Estimating Topic Difficulty Using Normalized Discounted Cumulated Gain

Lukas Gienapp
Leipzig University

Benno Stein
Bauhaus-Universität Weimar

Matthias Hagen
Martin-Luther-Universität Halle-Wittenberg

Martin Potthast
Leipzig University

ABSTRACT

Information retrieval evaluation has to consider the varying "difficulty" between topics. Topic difficulty is often defined in terms of the aggregated effectiveness of a set of retrieval systems to satisfy a respective information need. Current approaches to estimate topic difficulty come with drawbacks such as being incomparable across different experimental settings. We introduce a new approach to estimate topic difficulty, which is based on the ratio of systems that achieve an NDCG score that is better than a baseline formed as random ranking of the pool of judged documents. We modify the NDCG measure to explicitly reflect a system's divergence from this hypothetical random ranker. In this way we achieve relative comparability of topic difficulty scores across experimental settings as well as stability to outlier systems—features lacking in previous difficulty estimations. We reevaluate the TREC 2012 Web Track's ad hoc task to demonstrate the feasibility of our approach in practice.

ACM Reference Format:
Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Estimating Topic Difficulty Using Normalized Discounted Cumulated Gain. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412109

1 INTRODUCTION

Information retrieval systems are typically evaluated under the Cranfield paradigm with the following three components [21]: (1) a set D of documents; (2) a set T of topics (i.e., information needs), each comprising a query and a characterization of what constitutes relevant documents; and (3) relevance judgments for the k top-ranked results from each system for each topic. An evaluation's outcome in this scenario can be seen as a matrix $M = S \times T$ with the set S of participating systems $s_1, \ldots, s_n$ that contribute a run for each topic $t_1, ..., t_m \in T$. The cell M_{ij} of the matrix then denotes the effectiveness score of system s_i on topic t_j under some measure based on the relevance judgements. The overall effectiveness of a system is its row's average score, while a topic's difficulty is usually estimated by a column-wise aggregation [11, 16].

Inherently, some topics will be easier for some systems than for others [7]. While many studies aim at estimating a topic's/query's difficulty online, with the goal of adapting a system's retrieval strategy accordingly [2, 15, 19], estimating a topic's difficulty offline is important to analyze retrieval experiments. Damessie et al. [5] observe that topic difficulty influences annotator agreement, while Mizzaro and Robertson [11] have shown that a topic's difficulty correlates with its ability to predict system effectiveness.

Reviewing the existing aggregation approaches to estimate topic difficulty, we identify four issues that limit their usefulness (Section 2). To overcome these issues, we introduce a new ratio-based approach using on the widely accepted NDCG measure. We formally prove its shift and scale invariance (Section 3) and then adjust the measure to incorporate the divergence from the expected performance of the random ranker (Section 4). For illustration, we reevaluate the TREC 2012 Web Track's ad hoc task [4] (Section 5).[1]

2 RELATED WORK

The estimation of topic difficulty revolves around grouping topics into coarse-grained classes of difficulty, for example, "easy", "moderate", and "hard" topics. Following Mothe et al. [13], we distinguish three major strategies to assign such difficulty classes, namely size-based, threshold-based, and distribution-based strategies.

For size-based estimations, topics are assigned into difficulty classes such that all resulting classes are of equal size. For example, Eguchi et al. [6] decreasingly order topics by median average precision, and split them into three graded categories, such that category sizes are equal. This approach is also taken by Carterette et al. [3], but using average average precision (AAP) instead.[2] Damessie et al. [5] choose the two highest and lowest scoring topics by AAP to designate "easy" and "hard" topics in their experiment.

For threshold-based estimations, fixed difficulty thresholds discriminate difficulty classes. These thresholds are derived from the specific distribution of scores in an experiment. For a binary classification, Grivolla et al. [8] set their threshold of difficulty at the median AAP over all topics (coincidentally rendering their approach also size-based). Vercoustre et al. [20] set thresholds based on the mean AAP over all topics, and its standard deviation. Depending on where the AAP of a topic falls within the overall distribution, it is classified as one of four graded difficulty classes.

Instead of deriving thresholds, score distributions can be taken into account directly: Shtok et al. [19] and Pérez-Iglesias and Araujo [14] measure the standard deviation of retrieval scores in a run under the hypothesis that the score distributions for "easy" and "hard" topics differ. Similarly, Aslam and Pavlu [1] characterize difficulty as the Jensen-Shannon divergence between runs, with the expectation that the runs' score distributions of systems are more similar for "easy" topics than for "hard" ones.

