The Impact of Negative Relevance Judgments on NDCG

Lukas Gienapp
Leipzig University

Maik Fröbe
Martin-Luther-Universität
Halle-Wittenberg

Matthias Hagen
Martin-Luther-Universität
Halle-Wittenberg

Martin Potthast
Leipzig University

ABSTRACT

NDCG is one of the most commonly used measures to quantify system performance in retrieval experiments. Though originally not considered, graded relevance judgments nowadays frequently include negative labels. Negative relevance labels cause NDCG to be unbounded. This is probably why widely used implementations of NDCG map negative relevance labels to zero, thus ensuring the resulting scores to originate from the $[0, 1]$ range. But zeroing negative labels discards valuable relevance information, e.g., by treating spam documents the same as unjudged ones, which are assigned the relevance label of zero by default. We show that, instead of zeroing negative labels, a min-max-normalization of NDCG retains its statistical power while improving its reliability and stability.

ACM Reference Format:
Lukas Gienapp, Maik Fröbe, Matthias Hagen, and Martin Potthast. 2020. The Impact of Negative Relevance Judgments on NDCG. 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/10.1145/3340531.3412123

1 INTRODUCTION

Laboratory evaluations of information retrieval systems predominantly apply the Cranfield paradigm, especially for shared tasks [16]. An evaluation setup requires (1) a set of information needs (topics), (2) relevance judgments for document/topic pairs, and (3) effectiveness measures that calculate a system's success using the relevance judgments. The topic set must be sufficiently large to allow for insights into overall systems performance, and averaging the effectiveness scores must be a meaningful operation. Further, boundedness of scores and them being independent and identically distributed (i.i.d.) are crucial for statistical testing [12].

One of the most common effectiveness measures used is the Normalized Discounted Cumulative Gain (NDCG) [10]. NDCG produces effectiveness scores between zero and one when relevance labels are greater or equal than zero. In current experimental practice—most notably the TREC conferences—documents can be assigned negative relevance labels, causing a potential violation of the boundedness property of NDCG. According to the assessment guidelines of the TREC web tracks, negative labels indicate spam or documents without use for any reasonable purpose [3–7]. Table 1 provides an overview of the judgments of the TREC web tracks between 2010 and 2014.[1] All web track poolings contain a mean ratio of negative judgments per topic between two and six percent. Note that similar or even lower percentages are observable for the higher relevance grades, whereby negative labels become a major part of the evaluation process. Furthermore, the penalization of spam is justified: spam or documents without use are detrimental to the quality of a search result and should be labeled accordingly.

Table 1: Number of runs and mean label ratio per topic for the TREC Web Ad-Hoc Tracks between 2010 and 2014.

TREC	Runs	Mean Label Ratio per Topic					
		-2	0	1	2	3	4
2010	56	0.05	0.73	0.17	0.05	0.01	0.00
2011	37	0.06	0.77	0.12	0.05	0.10	0.00
2012	32	0.05	0.73	0.14	0.04	0.02	0.06
2013	34	0.02	0.70	0.21	0.08	0.02	0.01
2014	30	0.06	0.57	0.26	0.11	0.03	0.02

Existing implementations of NDCG disregard negative relevance labels to ensure that the resulting scores are between zero and one. Thus, they comply with the critical requirement that scores are bounded and comparable, yet ignore the effect of spam and junk pages within web search. Implementations like trec_eval,[2] trectools [13], and gdeval[3] map negative labels to zero, while implementations of widely used LTR pipelines—like RankLib [8], LETOR 4.0 [14], and LTR-evaluate[4]—consider negative labels as invalid. Consequently, evaluations performed on the web tracks implicitly neglect information contained in up to six percent of annotated documents by assigning the default zero relevance score.

We study the impact of negative relevance labels on NDCG scores by reevaluating the TREC web tracks. We further propose a modified version of NDCG that incorporates negative labels while ensuring soundness of scores. Our experiments indicate that including negative labels into the original NDCG can substantially impact systems rankings, while our modified version produces more reliable, stable, and sensitive results than the current practice.[5]

2 FOUNDATIONS AND RELATED WORK

Terminology. Let $D = \{d_1, \ldots, d_n\}$ be a set of n documents and t a topic from a set of topics T with regard to D. A ranking $\pi : D \longrightarrow [1, n]$ is as a bijective function from D onto the n possible ranks; let D_n denote the set of all possible rankings (i.e., all permutations of the elements of D). An IR system s that indexes D can thus be defined as a mapping $s : T \rightarrow D_n$. Let S denote the set of systems.

