Balancing Relevance Criteria through Multi-Objective Optimization

Joost van Doorn[1]
joost.vandoorn@student.uva.nl

Daan Odijk[1]
d.odijk@uva.nl

Diederik M. Roijers[1,2]
diederik.roijers@cs.ox.ac.uk

Maarten de Rijke[1]
derijke@uva.nl

[1]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[2]Department of Computer Science, University of Oxford, Oxford, United Kingdom

ABSTRACT

Offline evaluation of information retrieval systems typically focuses on a single effectiveness measure that models the utility for a typical user. Such a measure usually combines a behavior-based rank discount with a notion of document utility that captures the single relevance criterion of topicality. However, for individual users relevance criteria such as credibility, reputability or readability can strongly impact the utility. Also, for different information needs the utility can be a different mixture of these criteria. Because of the focus on single metrics, offline optimization of IR systems does not account for different preferences in balancing relevance criteria.

We propose to mitigate this by viewing multiple relevance criteria as objectives and learning a set of rankers that provide different trade-offs w.r.t. these objectives. We model document utility within a gain-based evaluation framework as a weighted combination of relevance criteria. Using the learned set, we are able to make an informed decision based on the values of the rankers and a preference w.r.t. the relevance criteria. On a dataset annotated for readability and a web search dataset annotated for sub-topic relevance we demonstrate how trade-offs between can be made explicit. We show that there are different available trade-offs between relevance criteria.

Keywords

Multi-objective optimization; Learning to rank

1. INTRODUCTION

The primary goal of information retrieval (IR) systems is to satisfy the information need of a user. Search engines today are fairly successful in finding topically relevant pages. To achieve this most search engines are optimized to rank documents based on their topical relevance to the query. In an offline optimization setting relevance is typically determined by experts, and evaluated with utility-based metrics such as nDCG, which tends to focus on optimizing a single aspect of utility. In online optimization, feedback is collected implicitly for all relevance criteria. However, this approach may ignore differences between individual users and information needs by aggregating across all users and queries. Often aggregation works well, but not always. E.g., users that have limited vocabularies (e.g., children) can benefit from search results optimized for their reading level. When people look for medical information on the web they would benefit from accurate and reliable information, more so than when looking for information on a Star Wars movie.

Utility depends on many factors aside from topicality; criteria such as credibility, reputability and readability are also important [14]. While their importance is typically secondary to topicality, there clearly is a benefit in many use cases. A learning to rank approach [16] can be used to learn an optimal ranker for a specified weighted preference over criteria. Similarly, data fusion techniques can combine ranked lists that are optimized for a certain notion of utility. But what if we want to optimize for multiple criteria, without knowing their relative importance beforehand? For instance, how should we balance relevance and readability if we do not know who our user will be? Or relevance and sub-topic relevance?

We draw inspiration from multi-objective optimization techniques to answer these questions, i.e., to find a set of rankers for which each solution is optimal for a different trade-off in the relevance criteria. We combine the multi-objective technique *Optimistic Linear Support (OLS)* [12] with multiple utility-based metrics in a learning-to-rank setting. We consider two scenarios with two relevance criteria for which we optimize a set of rankers. We evaluate our approach on two datasets, one annotated for relevance and readability, and one annotated for relevance and diversity. To learn our rankers we apply dueling bandit gradient descent with a point-wise ranking function. To optimize for diversity we subsequently apply MMR and cluster-based ranking.

2. BACKGROUND

The concept of relevance is core to IR and a much debated subject. Park [9] gives an extensive analysis on the nature of relevance in IR, and argues that relevance is intrinsically related to the selection process by the user. Cooper [4] states that each query could represent a set of specific questions as part of the information need, where documents are relevant if they provide an answer to one of these specific questions. Schamber [14] identifies major criterion groups for relevance: aboutness, currency, availability, clarity and credibility. There is a general trend that relevance cannot be attributed to just one factor such as topicality, but is multi-factored [9].

