's limitation (Section 4). We further empirically compare specific measures instantiated from the two frameworks using real data to study system ranking correlations across UBIRE and *MM* (Section 5). The results show that while system correlations measured with *MM* are aligned with UBIRE, *MM* provides richer information to researchers, allowing them to assess and control how each relevance dimension contributes to the evaluation score of a system.

2 INCORPORATING UNDERSTANDABILITY INTO EVALUATION METRICS

The understandability-biased IR evaluation framework (UBIRE) [12, 13] is based on the gain-discount framework [2] which models an

[1]This search task involves common people with no or limited medical knowledge searching for health advice on the web. This task is often carried out in time-sensitive and emotion-pressured circumstances [4, 6].

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269261

evaluation measure $\mathcal{M}$ as:

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

where $\mathsf{g}(d@k)$ and $\mathsf{d}(k)$ are respectively the *gain function* computed for the document at rank k and the *discount function* computed for the rank k. K is the depth of assessment at which measure $\mathcal{M}$ is evaluated, and $1/\mathcal{N}$ is a normalization factor, which serves to bound the value of the sum into the range [0,1] (details in [2]).

The gain-discount framework encompasses measures such as the normalized Discounted Cumulative Gain (nDCG) [5] with $\mathsf{g}(d@k) = 2^{P(R|d@k)} - 1$ and $\mathsf{d}(k) = 1/(log_2(1+k))$; the expected reciprocal rank (ERR) [3] with $\mathsf{g}(d@K) = (2^{P(R|d@k)} - 1)/2^{max(P(R|d))}$ and $\mathsf{d}(k) = 1/k$; and the Rank Biased Precision (RBP) with $\mathsf{g}(d@k)$ equal to 1 if $d@k$ is relevant and 0 otherwise and $\mathsf{d}(k) = \rho^{k-1}$ (with ρ representing the user persistence).

The gain provided by a document at rank k can be expressed as a function of its probability of relevance. Without loss of generality, $\mathsf{g}(d@k) = f(P(R|d@k))$, where $P(R|d@k)$ is the probability of relevance given the document at k. When only topical relevance is modelled, then $P(R|d@k) = P(T|d@k)$, i.e., the probability that the document at k is topically relevant. For binary relevance, this probability is 1 for relevant documents and 0 for non-relevant documents. For non-binary relevance, this probability can be distributed according to the number of relevance levels.

UBIRE extends this framework to consider cases where relevance is modelled beyond topicality so as to explicitly model other dimensions, such as understandability. This is done by modelling the probability of relevance $P(R|d@k)$ as the joint distribution over all considered dimensions, $P(\delta_1, \cdots, \delta_n|d@k)$, where each $\delta_i \in \mathcal{D}$ represents a dimension of relevance, e.g., topicality, understandability. The computation is simplified by assuming that dimensions are compositional events and their probabilities independent (see [12] for more details). The gain function with respect to different dimensions of relevance can then be expressed as:

$$\begin{aligned} \mathsf{g}(d@k) &= f(P(R|d@k)) \\ &= f\big(P(\delta_1, \cdots, \delta_n|d@k)\big) = f\Big(\prod_{i=1}^{n} P(\delta_i|d@k)\Big) \end{aligned}$$

Evaluation metrics developed within this framework differ by means of the instantiations of $f\big(P(\delta_1, \cdots, \delta_n|d@k)\big)$, other than by which dimensions are modelled. Zuccon provided an instantiation that considers both topicality and understandability [12]:

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

Specific implementations of the UBIRE framework that have been developed in previous work considered the basic gain and discount functions from RBP [7]; an instantiation with understandability [12, 13] has been later extended by jointly considering also trustworthiness [8]. For ease of explanation, we consider the formulation with topicality and understandability; similar considerations apply when also trustworthiness is modelled (as well as other dimensions). In this case, the understandability-biased RBP, *uRBP*, is defined as:

$$\begin{aligned} uRBP(\rho) &= (1-\rho) \sum_{k=1}^{K} \rho^{k-1} P(T|d@k) \cdot P(U|d@k) \\ &= (1-\rho) \sum_{k=1}^{K} \rho^{k-1} \mathsf{g}_{RBP}(d@k) \cdot \mathsf{g}_{U}(d@k) \end{aligned}$$