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

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412123

[1] TREC 2009 was omitted as it uses a different annotation scheme
[2] https://github.com/usnistgov/trec_eval
[3] https://trec.nist.gov/data/web/12/gdeval.pl
[4] http://learningtorankchallenge.yahoo.com/evaluate.py.txt
[5] Code and data underlying this paper: https://github.com/webis-de/CIKM-20

2.1 Normalized Discounted Cumulative Gain

The normalized discounted cumulative gain (NDCG) by Järvelin and Kekäläinen [10] 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 set of documents D, a topic t, a ranking π from D_n, 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 cumulative 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:[6]

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

where $g : T \times D \rightarrow \mathbb{R}$ returns a real-valued information gain score dependent on the relevance of d to t, and $\lambda : [1,n] \rightarrow \mathbb{R}$ a real-valued discount factor dependent on the rank $\pi(d)$ of d.[7]

2.2 Properties of Effectiveness Measures

Numeric Properties. Moffat [12] identifies seven numeric properties of effectiveness measures, and shows that NDCG is (1) bounded (scores reside in a defined interval), (2) convergent (if any relevance labels increase, scores strictly increase), (3) top-weighted (if a document in the rankings' top-k is switched with another of higher relevance outside the top-k, resulting scores strictly increase) (4) realizeable (if at least one relevant document exists, scores can be maximal). However, it is not (5) monotonous (if k is increased, scores never decrease), (6) localized (scores only depend on the information in the k documents of the ranking), or (7) complete (scores can be calculated even if there are no relevant documents). Further, Gienapp et al. [9] formally prove that NDCG is scale invariant, but not shift invariant, and that NDCG scores change linearly if a linear transformation is applied to the pooled relevance labels.

Reliability. The reliability of a measure denotes its ability to reflect the actual performance differences of systems in its score, minimizing deviation from the "true" performance rating. Generalizability Theory (GT) has been proposed to evaluate reliability in IR experiments [1], and been used to investigate the optimal gain and discount functions for NDCG [11]. Assuming that overall variance in performance scores of systems can be decomposed into a system variance σ_s^2, a topic variance σ_t^2, and a system-topic-interaction variance $\sigma_{s:t}^2$ [11], the goal is to draw conclusions about the proportion of variance in evaluation results that stems from actual performance differences. Variance components can be estimated by fitting an ANOVA model on the NDCG scores [1, 11]. The reliability coefficient Φ, i.e., the ratio of system to overall variance is calculated as follows [11, Eq. 2]:

$$\Phi = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_t^2 + \sigma_{s:t}^2}{|T|}}.$$

Stability. The stability of a measure denotes the dependency of scores on the number of topics it is calculated on. Buckley and Voorhees [2] quantify stability as the error rate associated with the decision about which of two tested systems is better, varying the topic set size $m \in [1, |T|]$. This rate is defined as ratio of erroneous decisions to total decisions across all system pairs [2, Eq. 1]:

$$\mathrm{ErrorRate} \;=\; \frac{\sum\limits_{s_a, s_b \in S} \min(|s_a > s_b|, |s_a < s_b|)}{\sum\limits_{s_a, s_b \in S} (|s_a > s_b| + |s_a < s_b| + |s_a = s_b|)},$$

where $|s_a > s_b|$ denotes the number of topics a system s_a scored better on than a system s_b, with $|s_a < s_b|$ and $|s_a = s_b|$ defined analogously. Two systems are deemed equal when their absolute score difference does not exceed a threshold given by a fuzzyness factor f. Buckley and Voorhees [2] conduct the stability analysis by randomly sampling topics 100 times for each m with $f = 5\%$.

Sensitivity. The sensitivity, or discriminative power of an evaluation measure denotes its ability to successfully tell two systems apart, given that their true performance differences are significant. To test the sensitivity of evaluation measures, Sakai [15] proposed the bootstrap method, where the sensitivity of a measure is quantified by testing all pairs of systems under consideration using a bootstrap hypothesis test and counting the number of system pairs achieving a significance level lower than a given threshold.

3 NDCG FOR NEGATIVE RELEVANCE LABELS

The boundedness of NDCG depends on a narrow definition of allowed relevance labels. Järvelin and Kekäläinen [10] only considered positive values when they initially described the measure, and Moffat [12] restricts his considerations to positive values in the range between 0 and 1. The use of negative relevance labels results in a violation of the boundedness property—neither the upper nor the lower bound can be guaranteed anymore.