[1] Code and data underlying this paper: https://github.com/webis-de/CIKM-20
[2] Average average precision (AAP) is introduced to denote a column-wise average of matrix M, distinguishing it from mean average precision's (MAP) row-wise average.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412109

The existing approaches suffer from one or more of the following four shortcomings: (1) *Local consistency.* When assigning topic difficulty classes based on score distributions of a given experiment, a different experiment might result in a different classification. This issue is inherent to offline difficulty estimation, as it depends on the system set, yet can be mitigated using a sensible approach to class assignment. (2) *Topic set stability.* The topic difficulty estimation depends on the topic set employed. Adding or removing topics can alter the difficulty assessment. (3) *Relevance and measure inconsistency.* Using average precision as a base for difficulty estimations restricts them to binary relevance labels—thus, valuable information is not considered for topic difficulty; also, given the importance of NDCG-based evaluations, using different measures to judge systems and topics seems inconsistent. (4) *Discrete class labeling.* Instead of ordinal labels such as "easy" and "hard", a numerical scale to reflect topic difficulty is desirable. Besides using such values in other situations, denoting difficulty on ratio scale allows for inferring all lower-scale levels.

3 DEFINITION AND INVARIANCE PROOFS

In this section, we briefly recap the NDCG measure and then analyze its shift and scale invariance properties as necessary prerequisite to implement our notion of topic difficulty.

Terminology. Let d be a document from the set D of n documents, and let t be a topic from the set T of topics formulated for a retrieval experiment using D. By D_n we denote the set of all possible rankings $\pi : D \to [1, n]$ (bijective functions from D onto the n possible ranks), i.e., D_n is the set of all permutations of the elements of D. An IR system $s \in S$ indexing D and taking part in a retrieval experiment on T can thus be seen as a mapping $s : T \to D_n$.

3.1 Normalized Discounted Cumulated Gain

The normalized discounted cumulated gain (NDCG) by Järvelin and Kekäläinen [9] has become one of the most widely used measures in IR evaluations. In contrast to other measures, it takes into account both the degree of relevance of documents via an information gain function g, and their ranking position via a discount function λ. Given a ranking π from D_n for a topic t and a gain function g, the NDCG score of π is computed as follows:

$$\mathrm{NDCG}(D, t, g, \pi) = \frac{\mathrm{DCG}(D, t, g, \pi)}{\mathrm{IDCG}} = \frac{\mathrm{DCG}(D, t, g, \pi)}{\max\limits_{\tau \in D_n} \mathrm{DCG}(D, t, g, \tau)},$$

where the discounted cumulated gain (DCG) of π given D, t, and g is normalized by the maximal, ideal score (IDCG) attainable for the possible rankings D_n. The DCG is defined as follows:[3]

$$\mathrm{DCG}(D, t, g, \pi) = \sum_{d \in D} \frac{g(d, t)}{\lambda(\pi(d))},$$

where $g : T \times D \to \mathbb{R}$ returns a real-valued information gain score dependent on the relevance of document d to t, and $\lambda : [1, n] \to \mathbb{R}$ a real-valued discount factor dependent on the rank $\pi(d)$ of d.[4]

[3] To ease later steps, we formalize NDCG based on the tuple (D, t, g, π) instead of the original vector of relevance labels [9]; both formalizations are equivalent.

[4] Many choices of discount functions are conceivable [22]; logarithmic discount being the most widespread one: $\lambda(i) := \log_2(i + 1)$ for rank $i \in [1, n]$.

The gain function g is usually obtained from manual, graded relevance judgments (e.g., on a 5-point Likert scale) of the documents in D with regard to topic t. Typically, not all documents in D are judged, but only a pooled subset $D_S = \{d \mid d \in D \text{ and } \exists s \in S : \pi_s(d) \leq k\}$, comprising the union of the k top-ranked documents from the systems S under consideration. Documents ranked below rank k are considered irrelevant by default but NDCG proved to be sufficiently robust against incomplete judgments [17, 23].