Many metrics have been proposed to measure the effectiveness of an IR system; we focus on metrics based on the concept of utility [2, 4]. The utility of an IR system depends on all factors that determine the usefulness for each specific user. Cooper [4] defines utility as "A catch all concept involving not only topic relatedness

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914708

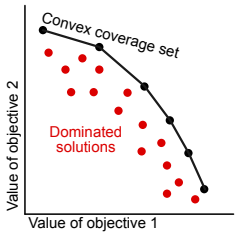

Figure 1: The points on the line represent solutions in the coverage set, the others are dominated.

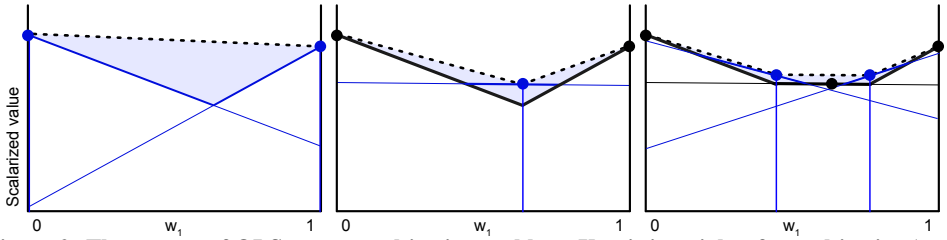

Figure 2: Three steps of OLS on a two-objective problem. X-axis is weight of one objective (note: $w_2 = 1 - w_1$), y-axis the scalarized value. Blue area highlights difference between upper bound and convex value. As more solutions are added to the CCS, the difference is iteratively reduced.

but also quality, novelty, importance, credibility, and many other things." Utility-based metrics combine a notion of utility with specific assumptions about user behavior [2]. Each document has a specific numerical utility value for an information need. Additionally, a discount function is used on the document's rank, under the assumption that more effort is needed to reach lower ranked documents, and it is less likely for the user to reach these documents. Many metrics, therefore, boil down to the same basic formula to estimate the utility of the ranked list, composed of a sum of the product of document utility and a discount factor [3]:

$$M = \sum_{k=1}^{K} gain(rel_k) \times discount(k) \quad (1)$$

Extensions focus on multiple criteria. E.g., Dai et al. [5] present an extension of nDCG for freshness and relevance. Zuccon [17] proposes on an extension of rank-based precision for readability.

Similarly, diversity and novelty metrics also take into account multiple criteria in the form of subtopic relevance [3]. The underlying assumption is that there are multiple subtopics (or categories) for each query, and each user will be interested in results for at least one of these subtopics. Relevance assessments are provided for each of the subtopics belonging to a query separately. These are combined based on the probability $p(i|q)$ of intent i being intended by the user for query q [1]. α-nDCG extends (1) with a weighted sum over subtopics, given p_i as the probability of each subtopic:

$$\alpha\text{-nDCG} = \frac{1}{\mathcal{N}} \sum_{i=1}^{M} p_i \sum_{k=1}^{K} gain_i^k \times discount(k) \quad (2)$$

While there has been previous work that combines multiple relevance criteria in the utility-based evaluation framework, to the best of our knowledge, no previous work uses multi-objective optimization techniques on information retrieval problems, i.e., existing methods do not return a *set* of alternative rankers with different available trade-offs with respect to the different relevance criteria.

3. MULTI-OBJECTIVE OPTIMIZATION

Scalarization function. We assume that a ranker has a value for each different relevance criterion, i.e., each ranker has an associated value vector $\mathbf{V}$, with a value, V_i in each criterion i. We follow [11] and assume that the preference of an individual user with respect to these criteria can be expressed in terms of an unknown *scalarization function* f, that collapses the value vector to a scalar utility: $f(\mathbf{V}, \mathbf{w})$, where $\mathbf{w}$ is a vector that parameterizes f. We are unable to observe this function directly. Instead, we aim to find a *coverage set* [11] of rankers, that contains an optimal trade-off for each possible preference (i.e., f and $\mathbf{w}$) that a user might have, see Fig. 1.

We assume that f is linear (where weighted means: $\sum_i^C w_i = 1$):

$$f(\mathbf{V}, \mathbf{w}) = \mathbf{w}^T \mathbf{V}, \quad (3)$$

i.e., the utility for the user is a *weighted* sum over relevance criteria.