Two violations are possible if both numerator (DCG) and denominator (IDCG) of the NDCG formula are allowed to extend into the negative domain: (1) The lower bound can fall below 0, if the DCG value is negative, i.e., when enough documents with negative gain values appear at high-ranking positions. (2) The upper bound can exceed 1, if both the DCG and the IDCG are negative. In such a case, the absolute value of the IDCG is smaller than the absolute value of the DCG, thus allowing NDCG to become larger than 1.

As outlined in the Introduction, boundedness is essential to compute score means. Two approaches to restore boundedness can be distinguished: either, the negative scores are eliminated, or the metric itself is adapted to be well-defined for negative relevance labels. To distinguish the unmodified formulation (i.e., calculating NDCG directly on all scores), we refer to it as $\mathrm{NDCG}_{\mathrm{org}}$.

We further consider one variant of $\mathrm{NDCG}_{\mathrm{org}}$ for each of the possible approaches: (1) NDCG_0, where negative scores are eliminated by mapping them to zero. Since the $\mathrm{NDCG}_{\mathrm{org}}$ measure itself is unaffected by this score transformation, all properties of $\mathrm{NDCG}_{\mathrm{org}}$, as well as boundedness remain valid. While mapping is a commonly employed strategy, it is problematic, as it can significantly impact

[6] To ease later steps, we formalize NDCG based on the tuple (D, t, g, π) instead of the original vector of relevance labels [10]; both formalizations are equivalent.

[7] Many choices of discount functions are conceivable [17]; but logarithmic discount is the most widespread one: $\lambda(i) := \log_2(i+1)$ for rank $i \in [1, n]$. We use logarithmic discount throughout this paper.

evaluation results; (2) we define a new, more constrained notion of $\mathrm{NDCG}_{\mathrm{org}}$ to explicitly normalize the range of values the measure can attain to the $[0, 1]$ interval, where 1 represents the perfect ranking, but additionally, 0 is now defined as the worst possible ranking of documents for the topic in question. To achieve this standardized range, we extend $\mathrm{NDCG}_{\mathrm{org}}$ to adopt full min-max-normalization. While $\mathrm{NDCG}_{\mathrm{org}}$ traditionally normalizes using only the maximum value (IDCG), our constrained version ($\mathrm{NDCG}_{\mathrm{min}}$) also incorporates the worst possible ranking of the pooling, formalized as follows, with function signatures shortened for brevity:

$$\mathrm{NDCG}_{\mathrm{min}}(\pi) = \frac{\mathrm{DCG}(\pi) - \min\limits_{\tau \in D_n} \mathrm{DCG}(\tau)}{\max\limits_{\tau \in D_n} \mathrm{DCG}(\tau) - \min\limits_{\tau \in D_n} \mathrm{DCG}(\tau)}$$

Since $\min \mathrm{DCG}$ is constant within a topic, substracting it from both the numerator and the denominator of the $\mathrm{NDCG}_{\mathrm{org}}$ formula is a linear transformation. Thus, the intra-topic relationship between $\mathrm{NDCG}_{\mathrm{org}}$ and $\mathrm{NDCG}_{\mathrm{min}}$ is linear and all the properties noted for $\mathrm{NDCG}_{\mathrm{org}}$ by Moffat [12] are also applicable to $\mathrm{NDCG}_{\mathrm{min}}$, with boundedness now explicitly ensured for all relevance labels.

4 COMPARISON OF EVALUATION RESULTS

Experimental Setup. To study the impact of negative gain values as well as the different strategies to deal with them on IR evaluations, we reevaluate the runs submitted to the TREC Web Tracks 2010 to 2014 [3–7]. For each track, we calculate the $\mathrm{NDCG}_{\mathrm{org}}$, NDCG_{0}, and $\mathrm{NDCG}_{\mathrm{min}}$ scores for each topic and run using the original pooling depths ($k = 20$ for 2010/11/12; $k = 15$ for 2013/14).

Boundedness. We first investigate the lower bound of attainable $\mathrm{NDCG}_{\mathrm{org}}$ scores per topic by calculating the $\mathrm{NDCG}_{\mathrm{org}}$ score of the worst possible ranking (i.e., the inverse IDCG). The widespread violation of the boundedness property is apparent in Table 2a, with a large portion of topics in each year falling below zero regarding attainable $\mathrm{NDCG}_{\mathrm{org}}$ scores. This is not surprising, as only one negative relevance label suffices for this violation. Yet, a high amount of topics effectively double the range of possible $\mathrm{NDCG}_{\mathrm{org}}$ scores, reaching beyond -1. While none of the submitted runs achieve a negative score, the different attainable score ranges across topics increase the score variance, impacting the reliability of experiments.