3.2 Shift and Scale Invariance of NDCG

The properties of NDCG have been studied in great detail: it has been shown that the measure is bounded, non-monotonic, non-converging, top-weighted, non-localized, incomplete, and realizeable [12]. We formally investigate the shift and scale invariance of NDCG—i.e., whether system rankings change for shifted and/or scaled scores—and prove that scaling and shifting NDCG does not result in different rankings. Scale invariance has been informally considered before [10]; shift invariance not at all. Let $g \odot c$ and $g \oplus c$ denote the element-wise multiplication and addition of the relevance scores of the documents in D with regard to a topic t with a constant $c \in \mathbb{R} \setminus 0$. For example, if the domain of relevance scores of g is $\{1, 2, 3, 4, 5\}$, then for $c = 2$ the domain of $g \odot 2$ is $\{2, 4, 6, 8, 10\}$, and that of $g \oplus 2$ is $\{3, 4, 5, 6, 7\}$. For legibility, the signature of the DCG function is omitted below.

Lemma 1. NDCG is scale invariant.

$$\mathrm{NDCG}(D, t, g \odot c, \pi) = \frac{\sum\limits_{d \in D} \frac{g(d) \cdot c}{\lambda(\pi(d))}}{\max\limits_{\tau \in D_n} \sum\limits_{d \in D} \frac{g(d) \cdot c}{\lambda(\tau(d))}} = \frac{c \cdot \sum\limits_{d \in D} \frac{g(d)}{\lambda(\pi(d))}}{c \cdot \max\limits_{\tau \in D_n} \sum\limits_{d \in D} \frac{g(d)}{\lambda(\tau(d))}} = \frac{c \cdot \mathrm{DCG}}{c \cdot \mathrm{IDCG}}$$

$$= \frac{\mathrm{DCG}}{\mathrm{IDCG}} = \mathrm{NDCG}(D, t, g, \pi) \qquad \square$$

Lemma 2. NDCG is not shift invariant.

$$\mathrm{NDCG}(D, t, g \oplus c, \pi) = \frac{\sum\limits_{d \in D} \frac{g(d) + c}{\lambda(\pi(x))}}{\max\limits_{\tau \in D_n} \sum\limits_{d \in D} \frac{g(d) + c}{\lambda(\pi(d))}} = \frac{\sum\limits_{d \in D} \frac{g(d)}{\lambda(\pi(d))} + \sum\limits_{d \in D} \frac{c}{\lambda(\pi(d))}}{\max\limits_{\tau \in D_n} \left(\sum\limits_{d \in D} \frac{g(d)}{\lambda(\tau(d))} + \sum\limits_{d \in D} \frac{c}{\lambda(\tau(d))} \right)}$$

Note that $c' = \sum\limits_{d \in D} \frac{c}{\lambda(\pi(d))} = \sum\limits_{d \in D} \frac{c}{\lambda(\tau(d))}$. Since $\forall c \in \mathbb{R} \setminus 0 : c' \neq 0$:

$$\mathrm{NDCG}(D, t, g, \pi) = \frac{\mathrm{DCG}}{\mathrm{IDCG}} \neq \frac{\mathrm{DCG} + c'}{\mathrm{IDCG} + c'} = \mathrm{NDCG}(D, t, g \oplus c, \pi) \ \square$$

Corollary 3. NDCG's score differences due to scaling and shifting are linear such that system rankings will not change.

Proof. From Lemma 1 we have $g \odot c_1 \oplus c_2 = g \oplus c_2 \odot c_1 = g \oplus c_2$. With Lemma 2, it follows for $\Delta(\,\mathrm{NDCG}(D, t, g, \pi), \mathrm{NDCG}(D, t, g \oplus c, \pi)\,)$:

$$\Delta(\cdot, \cdot) = \frac{\mathrm{DCG}}{\mathrm{IDCG}} - \frac{\mathrm{DCG} + c'}{\mathrm{IDCG} + c'} = \frac{c'}{\mathrm{IDCG} \cdot (\mathrm{IDCG} + c')} \cdot \mathrm{DCG} - \frac{c'}{\mathrm{IDCG} + c'}$$

Since the IDCG and c' are constant, Δ is linear. $\square$

A linear transformation $g \odot m \oplus n$ of relevance labels thus changes the resulting NDCG scores by a linear factor. NDCG retains all of its other properties, except its boundedness to the $[0, 1]$ range.

4 RATIO-BASED TOPIC DIFFICULTY

To address the issues of the existing topic difficulty estimation strategies identified in Section 2, we propose a new, ratio-based strategy: We measure topic difficulty as the ratio of systems scoring higher than a baseline compared to the overall number of systems. As a consequence, difficulty scores are continuous numerical values in the domain $[0, 1]$, solving Issue 4 (discrete class labeling).

