Evaluating Multi-Dimensional Cumulated Utility in Information Retrieval

Francesco Luigi De Faveri*
University of Padova
Padova, Italy
francescoluigi.defaveri@phd.unipd.it

Guglielmo Faggioli
University of Padova
Padova, Italy
guglielmo.faggioli@unipd.it

Nicola Ferro
University of Padova
Padova, Italy
ferro@dei.unipd.it

Kalervo Järvelin
Tampere University
Tampere, Finland
kalervo.jarvelin@tuni.fi

ABSTRACT

Traditional Information Retrieval (IR) effectiveness metrics assume that a relevant document satisfies the information need as a whole. Nevertheless, if the information need is faceted or contains subtopics, this notion of relevance cannot model documents relevant only to one or a few subtopics. Furthermore, faceted documents in a ranked list may focus on the same subtopics, and their content may overlap while neglecting other subtopics. Hence, a search result, where top-ranked documents deal with different subtopics should be preferred over a result where documents are thematically limited and provide overlapping information. The Multi-Dimensional Cumulated Utility (MDCU) metric, recently formulated theoretically by Järvelin and Sormunen, extends the evaluation of novelty and diversity by considering content overlapping among documents. While Järvelin and Sormunen described the theory of MDCU and illustrated its application on a toy example, they did not investigate its empirical use. In this paper, we show the practical feasibility and validity of the MDCU by applying it to publicly available TREC test collections. Furthermore, we analyse its relation with the well-established α-nDCG, and finally, we provide a Python implementation of the MDCU, fostering its adoption as an evaluation framework. Our results indicate a positive correlation between α-nDCG and MDCU, suggesting that both measures correctly identify similar trends when evaluating the IR systems. Finally, compared to α-nDCG, MDCU exhibits a stronger statistical power and identifies up to 9 times more statistically significantly different pairs of systems.

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**; **Novelty in information retrieval**; **Information retrieval diversity**; **Retrieval effectiveness**; **Presentation of retrieval results**.

*Corresponding Author.

ACM SIGIR '25, July 13–18, 2025, Padova, Italy

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

Evaluation Metrics, Multidimensional Cumulated Utility

ACM Reference Format:
Francesco Luigi De Faveri, Guglielmo Faggioli, Nicola Ferro, and Kalervo Järvelin. 2025. Evaluating Multi-Dimensional Cumulated Utility in Information Retrieval. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Traditional Information Retrieval (IR) evaluation metrics primarily rely on mono-dimensional relevance judgments to assess the quality of retrieved documents. Mono-dimensional relevance judgements assume a query to be mono-thematic: the retrieved documents either concern the (single) theme relevant to the query — possibly with a degree of quality— or they do not. Moreover, most of the IR measures focus exclusively on the topical relevance, ignoring other aspects that might play a central role in determining how accessible is the information in the document. An example of a family of measures based on these principles is the Cumulated Gain IR evaluation metrics family introduced by Järvelin and Kekäläinen [9].

This abstraction is instrumental in facilitating offline evaluation. However, users' information needs often extend beyond singular relevance assessments, considering factors such as novelty, redundancy, and thematic diversity [2, 13, 16, 19, 22]. To address this fact, different evaluation measures [6, 10] have been proposed to assess IR systems by blending multi-dimensional relevance judgments. These measures reward the novelty of retrieved documents' content, penalizing redundant information and overlap between themes. A recent advance in the multi-theme and multi-attribute IR evaluation domain is the Multi-Dimensional Cumulated Utility (MDCU), proposed by Järvelin and Sormunen [10]. MDCU operates under four assumptions: i) An information need might be multi-thematic, and documents might satisfy one, many, or none of such themes; ii) The relevance of a document to each sub-theme of the information need can be multi-graded; iii) When crossing the ranked list of documents, the user experience decreasing gain proportionally to the information accrued up to that point: i.e., a partially relevant document inspected after a highly relevant one contributes less to the user's total gain; iv) The contribution of the document to

the user's utility gain depends on its attributes, e.g., the language, complexity, recency.

Although Järvelin and Sormunen [10] formalise theoretically the MDCU, they did not test it empirically on real collections. Therefore, the contributions of this paper are: i) we adapt the theoretical MDCU formulation to use it as a practical IR measure; ii) we gauge empirically its properties using the well-known TREC Web diversity track collections; iii) we compare it with α-nDCG [6], another multi-theme measure; iv) we release publicly the MDCU code, fostering its adoption and reproducibility[1]. Empirically, we observe a reasonable but not pathological correlation between MDCU and α-nDCG. This shows the MDCU stability, offering another perspective on retrieval systems and enabling more powerful statistical testing than α-nDCG.