Change in System Rankings. To compare the different system rankings as given by the mean system scores across topics with each measure, we calculate the correlation coefficient Spearman's ρ between the ranking given by mean $\mathrm{NDCG}_{\mathrm{org}}$ scores to the rankings given by mean NDCG_{0} and mean $\mathrm{NDCG}_{\mathrm{min}}$ (Table 2b). While the resulting rankings are virtually the same between all measures in 2013 and 2014, the system rankings as produced by NDCG_{0}-based evaluation substantially differ from the $\mathrm{NDCG}_{\mathrm{org}}$ rankings in the other years, with correlation falling as low as 0.79 in 2011. Nevertheless, the rankings given by $\mathrm{NDCG}_{\mathrm{min}}$ almost perfectly reproduce the system rankings obtained with traditional $\mathrm{NDCG}_{\mathrm{org}}$ in all instances. To closely examine the differences between $\mathrm{NDCG}_{\mathrm{org}}$ and NDCG_{0} rankings, we plot the attained ranks for each system in 2010/11/12 in Table 2c. Most of the divergence between rankings occurs at lower ranks, however, some deviations in top ranks are observable. This is problematic: the best-performing system may depend on whether negative gain values are considered or not.

5 COMPARISON OF MEASURE PROPERTIES

Experimental Setup. In this section, we compare the reliability, stability, and sensitivity of $\mathrm{NDCG}_{\mathrm{org}}$, NDCG_{0}, and $\mathrm{NDCG}_{\mathrm{min}}$. We reevaluate the runs submitted to the Web Tracks 2011 and 2012 [4, 5]. These two were selected since the 2011 track has shown a clear divergence between $\mathrm{NDCG}_{\mathrm{org}}$ and NDCG_{0} scores ($\rho = 0.79$), indicating that negative relevance labels are especially problematic here, while the results are much closer for 2012 ($\rho = 0.91$), providing complementary insight for non-problematic settings. Also, both years are judged up to a pooling depth $k = 20$ and they had a similar number of submitted runs. For both tracks, we analyze the scores of $\mathrm{NDCG}_{\mathrm{org}}$, NDCG_{0}, and $\mathrm{NDCG}_{\mathrm{min}}$ at different pooling depths to provide an intuition about their benefits and shortcomings. We calculate scores for each topic and run on poolings of depth k of 5, 10, 15, and 20, respectively.

Reliability. To compare the reliability of the measures, we calculate Φ for different pooling depths per year. Results are given in Table 2d. Throughout all years and pooling depths, $\mathrm{NDCG}_{\mathrm{min}}$ achieves the highest reliability score, indicating that existing NDCG implementations can improve their reliability by switching from NDCG_{0} to $\mathrm{NDCG}_{\mathrm{min}}$. The reliability of $\mathrm{NDCG}_{\mathrm{org}}$ is much lower than that for $\mathrm{NDCG}_{\mathrm{min}}$ and NDCG_{0}. This supports two key conclusions: (1) the usage of negative scores (in violation of boundedness) is indeed problematic in $\mathrm{NDCG}_{\mathrm{org}}$-based evaluations, as the reliability of the metric is decreased; (2) adopting $\mathrm{NDCG}_{\mathrm{min}}$ instead of NDCG_{0} yields benefits, as it achieves the highest reliability among the tested measures. The increased reliability could be attributed to decreasing the impact of score variance across topics, as they are normalized to the same standardized range. The variance due to topic-system-interaction (i.e., different systems have difficulties with different topics) is not influenced by this normalization.

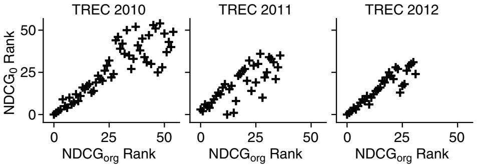
Stability. To compare the stability of the measures, we calculate the error rates as proposed by Buckley and Voorhees [2]. However, we increase the number of samples to $n = 200$ to accommodate for the higher number of compared systems and topics. Results are given in Figure 2e, plotting the error rate by increasing number of topics, for each year and value of k, respectively. Two key insights become apparent: (1) the error rate is inversely related to the number of topics in all three measures; (2) while the three measures are similar at lower values of k, $\mathrm{NDCG}_{\mathrm{min}}$ consistently achieves a lower error rate across all numbers of topics at $k = 20$. Also, $\mathrm{NDCG}_{\mathrm{org}}$ and NDCG_{0} seem to suffer from an increased error rate for higher values of k, an effect that is not observable for $\mathrm{NDCG}_{\mathrm{min}}$.