In the *uRBP*, the function $\mathsf{g}_{RBP}(d@k)$ is the same as the gain in RBP and transforms relevance values into the corresponding gains and, likewise, $\mathsf{g}_U(d@k)$ transforms understandability values into the corresponding gains. If $\mathsf{g}_U(d@k) = 1$ for every document, then only topical relevance affects retrieval evaluation, i.e. every document is considered as having equal understandability and we obtain the original RBP. Two instantiations of the gain function $\mathsf{g}_U(d@k)$ have been explored in previous work: one binary (*uRBP*) and the other graded (*uRBPgr*). In the binary version $\mathsf{g}_U(d@k) = 1$ if $P(U|d@k) \geq th_U$, where th_U is a threshold on the assessments of understandability (every assessment that is greater than or equal to th_U would generate a gain of 1), and $\mathsf{g}_U(d@k) = 0$ otherwise. In the graded version, understandability assessments are transformed into estimations of the probability function $P(U|d@k)$.

3 A NEW FRAMEWORK FOR MULTIDIMENSIONAL IR EVALUATION

A limitation of UBIRE is that it prematurely combines the gains contributed by each dimension of relevance in **one** single step, providing a unique evaluation score [12, 13]. While this allows for the comparison of systems, it does not permit to understand the contribution each dimension had on the evaluation measure. To overcome this limitation, we aim to create a measure which, while still allowing the modelling of multidimensional relevance, is of easy interpretation and for which it is straightforward to track the contribution each relevance dimension had on the final effectiveness score. This is achieved by separating the evaluation of each dimension such that a value for each dimension is calculated separately with respect to its gain and discount, and then these are combined into a unique effectiveness measure. Note that we assume that it is possible to evaluate each measure separately: while this is akin to the compositionality assumption in UBIRE, if that failed, UBIRE would use mixture models to compute the related probabilities, while the proposed measure would be instead likely undefined.

The evaluation of each relevance dimension separately is trivial, as it consists in applying the discount and gain function of the underlying evaluation measure, e.g. RBP, to each relevance dimension $\delta \in \mathcal{D}$, where the gains are those associated with the criteria for that specific dimension.

While the outputs of each relevance dimension could be combined with a linear or geometric combination of values, we opt to use the weighted harmonic mean, as it is particularly sensitive to a single lower-than-average value. The same intuition is used to combine recall and precision in the widely used F-measure. Given a (discount-gain) evaluation measure $\mathcal{M}$, we apply the measure to evaluate a list of documents l_δ which have been labeled with respect to dimension δ (i.e., we compute $\mathcal{M}(l_\delta)$). Then, to compute the proposed measure $MM_{\mathcal{M}}$, we combine all $\mathcal{M}(l_\delta)$ for each relevance dimension using the harmonic mean, where each dimension is weighted according to a preferential weight w_δ assigned to each dimension; formally:

$$MM_{\mathcal{M}} = \left(\frac{\sum_{\delta=1}^{n} w_\delta \cdot \mathcal{M}(l_\delta)^{-1}}{\sum_{\delta=1}^{n} w_\delta} \right)^{-1} = \frac{\sum_{\delta=1}^{n} w_\delta}{\sum_{\delta=1}^{n} \frac{w_\delta}{\mathcal{M}(l_\delta)}} \qquad (1)$$

Without loss of generality, we instantiate $\mathcal{M} = RBP$ and define the following modification of RBP [7] for each dimension:

- $RBP_t(\rho)$: uses binary topicality assessments (i.e. the usual RBP).
- $RBP_u(\rho)$: uses understandability assessments (either graded or binary; see below for specific instantiations).