But which system can possibly serve as a meaningful, reliable, and standardized baseline, given the steady progress made on new systems? While it is conceivable that the community decides to pick an implementation of a widespread retrieval model, such as BM25, for a baseline, we introduce another notion which has not been considered so far: the random ranker. We propose the expected NDCG score of a hypothetical random ranking function as reference point, providing for a clear, binary separation between "difficult" and "not difficult": if a system scores worse than the random ranker on a topic, this topic is presumed difficult for that system, and otherwise not. Using this fixation point provides higher consistency across different experimental settings, improving on Issue 1 (local consistency). However, due to the dependence on pooled documents, Issue 1 cannot be fully solved for offline topic difficulty estimation.

We characterize the expected performance of a hypothetical random ranking function $s_{\text{rand.}} : T \times D \rightarrow_{\text{iid.}} D_n$, which picks a ranking at random from D_n, as its expected NDCG score. The gain function g forms a discrete gain distribution $\mathcal{G}(\mu, \sigma)$ over its domain. Therefore, the expected gain value on every rank position approaches μ, which allows for characterizing the expected NDCG score of $s_{\text{rand.}}$, RNDCG, as follows:

$$\text{RNDCG} = \frac{\sum_{r=1}^{n} \frac{\mu}{\lambda(r)}}{\text{IDCG}}.$$

Clearly, computing RNDCG over the entire D_n is nonsensical in practice, given the high class imbalance between relevant and irrelevant documents, where the set of relevant documents is usually dwarfed by the set of irrelevant ones. However, when applying depth-k pooling on all systems in S, and computing the random rankings only on the set of pooled documents D_S, thus obtaining $D_{S,m}$, where $m = |D_S|$, this becomes feasible. Given a sufficient number of systems and pooling depth, the pooling can be assumed to contain a representative sample of the relevant documents in the collection [18]. Thus, the notion of "random ranker" can be redefined as "random reranker". This increases usability, since no extra judgments have to be collected to establish such a baseline, as all necessary information is already contained in the pooling.

As shown in Section 3.2, NDCG retains its properties and the contained information when scaling and shifting the distribution of relevance labels, since this only induces a linear change of the resulting NDCG scores. We can therefore apply a standardization which transforms each value z in $\mathcal{G}$ as follows:

$$z' = \frac{z - \mu}{\sigma},$$

resulting in $\mu' = 0$ and $\sigma' = 1$ for $\mathcal{G}'$. The key benefit of applying standardization is that, since $\mu' = 0$, $\sum_{r=1}^{n} \frac{\mu'}{\lambda(r)} = 0$ follows for all values of k and all possible discount functions λ. Therefore, RNDCG is standardized as zero, assigning a fixed and explicit baseline score. Furthermore, using NDCG with standardized scores, Issue 3 (relevance and measure inconsistency) is solved.

Altogether, we obtain a non-binary topic difficulty score by computing the ratio of positive-scoring system runs to the overall number of systems that contributed to a topic's pool when calculating NDCG scores on standardized score distributions (NDCG_{std}):

$$\text{Difficulty}(t) = \frac{1}{|S|} \sum_{s \in S} \delta(s, t)$$

$$\delta(s, t) = \begin{cases} 1 & \text{if } \text{NDCG}_{std}(D, t, g, \pi_s) > 0 \\ 0 & \text{if } \text{NDCG}_{std}(D, t, g, \pi_s) \leq 0 \end{cases}$$

Difficulty scores are in $[0, 1]$, where 1 denotes the easiest possible kind of topic (all systems score better than random) and 0 denotes the hardest possible kind of topic (all systems score worse than random). This process is only dependent on the gain distribution of the topic in question, solving Issue 2 (topic set stability).

5 VALIDATION

This section aims to illustrate our ratio-based topic difficulty estimation in a practical setting, and in comparison to median or mean aggregation. While our NDCG-based approach of assigning difficulty scores can be motivated purely from a theoretical standpoint, given that it solves most issues prevalent in existing approaches, its practical ramifications are explored in more detail here.