Sensitivity. To compare the sensitivity of the measures, we use the paired bootstrap method of Sakai [15]. For each pair of runs within one year, the achieved significance level was calculated at different pooling depths ($n = 1000$). Results are given in Figure 2f, where we plot the cumulative ratio of systems achieving a certain significance level with testing based on $\mathrm{NDCG}_{\mathrm{org}}$, NDCG_{0}, and $\mathrm{NDCG}_{\mathrm{min}}$. The sensitivity curves of $\mathrm{NDCG}_{\mathrm{org}}$ and $\mathrm{NDCG}_{\mathrm{min}}$ are nearly indistinguishable, while NDCG_{0} is consistently less sensitive. This provides evidence that the information lost by ignoring negative labels is vital to achieve evaluation results with high discriminative power—a finding that implementations of NDCG can directly employ by switching from NDCG_{0} to $\mathrm{NDCG}_{\mathrm{min}}$.

Table 2: (a) Proportion of topics with an $\mathrm{NDCG_{org}}$ bound lower than 0/-1. (b) Spearman's ρ correlation coefficient of mean system scores for all measure combinations. (c) $\mathrm{NDCG_{org}}$ system rank by $\mathrm{NDCG_0}$ system rank for 2010, 2011, 2012. (d) Φ-coefficients of $\mathrm{NDCG_{org}}$, $\mathrm{NDCG_0}$, and $\mathrm{NDCG_{min}}$ at different pooling depths. Highest per column segment marked. (e) Error rate by number of topics per year and pooling depth k. (f) Cumulative ratio of system pairs at each significance level per year and pooling depth k. Key: $\mathrm{NDCG_{min}}$ ——; $\mathrm{NDCG_{org}}$ - - -, $\mathrm{NDCG_0}$ ······.

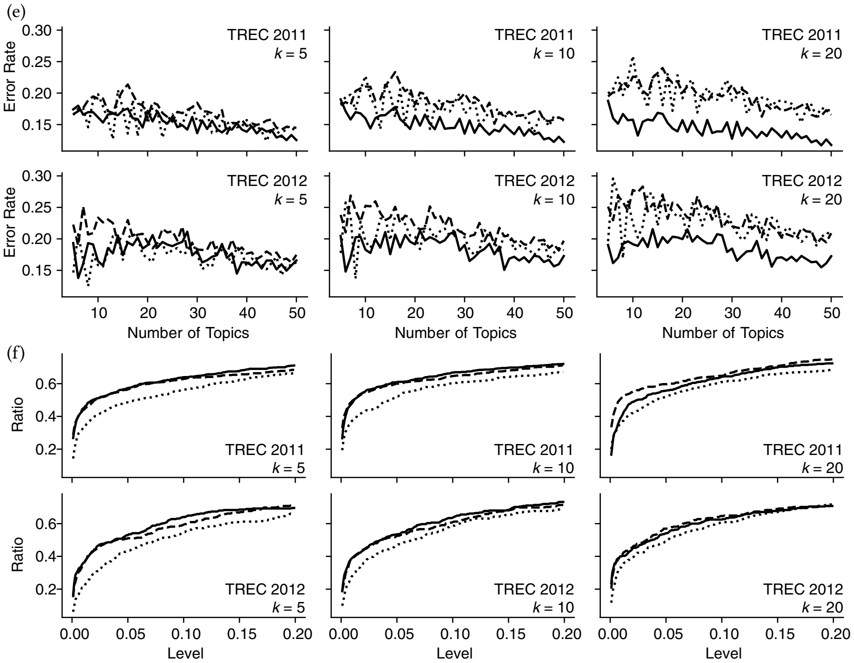
	(a)		(b)		
TREC	Topics		Spearman's ρ Correlation		
	$\mathrm{NDCG_{org}}$		$\mathrm{NDCG_0}$	$\mathrm{NDCG_{min}}$	$\mathrm{NDCG_{min}}$
	< 0	≤ -1	$\mathrm{NDCG_{org}}$	$\mathrm{NDCG_{org}}$	$\mathrm{NDCG_0}$
2010	100%	44%	0.87	1.00	0.88
2011	94%	68%	0.79	0.99	0.82
2012	96%	08%	0.91	0.99	0.93
2013	74%	18%	0.93	0.96	0.97
2014	70%	12%	0.95	0.97	0.99