Thus Equation 1 becomes (we assumed $w_t = w_u$):

$$MM_{RBP(\rho)} = 2 \cdot \frac{RBP_t(\rho) \cdot RBP_u(\rho)}{RBP_t(\rho) + RBP_u(\rho)} \qquad (2)$$

4 COMPARING FRAMEWORKS THROUGH SYSTEM SIMULATIONS

To understand the behaviour of UBIRE and *MM* when facing different IR systems, we first employed synthetic systems so as to have a fine-grained control over our experiments. This allowed to know a priori what has changed between two system instances and study the effect these changes had on evaluation. In our experiments, along with topicality, we considered understandability, leaving the (trivial) extension to other dimensions to later work. In the following simulations we controlled the amount of topical documents and understandable documents retrieved. We did so by following this two-phase procedure:

(1) **Topicality Phase:** we controlled the amount of topical documents in a simulated run using a random variable T, $0 \leq T \leq 1$. We constructed a synthetic run by drawing a real number N_i, $0 \leq N_i \leq 1$, for each position i in a ranking. If $N_i \leq T$, we marked the document at position i as relevant, otherwise, we marked it as not relevant. It is expected that a run generated with $T = 0.1$ has 10% of the documents assessed as relevant (90% as non-relevant), while a run with $T = 0.5$ has as many relevant as non-relevant documents.

(2) **Understandability Phase:** we controlled the level of understandability of the documents in a synthetic run. In order to create and control the randomness of our synthetic systems, we generated understandability labels using a Gaussian distribution with pre-defined mean μ and variance σ. As previously done in consumer health search collections [8, 14], we forced the understandability labels to be in the interval $[0, 100]$. We fixed a relatively large variance, $\sigma = 40$, to mimic results of previous collections in which the understandability labels had a large variance [14], and we varied the mean μ of the Gaussian from 0 to 100. Figure 1 shows the expected label distribution for $\mu = 20, 50, 80$, i.e., $\mathcal{N}(20, 40)$, $\mathcal{N}(50, 40)$ and $\mathcal{N}(80, 40)$. In Figure 1 we also included the threshold U used to compute RBP_u (Section 3).

We executed these two phases in succession. In total, we generated 1,000 runs for each value of T and μ.

We calculated *uRBP* (using UBIRE) and MM_{RBP} for each synthetic system. Table 1 shows the average result for different values of T (rows), i.e., different expected number of topical documents retrieved, and values of μ (cols), i.e., different understandability distributions. A smaller μ means that more understandable documents were retrieved. The results show that as the expected number of topical documents (T) increases, *RBP* increases. Likewise, *uRBP* increases, as it is bounded by topical relevance. In turn, increasing T has no effect on RBP_u, but increases MM_{RBP}, as it is also directly dependant on *RBP*. When the number of understandable documents retrieved is increased (i.e., μ decreased), *RBP* stays constant, as it does not measure how understandable documents are. In turn, *uRBP*, RBP_u and MM_{RBP} increase. These are the expected behaviours of the considered measures.

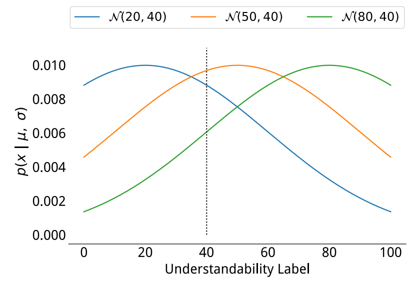

Figure 1: Gaussian distribution for different μ: higher μ generates higher understandability labels (more difficult documents were retrieved). Here, only documents with understandability lower than 40 are considered easy-to-understand (threshold shown as dotted line).

We further focused our attention to the results of specific experiments highlighted in blue and yellow in Table 1. These cases simulated an initial system S1 that exhibited the results in blue (condition $T = 0.6$ and $\mathcal{N}(40, 40)$) being modified to improve the understandability of retrieved documents ($\mathcal{N}(30, 40)$) at the expenses of topicality ($T = 0.5$), producing a new system S2. The effectiveness of S2 is highlighted in yellow.

If *RBP* and *uRBP* were used to decide whether S2 should be preferred over the initial system S1, then S2 would be discarded and S1 preferred, as S2 produced a 16% reduction in *RBP* and a 6% reduction in *uRBP*. With these results, an IR researcher would conclude that the modifications in S2 did not pay off.