Section 2 introduces the state-of-the-art measures focusing on α-nDCG and MDCU. Section 3 explains the challenges of measuring the MDCU in practice and details the proposed solutions. Section 4 shows the experimental analysis of MDCU statistical stability and power, and Section 5 outlines the conclusion and future work.

2 BACKGROUND

Before introducing α-nDCG and MDCG, we define the used notation. Assume an IR system has produced a ranked list of documents $\mathcal{D} = \{d_1, ..., d_k\}$ for a given information need. We consider multi-theme information needs, meaning each information need can be split into themes $t \in \mathcal{T}$. Therefore, the relevance judgement for the document d_i is a vector $(r_{i,1}, ..., r_{i,|\mathcal{T}|})$ were element $r_{i,t}$ describes the relevance of d_i to theme t. This relevance can be binary or graded.

α-nDCG. The α-Discounted Cumulative Gain (α-DCG) [6] is an extension of the standard Discounted Cumulative Gain (DCG) [9] designed to evaluate IR systems while accounting for diversity in retrieved documents. Compared to DCG, the gain is modified to discount redundant information in the ranked list of documents. Such discount is controlled by a parameter $\alpha \in [0, 1]$. The utility gain $G(i)$ due to the i-th document of the ranked list is defined as $G(i) = \sum_{t \in \mathcal{T}} r_{i,t}(1-\alpha)^{\sum_{j=1}^{i-1} r_{j,t}}$. The exponent $\sum_{j=1}^{i-1} r_{j,t}$ accounts for the information on sub-theme t that was accrued by looking at the first i-1 documents to discount the relevance of the document. The α-DCG at a given rank k is defined as $\alpha\text{-}DCG@k = \sum_{i=1}^{k} \frac{G(i)}{\log_2(i+1)}$.

The normalized version α-DCG (α-nDCG) is computed by dividing the α-DCG by the ideal α-DCG, ensuring that the values remain in the $[0, 1]$ range, as done in the traditional nDCG [9]. The ideal ranking is constructed by sorting the documents according to the average of their relevance judgements vector. This heuristic can lead to sub-optimal ideal runs, hence lower ideal DCG, but, according to Clarke et al. [6], it provides a good approximation.

The MDCU Framework. The Multi-Dimensional Cumulated Utility (MDCU) framework introduced in [10] assesses cumulative gain by considering multi-dimensional relevance judgments and usability attributes of the documents retrieved by the system in an IR search task. Therefore, document d_i is associated with a vector of usability attributes $(a_{i,1}, \ldots, a_{i,m})$, $(a_{i,j})_{1 \le j \le m} \in [0, 1]$. Algorithm 1 reports the pseudo-code for computing the MDCU. The algorithm takes as input the ranked list of search results, denoted as $\mathcal{D}$, and a discounting parameter b, accounting for the information overlap.

Algorithm 1: MDCU Evaluation Framework [10]

Input: Ranked Search Results List $\mathcal{D}$, overlapping base b
Output: *MDCU*

```
c = zeros(|T|);
for d_i ∈ D with d = [(r_{i,1}, ..., r_{i,|T|}), (a_{i,1}, ..., a_{i,m})] do
    Define a = ∏_{j=1}^{m} a_{i,j};
    for t ∈ T do
        c_t = c_t + a · r_{i,t} / max(1, max(0, log_b(c_t)));
MDCU = Σ_{t∈T} c_t;
return MDCU
```

The cumulated relevance vector c is initialized as the 0-vector of length $|\mathcal{T}|$. For each document $d_i \in \mathcal{D}$, MDCU defines the attribute factor $a = \prod_{j=1}^{m} a_{i,j}$ that describes its usability. After each document's inspection, the cumulated relevance vector c is updated considering the multi-dimensional relevance of the document d_i. The contribution of the i-th document combines its relevance to the various sub-themes and weighs it by the usability attribute. Furthermore, the contribution on theme t is discounted by the cumulated contribution c_t accrued until that point. The MDCU is the sum of all cumulated contributions c_t across the relevance themes $\mathcal{T}$.

3 MDCU: ISSUES & POSSIBLE SOLUTIONS

3.1 Collections for Computing MDCU

Identifying a suitable test collection to validate empirically MDCU represents a preliminary experimental challenge. As noted by Järvelin and Sormunen [10], no collection contains document annotations for multi-theme relevance and usability. Consequently, we focus on the multi-theme evaluation, leaving the investigation of the role of the usability attributes as future work. In detail, we consider the TREC Web collections spanning 2009 to 2012 and use MDCU to evaluate systems submitted to the Web Diversity cat-B task [1, 3–5]. Within the Web Diversity challenge, relevance judgments for the documents are provided considering distinct subtopics relevant to the queries, thus representing themes' relevance. The objective of the diversity task was to produce a ranked list of pages that collectively offer comprehensive coverage for a query, minimizing excessive redundancy in the resultant list.