Metrics as objectives. To formulate our own scalarization function we can combine (1) with (3):

$$M = \sum_{i=1}^{C} w_i \sum_{k=1}^{K} gain_i(doc_k) \times discount(k) \quad (4)$$

This definition is similar to the definition of α-nDCG of (2), where instead of a sum over topics we have a sum over C relevance criteria. In α-nDCG, the metric M would subsequently be normalized. It is, however, not desirable to normalize the linear scalarized value function as this would remove the convex property of the value vector that is required for efficient optimization using OLS. The linear scalarization function does require values that are comparable in their magnitude, therefore, the individual value functions are normalized with normalization value N_i instead, giving:

$$V_i = \frac{1}{N_i} \sum_{k=1}^{K} gain_i(doc_k) \times discount(k) \quad (5)$$

Convex coverage set. Because each criterion contributes positively to the scalarization function, and we are interested in the *relative* importance of each criterion, we can assume that $\mathbf{w}$ is a positive vector that sums to 1 in order to determine a coverage set. A coverage set that covers all possible linear scalarizations is called a convex coverage set (CCS) [11]. To compute the (approximate) CCS, we build on the Optimistic Linear Support (OLS) framework for multi-objective optimization [12]. Fig. 2 illustrates the OLS procedure; OLS computes a CCS by solving a multi-objective problem as a series of single-objective problems, i.e., problems that are scalarized using different $\mathbf{w}$. At each iteration, OLS tries to find a new ranker, thereby incrementally building a CCS. We can use existing single-objective optimization techniques to find rankers for a given $\mathbf{w}$. We use Dueling Bandit Gradient Descent (DBGD) [16].

Each ranker found in an iteration of OLS has an associated value vector $\mathbf{V}$. For each $\mathbf{w}$, the scalarized value of a ranker is $V_{\mathbf{w}} = \mathbf{w} \cdot \mathbf{V}$. Given a partial CCS, i.e., the set S of rankers and associated $\mathbf{V}$ found so far, we define the scalarized value function as a function of $\mathbf{w}$:

$$V_S^*(\mathbf{w}) = \max_{\mathbf{V} \in S} \mathbf{w} \cdot \mathbf{V},$$

i.e., the scalarized value function is the convex upper surface of the vectors in Fig. 2. OLS selects the next $\mathbf{w}$ from the *corner weights* of $V_S^*(\mathbf{w})$, i.e., those $\mathbf{w}$ where $V_S^*(\mathbf{w})$ changes slope. In Fig. 2 the corner weights evaluated in that iteration are indicated by the blue vertical lines. The maximal possible improvement in scalarized value on the partial CCS is indicated by the dashed lines above $V_S^*(\mathbf{w})$. Once it reaches a corner weight, OLS is provably optimal as long as the single-objective method it employs to solve the scalarized problems is exact [12]. In practice, exact single-objective subroutines are not required; we can safely use DBGD, but with lesser guarantees of the optimality of the solution [13].

Reuse and iterated search scheme. A limitation of applying standard OLS is that for every corner weight DBGD needs to be run. This can be expensive, depending on the size of the dataset. However, this can be mitigated by hot-starting the single-objective optimization algorithm with parts of previously found solutions (following [13]). For each new corner weight, we multi-start DBGD, starting from the rankers that were found at the 3 closest corner weights so far. It is possible that DBGD does not find a new best

solution, even though such a solution might still exist. If this is the case for a number of iterations, we take a random pertubation step. DBGD is stopped automatically after 40,000 iterations, or if no improvement has been found after 20 random pertubations. To our knowledge, this is the first time such a Multi-Start/Iterated Local Search scheme [7] has been combined with OLS.

4. EXPERIMENTAL SET-UP

To demonstrate how multi-objective optimization for balancing multiple relevance criteria works in practice, we perform experiments on two datasets: (i) balancing readability and topical relevance in a health setting (CLEF eHealth 2015 task 2 [8]), and (ii) balancing diversity and topical relevance in a web search dataset annotated for sub-topic relevance (TREC 2012 Web Track diversity task). While our runs are competitive, our main goal is to find multiple solutions that balance different relevance criteria, which we report in the form of a CCS.