If MM_{RBP} was used instead, the IR researcher would have been able to gain more insights about system effectiveness and the trade-off between understandability and topicality. To use MM_{RBP}, RBP_t (= *RBP*) and RBP_u needed to be computed. Between S1 and S2, there was a decrease in RBP_t of 16%; but conversely RBP_u increased by 20%: this clearly allows the trade-off between topicality and understandability to be gauged.

When *RBP* and RBP_u were combined within MM_{RBP}, if both dimensions were given equal weight, then systems S1 and S2 obtained the same MM_{RBP}. Note that *MM* can be adapted to specific circumstances: if topicality is more important than understandability, then the weights of each dimension can be changed accordingly in the harmonic mean computation.

5 RANK CORRELATIONS

Next, we compared the behaviours of *MM* and UBIRE using real data. For this, we used the systems participating to the CLEF eHealth

Table 1: We varied T, the expected proportion of topically relevance documents (rows), and the mean μ of Gaussian distribution used to generate understandability labels (columns). A smaller μ means that easier to read documents are retrieved. We showed the average and standard deviation of each experiment.

T	Understandability $\mathcal{N}$(50,40)				Understandability $\mathcal{N}$(40,40)				Understandability $\mathcal{N}$(30,40)			
	RBP	uRBP	RBP_u	MM_{RBP}	RBP	uRBP	RBP_u	MM_{RBP}	RBP	uRBP	RBP_u	MM_{RBP}
.3	.29 ± .15	.15 ± .09	.39 ± .17	.30± .12	.29 ± .15	.17 ± .11	.50 ± .16	.34 ± .14	.29 ± .15	.19 ± .12	.61 ± .16	.36 ± .15
.4	.39 ± .17	.20 ± .11	.40 ± .17	.36 ± .14	.39 ± .17	.22 ± .12	.48 ± .17	.40 ± .13	.39 ± .17	.25 ± .13	.60 ± .16	.44 ± .14
.5	.50 ± .17	.25 ± .11	.42 ± .16	.42 ± .13	.50 ± .17	.29 ± .12	.50 ± .17	.47 ± .13	.50 ± .17	.33 ± .14	.60 ± .17	**.52 ± .13**
.6	.60 ± .16	.30 ± .12	.41 ± .16	.46 ± .14	.60 ± .16	.35 ± .12	.50 ± .17	**.52 ± .13**	.60 ± .16	.40 ± .13	.61 ± .17	.58 ± .13
.7	.70 ± .15	.36 ± .12	.41 ± .17	.49 ± .15	.70 ± .15	.41 ± .13	.51 ± .17	.56 ± .14	.70 ± .15	.46 ± .13	.59 ± .16	.62 ± .12

Table 2: Kendall-τ correlation for systems participating in CLEF eHealth 2015 and 2016.

	CLEF 2015				CLEF 2016			
	RBP	uRBP	RBP_u	MM_{RBP}	RBP	uRBP	RBP_u	MM_{RBP}
RBP	1.000	0.901	0.483	0.843	1.000	0.948	0.497	0.850
uRBP	0.901	1.000	0.563	0.901	0.948	1.000	0.456	0.866
RBP_u	0.483	0.563	1.000	0.610	0.497	0.524	1.000	0.633
MM_{RBP}	0.843	0.901	0.610	1.000	0.850	0.866	0.633	1.000

IR Lab evaluations in 2015 and 2016 [9, 14]. In both these evaluation challenges, systems were officially evaluated using *uRBP* – we further evaluated each system using *MM* and studied the correlations among system rankings obtained using RBP (thus considering topicality only), *uRBP* (UBIRE), and our proposed RBP_u (thus considering only understandability) and MM_{RBP}. This investigation of correlations is a common approach to compare and understand relative behaviour of evaluation measures [12].