We reevaluated the TREC 2012 ad hoc runs [4] with a pooling depth of $k = 20$, and applied a standardization to the annotated labels, for a total of 26 systems across 40 topics. We then calculated NDCG scores for each system on each topic. We further calculated the difficulty ratio per topic, and inferred difficulty classes by dividing the $[0, 1]$-range into the four equisized intervals $[0, 0.25]$, $(0.25, 0.5]$, $(0.5, 0.75]$, and $(0.75, 1]$, resulting in four discrete classes in ascending order. In Figure 1, we plot the performance of runs as boxplot for each topic, with quartiles, outliers and median shown. A red line is drawn at 0 as visual guide for the baseline. On the right side of the figure, the topic difficulty score is denoted alongside the associated difficulty classes, color-coded from top to bottom as green ("easy") , yellow ("moderately easy"), orange ("moderately hard") and red ("hard"). The general trend is captured by the ratio-based score, as we can see the overall shift of the topics' score distributions being reflected.

Given that our ratio takes into account only whether a score is higher or lower than the random baseline, not the absolute difference in scores, the local precision of the decisions is lower than for median or mean-based approaches. For example, take topics 159 and 174: both have the same difficulty score of 0.81, yet the upper 50% of systems perform better on topic 159 than all of the systems in topic 174. While calculating the mean score instead of our ratio would differentiate between both topics, the global consistency of the ratio method would be lost. Thus, we deem this accuracy tradeoff as favorable. Moreover, the decisions of our ratio seem to be more robust on a local level when dealing with highly

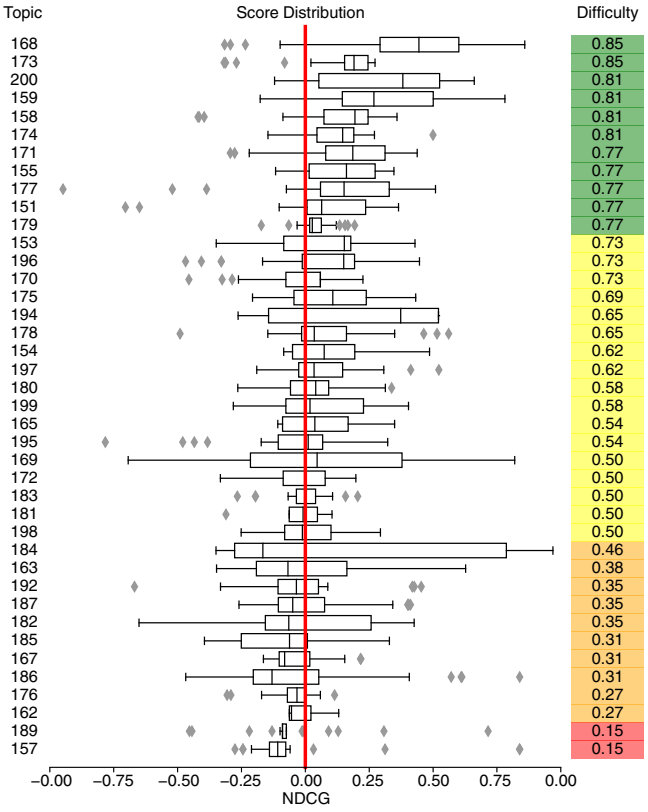

Figure 1: NDCG scores of TREC runs, difficulty ratio and associated class per topic, ordered by difficulty.

skewed distributions. For example, the median value in topic 194 is one of the highest in the topic set, yet the whole 25%-quartile of systems scores worse than 0—indicating that a significant portion of systems have performance difficulties on this topic. When inferring topic difficulty by median, topic 194 would score one of the highest; using our ratio, however, assigns a more reasonable difficulty. This further illustrates the improvement on Issue 1: while inconsistencies between experiments are given due to dependence on a system-specific pooling, the newly introduced measure is more robust in that regard than existing approaches.

The stability of the assessment to outlier runs is demonstrated in topics 177 and 195. Both exhibit extreme outliers towards the lower end of the NDCG distribution, while the main cluster of runs performs considerably better. Yet, when comparing the score distributions of topics with a similar or equal difficulty ratio, both topics are classified correctly. In average-based aggregation approaches, these outlier runs would have had greater impact on the difficulty estimation, skewing the results. If we model changing experiment conditions, for example by sampling only half of the topics and repeat the calculation, the scores stay fixed, since the score of a single topic only depends on the topic itself, not the context of other topic scores. This allows for repeatable assessments within an experiment and comparable assessments across experiments. The redundancy of class labels for different levels of difficulty is further substantiated: We can not observe great changes of performance across inter-class borders.