CLEF eHealth. CLEF eHealth 2015 task 2 provides annotations for two objectives. It is composed of 5 training queries and 67 test queries; annotations are provided for relevance only for the training queries, and both relevance and understandability for the test queries. As the extra understandability annotations are required to optimize for both relevance and understandability at the same time we only use the test set queries for optimization. To measure readability document text is extracted using boilerpipe [6]. Since simple readability metrics do not correlate well with actual readability in the medical domain [15], another approach to readability is required. We compiled a list of medical terms and counted their occurrences. This list was taken from an English Wikipedia page, for which all words contained in the 20k English word list from the Google Trillion Word Corpus were filtered out.[1] Using this feature, and, additionally, the Coleman-Liau index, Gunning fog index, and document length, we trained an SVM to predict the understandability score. For the CLEF eHealth 2015 task 2 the usual metrics are RBP, uRBP and uRBPgr. In preliminary experiments we found a strong correlation between RBP and uRBP, like [17]. Hence, we optimize for nDCG using relevance annotations (nDCG_{rel}), and also for nDCG using understandability annotations ($\text{nDCG}_{\text{read}}$). We normalize $\text{nDCG}_{\text{read}}$ so that the value is in the same range as nDCG_{rel}.

TREC Web Diversity. The TREC 2012 Web Track diversity task comes with 50 queries, with sub topic relevance assessments provided for the first 20 documents produced by the participating systems. TREC 2010 and 2011 Web Track diversity task queries were used for training. For diversity we use MMR and cluster-based ranking [10] with cosine similarity on TF-IDF vectors. We only apply MMR on the first T clusters. Documents are first scored based on relevance, subsequently, MMR and cluster-based ranking rerank the documents for diversity, which produces a $rank_i$ for each document i. Using $rank_i$, the final ranking is determined based on $(1 - w_d)score_i + w_d \frac{1}{rank_i}$, where w_d is a parameter that balances diversity and relevance. The usual metrics reported for the TREC Diversity task are nDCG and α-nDCG. For optimization we use both nDCG and α-nDCG to optimize for both relevance and diversity. As clustering introduces a lot of randomness, we average over 5 runs of DBGD. For value functions V_{rel} and V_{div}, we normalize the values of nDCG and α-nDCG, respectively, to $[0, 1]$.

5. RESULTS

CLEF eHealth 2015 task 2. For this task, we simultaneously optimize for readability and relevance, using nDCG for both metrics.

[1]For this list see: github.com/JoostvDoorn/moo-sigir2016

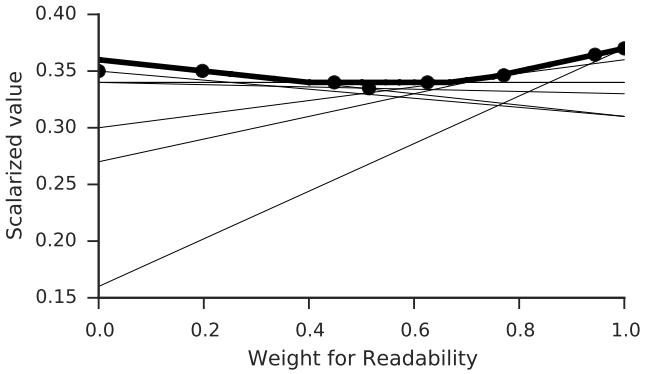

Figure 3: The CCS found on CLEF eHealth 2015, with a scalarized value based on relevance and readability. Absolute left having maximum weight on relevance, and absolute right maximum weight on readability.

The scalarized value of a solution was calculated using a weighted interpolation between value functions V_{rel} and V_{read} (Eq. 3).