(c)



(d)

TREC	Measure	$k = 5$	$k = 10$	$k = 15$	$k = 20$
2011	$\mathrm{NDCG_{org}}$	0.903	0.937	0.950	0.924
	$\mathrm{NDCG_0}$	0.977	0.973	0.978	0.975
	$\mathrm{NDCG_{min}}$	**0.996**	**0.993**	**0.988**	**0.984**
2012	$\mathrm{NDCG_{org}}$	0.969	0.930	0.942	0.967
	$\mathrm{NDCG_0}$	0.958	0.975	0.959	0.940
	$\mathrm{NDCG_{min}}$	**0.994**	**0.995**	**0.996**	**0.996**



6 CONCLUSION

When calculating NDCG on a pooling that contains negative relevance labels, boundedness is violated. As unbounded metrics should not be used for mean computation, this is not only a theoretical issue, but also a widespread problem in practice: it opposes one of the central assumptions that system performance can be approximated by mean performance over a set of topics. All NDCG implementations provided by commonly used evaluation tools circumvent this problem by ignoring negative scores altogether. However, since crucial information about system performance is ignored, the ranking of systems can be significantly affected, even at top ranks.

As an alternative, we propose a more constrained version of $\mathrm{NDCG_{org}}$ by adopting full min-max-normalization to render NDCG well-behaved for arbitrary choices of relevance grades. Besides reestablishing theoretical consistency, this improves on several properties of the $\mathrm{NDCG_{org}}$ measure. Our experiments suggest $\mathrm{NDCG_{min}}$ as viable solution, as it exhibits higher reliability than the common practice of ignoring negative labels, while reproducing the system ranking as implied by $\mathrm{NDCG_{org}}$ with similar sensitivity to NDCG, and achieving a higher discriminative power than $\mathrm{NDCG_0}$. As it also exhibits increased stability, evaluation experiments adopting this constrained version could potentially achieve conclusive insights using less topics, thus reducing the cost overhead.

REFERENCES

[1] D. Bodoff and Pu Li. 2007. Test Theory for Assessing IR Test Collections. In *Proc. of SIGIR*. ACM, 367–374.

[2] C. Buckley and E. M. Voorhees. 2017. Evaluating Evaluation Measure Stability. *SIGIR Forum* 51, 2 (2017), 235–242.

[3] C.L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. 2010. Overview of the TREC 2010 Web Track. In *Proc. of TREC*.

[4] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. 2011. Overview of the TREC 2011 Web Track. In *Proc. of TREC*.

[5] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *Proc. of TREC*.

[6] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. 2013. Overview of the TREC 2013 Web Track. In *Proc. of TREC*.

[7] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. 2014. TREC 2014 Web Track Overview. In *Proc. of TREC*.

[8] V. Dang. 2013. The Lemur Project-Wiki-RankLib. *Lemur Project* (2013). Available: https://sourceforge.net/p/lemur/wiki/RankLib.

[9] L. Gienapp and B. Stein and M. Hagen and M. Potthast. 2020. Estimating Topic Difficulty Using Normalized Discounted Cumulated Gain. In *Proc. of CIKM*. ACM.

[10] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM TOIS* 20, 4 (2002), 422–446.

[11] E. Kanoulas and J. A. Aslam. 2009. Empirical Justification of the Gain and Discount Function for nDCG. In *Proc. of CIKM*. ACM, 611–620.

[12] A. Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *Information Retrieval Technology*, Springer, Berlin, Heidelberg, 1–12.

[13] J. Palotti, H. Scells, and G. Zuccon. 2019. TrecTools: An Open-Source Python Library for Information Retrieval Practitioners Involved in TREC-like Campaigns In *Proc. of SIGIR*. ACM.

[14] T. Qin, T. Liu. 2013. Introducing LETOR 4.0 Datasets. *CoRR* abs/1306.2597 (2013).

[15] T. Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proc. of SIGIR*. ACM, 525–532.

[16] E. M Voorhees, D. K. Harman, et al. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Vol. 63. MIT press Cambridge.

[17] 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*. 25–54.