6 CONCLUSION

We introduce a novel approach of estimating the difficulty of topics in IR evaluations. Our measure is based on a linear transformation of the NDCG measure and assigns continuous difficulty scores derived from the comparison of systems to a random baseline ranking. We demonstrate the usefulness of choosing such a ratio-based method over mean or median aggregation, and substantiated the redundancy of discrete class labels.

In addition to improving on several issues noted for previous measures of topic diffficulty, our approach is useful for several tasks. For example, selecting a specific set of topics, yielding the highest evaluation confidence, akin to Zhu et al. [24]; computing a weighted mean NDCG, placing more emphasis on the systems' performance on hard topics; and provide additional insight in emerging domains of IR, where no baseline systems are established: having a universal reference point at no extra cost is a valuable resource.

REFERENCES

[1] J. A. Aslam and V. Pavlu. 2007. Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions. In *Proc. of ECIR 2007*. Springer, 198–209.
[2] D. Carmel and E. Yom-Tov. 2010. Estimating the query difficulty for information retrieval. In *Proc. of SIGIR 2010*. ACM, 911.
[3] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. 2009. Overview of the trec 2009 million query track. In *Proc. of TREC 2009*.
[4] C. L. A. Clarke, N. Craswell, and E. M.. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *Proc. of TREC 2012*.
[5] T. T. Damessie, F. Scholer, K. Järvelin, and J. S. Culpepper. 2016. The Effect of Document Order and Topic Difficulty on Assessor Agreement. In *Proc. of ICTIR 2016*. ACM, 73–76.
[6] K. Eguchi, K. Kuriyama, and N. Kando. 2002. Sensitivity of IR systems Evaluation to Topic Difficulty. In *Proc. of LREC 2002*. ELRA, 585–589.
[7] N. Ferro, Y. Kim, and M. Sanderson. 2019. Using Collection Shards to Study Retrieval Performance Effect Sizes. *ACM TOIS* 37, 3 (2019), 30:1–30:40.
[8] J. Grivolla, P. Jourlin, and R. de Mori. 2005. Automatic classification of queries by expected retrieval performance. *Actes de SIGIR* 5 (2005).
[9] K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM TOIS* 20, 4 (2002), 422–446.
[10] E. Kanoulas and J. A. Aslam. 2009. Empirical justification of the gain and discount function for nDCG. In *Proceedings of CIKM 2009*. ACM, 611–620.
[11] S. Mizzaro and S. Robertson. 2007. Hits hits TREC: exploring IR evaluation results with network analysis. In *Proc. of SIGIR 2007*. ACM, 479–486.
[12] A. Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *Proc. of AIRS 2013*. Springer, 1–12.
[13] J. Mothe, L. Laporte, and A.-G. Chifu. 2019. Predicting Query Difficulty in IR: Impact of Difficulty Definition. In *Proc. of KSE 2019*. IEEE, 1–6.
[14] J. Pérez-Iglesias and L. Araujo. 2010. Standard Deviation as a Query Hardness Estimator. In *Proc. of SPIRE 2010*. Springer, 207–212.
[15] F. Raiber and O. Kurland. 2014. Query-performance prediction: setting the expectations straight. In *Proc. of SIGIR 2014*. ACM, 13–22.
[16] K. Roitero, E. Maddalena, and S. Mizzaro. 2017. Do Easy Topics Predict Effectiveness Better Than Difficult Topics?. In *Proc. of ECIR 2017*. Springer, 605–611.
[17] T. Sakai. 2007. On the reliability of information retrieval metrics based on graded relevance. *Inf. Process. Manage.* 43, 2 (2007), 531–548.
[18] G. Salton. 1992. The State of Retrieval System Evaluation. *Inf. Process. Manage.* 28, 4 (1992), 441–450.
[19] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Movits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM TOIS* 30, 2 (2012), 11:1–11:35.
[20] A.-M. Vercoustre, J. Pehcevski, and V. Naumovski. 2008. Topic Difficulty Prediction in Entity Ranking. In *Proc. of INEX 2008*. Springer, 280–291.
[21] E. M. Voorhees, D. K. Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT Press, Cambridge.
[22] Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu. 2013. A Theoretical Analysis of NDCG Type Ranking Measures. In *Proc. of COLT 2013*. JMLR.org, 25–54.
[23] E. Yilmaz, E. Kanoulas, and J. A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *Proc. of SIGIR 2008*. ACM, 603–610.
[24] J. Zhu, J. Wang, V. Vinay, and I. J. Cox. 2009. Topic (query) selection for IR evaluation. In *Proc. of SIGIR 2009*. ACM, 802–803.