The convex coverage set (CCS), constructed using OLS, is shown in Fig. 3. OLS finds eight solutions of which six are not dominated. The set of solutions is reported in Table 1 with their RBP and uRBP scores. We note that our best uRBP score is above the second run for the original task and the best nDCG_{rel} is in the top-5 out of 110 runs. We observe that we are able to find solutions that optimally combine the nDCG_{rel} and $\text{nDCG}_{\text{read}}$ objectives given different preferences for readability. E.g., with a w_{read} of 0.626, we obtain a 5% increase in $\text{nDCG}_{\text{read}}$, with an 8% loss compared the best solution in terms of nDCG_{rel} ($w_{read} = 0.197$). uRBP combines both objectives, and is highly correlated to RBP [17]. Due to this correlation, using RBP with uRBP would not find all rankers that offer the best available trade-offs between relevance and readability; the solution would be biased toward relevance. We therefore conclude that uRBP is not suitable for all possible preferences that a user might have.

Table 1: Evaluation of the solutions from the CCS on eHealth.

w_{read}	nDCG_{rel}	$\text{nDCG}_{\text{read}}$	RBP	uRBP	uRBPgr
0.000	0.350	0.783	0.392	**0.342**	0.337
0.197	**0.364**	0.777	**0.397**	0.340	**0.339**
0.448	0.344	0.807	0.371	0.327	0.324
0.514	0.343	0.804	0.372	0.327	0.324
0.626	0.335	0.814	0.369	0.326	0.324
0.771	0.298	0.832	0.333	0.294	0.294
0.944	0.266	**0.840**	0.304	0.270	0.269
1.000	0.157	**0.840**	0.189	0.160	0.159

To further analyze the effect of different weights for the readability objective, we analyze the annotations at each position in the ranking averaged over topics. Fig. 4 shows for three solutions in the CCS. The ranker optimized for readability does not show documents with higher relevance annotations in the top positions, whereas the other rankers are able to place more relevant documents at the top (similarly for readability). We observe that each ranker is suitable for their specific scalarization function, and as such our method is effective in balancing different relevance criteria.

TREC 2012 diversity task. For this task, we simultaneously optimize for overall relevance and sub-topic relevance by linearly combining value functions based on nDCG and α-nDCG. The CCS from OLS is shown in Fig. 5. The results from the points in the CCS on the TREC 2012 diversity task are shown in Table 2. Fewer solutions were found for the CCS compared to the readability task,

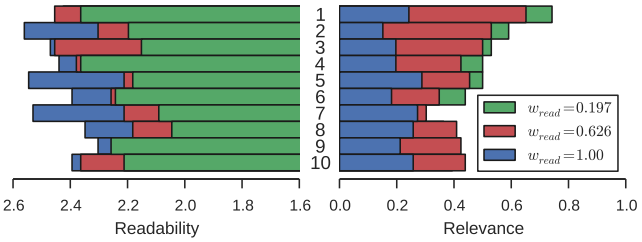

Figure 4: Average readability and relevance annotations on each rank for three different solutions in the CCS.

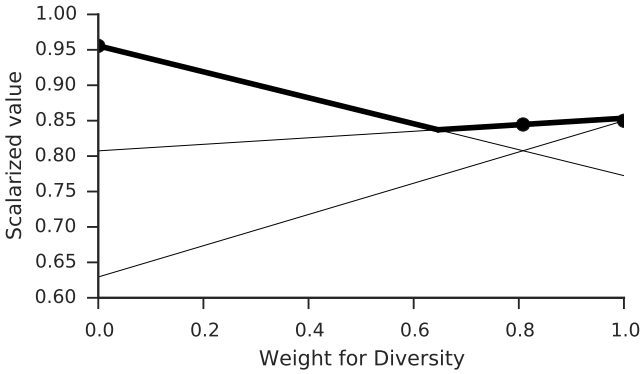

Figure 5: The CCS found on the TREC 2010 and 2011 datasets, with a scalarized value based on relevance and diversity. Absolute left having maximum weight on relevance, and absolute right maximum weight on diversity.

furthermore the differences between the values in Table 5 are also quite small, suggesting only a small trade-off between the objectives. As such this setting seems less suitable for our method. In terms of α-nDCG, the solutions on the test set (TREC 2012) are mid-performers compared to the original participants (the best solution is above the fourth of nine participants). The overall nDCG score would have ranked second. During training (TREC 2010–2011), the intermediate solution that OLS found obtains the same α-nDCG score with an increase in nDCG, compared to the solution optimized only for α-nDCG, see Fig. 5. We therefore conclude that our approach finds more balanced and better solutions, than if we would optimize for a single objective.