Specifically, we studied a setting where understandability was binary, akin to topicality, which also was considered as binary. For topicality, this was achieved using the common gain function for RBP that only models binary relevance: graded relevance labels were conflated to binary such that highly relevant and relevant assessments were mapped to relevant, and the rest to irrelevant. For understandability, the binarisation of the assessments was dependant on the year of the challenge. For 2015, understandability assessments were made on a 4-point scale (*very easy*, *easy*, *hard* and *very hard*) [9]: we made this binary by assuming that a document was understandable if assessed as *very easy* or *easy*, and not-understandable otherwise. For 2016, understandability assessments were made on an integer scale ranging from 0 (very easy) to 100 (very hard) [14]: we made this binary by assuming that documents with an assessment lower than or equal to 40 were understandable, while we made the remaining as not-understandable.

Table 2 shows the Kendall-τ rank correlations of systems according to RBP, *uRBP*, RBP_u and MM_{RBP}. Rank correlation between RBP and *uRBP* was high for both 2015 and 2016 data. This emphasises the tight relation between RBP and *uRBP*. On the other hand, MM_{RBP} exhibited the strongest rank correlation with RBP_u, while the correlation between RBP_u and RBP or *uRBP* is marginal. In addition, we found that MM_{RBP} strongly correlated with RBP, but not as strongly as *uRBP* does. Finally, MM_{RBP} and *uRBP* showed a generally high correlation among themselves, highlighting that the two measures provided similar evaluations of system effectiveness; however, MM_{RBP} had the advantage that the trade-off between topicality and understandability could be clearly identified and studied.

6 CONCLUSION

In this paper, we proposed a new framework, called *MM*, to evaluate search engines when multidimensional relevance should be considered. Using both synthetic and real data, we compared *MM* to the understandability-biased information retrieval evaluation framework (UBIRE), which has recently been used to evaluate search systems in the consumer health search domain.

Our experiments showed that *MM* correlated well with UBIRE and that both had an equivalent power to rank and distinguish good systems. However, *MM* has the advantage of being more intuitive and allowing experimenters to easily understand what relevance dimensions are affecting their systems performance, as well as carefully tune the trade-off between topical relevance and other dimensions. This is important because it allows search engine practitioners to better debug their systems and tackle the understandability/trustworthiness of the ranked information.

While our empirical experiments only considered understandability as an additional dimension to relevance, this was done for directly comparing with UBIRE, and by definition *MM* naturally accommodates an unlimited number of relevance dimensions. An open question is whether *MM* correlates with human preferences and how it compares with UBIRE in this respect. To answer this, future work will consider user-based validation and comparison of the two multidimensional evaluation frameworks.

REFERENCES

[1] P Borlund. 2003. The concept of relevance in IR. *JASIST* 54, 10 (2003), 913–925.
[2] B Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *SIGIR*. ACM, 903–912.
[3] Olivier Chapelle, Donald Meltzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *CIKM*. ACM, 621–630.
[4] W Hersh. 2008. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media.
[5] K Järvelin and J Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446.
[6] N Mishra, R White, S Ieong, and E Horvitz. 2014. Time-critical search. In *SIGIR*. ACM, 747–756.
[7] A Moffat and J Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (2008), 27 pages.
[8] J Palotti, G Zuccon, et al. 2017. CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In *CLEF*.
[9] J Palotti, G Zuccon, L Goeuriot, et al. 2015. ShARe/CLEF eHealth Evaluation Lab 2015, Task 2: User-centred Health Information Retrieval. In *CLEF*.
[10] T K Park. 1993. The Nature of Relevance in Information Retrieval: An Empirical Study. *The Library Quarterly* 63, 3 (1993), 318–351.
[11] L Schamber. 1994. Relevance and information behavior. *ARIST* 29 (1994).
[12] G Zuccon. 2016. Understandability biased evaluation for information retrieval. In *ECIR*. Springer, 280–292.
[13] G Zuccon and B Koopman. 2014. Integrating Understandability in the Evaluation of Consumer Health Search Engines. In *MedIR*.
[14] G Zuccon, J Palotti, L Goeuriot, et al. 2016. The IR Task at the CLEF eHealth evaluation lab 2016. In *CLEF*, Vol. 1609. 15–27.