Table 2: Evaluation of the solutions from the CCS on the TREC 2012 dataset.

w_{div}	nDCG	α-nDCG
0.000	**0.236**	0.489
0.808	0.229	**0.493**
1.000	0.204	0.480

6. DISCUSSION

We demonstrated how to optimize rankings for multiple objectives by proposing a multi-objective approach based on optimistic linear support and DBGD for learning to rank. Because DBGD may get stuck in a local minimum we proposed an iterated local search schema for DBGD, and reuse of rankers inside OLS in order to make our algorithm more efficient. Using this approach, we have found multiple optimal rankers on the CLEF eHealth 2015 task 2 and on the TREC diversity task that offer different trade-offs w.r.t. different relevance criteria. These multiple optimal rankers are more flexible than a one-size-fits-all ranker produced by a standard learning to rank approach, and our work therefore forms an important step for flexibly optimizing search when multiple criteria are in play.

As to future work, one important issue is exposing different solutions to the user, or using different solutions to select the desired one. Exposing the user to multiple solutions can be done using additional user interface elements, or based on profiling of the user or adapting per query. The number of user interface controls provided in generic search engines is very minimal; specialized search engines are more likely to benefit from optimizing their controls based on these multi-objective criteria. Future work may also investigate what a good scalarization function is, as others may exist and be more suited, and which metrics are more suitable for linear combination. Many current evaluation metrics are highly correlated with relevance and as such may not always provide the flexibility to get a large CCS.

Acknowledgments. This research was supported by Ahold, Amsterdam Data Science, the Bloomberg Research Grant program, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the ESF Research Network Program ELIAS, the Royal Dutch Academy of Sciences (KNAW) under the Elite Network Shifts project, the Microsoft Research Ph.D. program, the Netherlands eScience Center under project number 027.012.105, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.001.109, 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.-066.930, CI-14-25, SH-322-15, 652.002.001, 612.001.551, the Yahoo Faculty Research and Engagement Program, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM'09*, pages 5–14. ACM, 2009.

[2] B. Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *SIGIR'11*, pages 903–912. ACM, 2011.

[3] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *WSDM'11*, pages 75–84. ACM, 2011.

[4] W. S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971.

[5] N. Dai, M. Shokouhi, and B. D. Davison. Learning to rank for freshness and relevance. In *SIGIR'11*, pages 95–104. ACM, 2011.

[6] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *WSDM'10*, pages 441–450. ACM, 2010.

[7] H. R. Lourenço, O. C. Martin, and T. Stützle. *Iterated local search*. Springer, 2003.

[8] 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 *CLEF '15*. Springer, 2015.

[9] T. K. Park. The nature of relevance in information retrieval: An empirical study. *The Library Quarterly*, 63(3):318–351, 1993.

[10] F. Raiber and O. Kurland. The Technion at TREC 2013 web track: Cluster-based document retrieval. Technical report, Technion, Israel, 2013.

[11] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[12] D. M. Roijers, S. Whiteson, and F. A. Oliehoek. Computing convex coverage sets for faster multi-objective coordination. *Journal of Artificial Intelligence Research*, 52:399–443, 2015.

[13] D. M. Roijers, S. Whiteson, and F. A. Oliehoek. Point-based planning for multi-objective POMDPs. In *IJCAI'15*, 2015.

[14] L. Schamber and J. Bateman. User criteria in relevance evaluation: Toward development of a measurement scale. In *ASIS'96*, volume 33, pages 218–25. ERIC, 1996.

[15] X. Yan, D. Song, and X. Li. Concept-based document readability in domain specific information retrieval. In *CIKM'06*, pages 540–549. ACM, 2006.

[16] Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *ICML'09*, pages 1201–1208. ACM, 2009.

[17] G. Zuccon. Understandability biased evaluation for information retrieval. In *ECIR'16*. Springer, 2016.