3.2 Normalizing The MDCU Score

As noted by Clarke et al. [6], determining the optimal run to normalize $\alpha - nDCG$ in the multi-theme scenario is an NP-hard problem. The same challenge holds for MDCU. Indeed, finding such run would require evaluating all possible ranking permutations to find the one that maximizes the final MDCU score. Specifically, for a list of k documents, this entails considering the $n!(k-1)!$ rankings. To make the problem computationally tractable and avoid heuristics that might lead to inconsistent values, we propose two strategies to

[1] https://anonymous.4open.science/r/SIGIR-SP-MDCUEval-BE98/README.md

project MDCU scores in standard intervals: Z-Score standardization and MinMax normalization. These approaches have been demonstrated to map the values of a stochastic variable in equivalent intervals in [12].

We follow Webber et al. [21] to apply Z-score standardization. In detail, given $\mathcal{S}$ the set of systems to evaluate, a query q, and called $MDCU_{s,q}$ the MDCU score for the system $s \in \mathcal{S}$ on query q, to standardize the MDCU values we compute across the systems under evaluation the observed mean MDCU $\mu_q = \sum_{s \in \mathcal{S}} MDCU_{s,q}/|\mathcal{S}|$ and standard deviation $\sigma_q = \text{stdev}(\{MDCU_{s,q}, \forall s \in \mathcal{S}\})$. The standardized MDCU score of system s on query q is computed as:

$$MDCU_{ZScore}(s,q) = (MDCU_{s,q} - \mu_q)/\sigma_q. \quad (1)$$

To map the values of the MDCU in the interval $[0,1]$, we employ the MinMax normalization. Thus, for a query q, we first compute the minimum and maximum MDCU scores observed for that query across the retrieval systems as $min_q = \min_{s \in \mathcal{S}} MDCU_{s,q}$ and $max_q = \max_{s \in \mathcal{S}} MDCU_{s,q}$. The MinMax normalized MDCU for a system s on query q is computed as:

$$MDCU_{MinMax}(s,q) = \frac{MDCU_q - \min_q}{\max_q - \min_q} \quad (2)$$

4 EXPERIMENTAL EVALUATION

4.1 Methodology

We compare α-nDCG and MDCU when evaluating the systems submitted to the TREC Web Diversity cat-B track. To compute α-nDCG, we use the `pyndeval` package of the `ir_measure` Python library[2]. The α value has been maintained as the default specified in the package, i.e., $\alpha = 0.5$. To ensure transparency and reproducibility, we provide the MDCU code and results on the collections in the repository.

4.1.1 Assessing the MDCU Stability. We present the analysis concerning the stability of MDCU compared to α-nDCG on the TREC Web Diversity cat-B[3]. For the comparison cut-off points k, we select $k = 5$ and $k = 20$, which corresponds to the measurement standard adopted in the original challenges of the Web Diversity track [1, 3–5]. We analyze the correlation and agreement between different runs by computing Pearson's ρ and Kendall's τ correlation coefficients [11, 15] for pairs of evaluation measures. The ideal outcome of the stability analysis is to ensure that α-nDCG and MDCU positively correlate —indicating that the two measures are related— but they capture distinct aspects of the multi-dimensional evaluation, testified by the absence of pathologically high correlation.

4.1.2 Measuring the MDCU Statistical Power. In addition to the stability analysis of the MDCU framework, we conduct the ANalysis Of the VAriance (ANOVA) [17] and Siegel-Tukey's test [18] using the Pingouin package [20] to assess the concordance between α-nDCG and MDCU, as proposed in [7, 8, 14]. In detail, the concordance measures proposed by [7, 8, 14] consider pair-wise comparisons of systems carried out in different experimental settings—in our case, using different measures. Such measures consider two aspects of a system-system pair-wise comparison: "statistical significance" and

[2]https://github.com/terrierteam/ir_measures

[3]Here, we show the results for Pearson's and Kendall's correlation on the runs of the Web Diversity'12 [5]. The results on the other collections are reported in the repository.

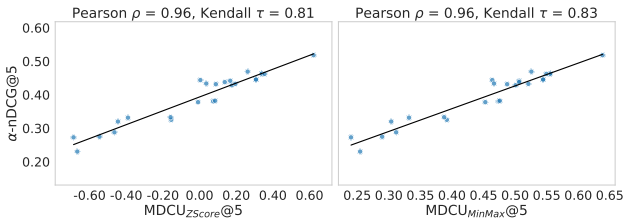

Figure 1: Comparison between the average α-nDCG@5 and Normalized MDCU@5 on the WebDiversity 2012 collection.

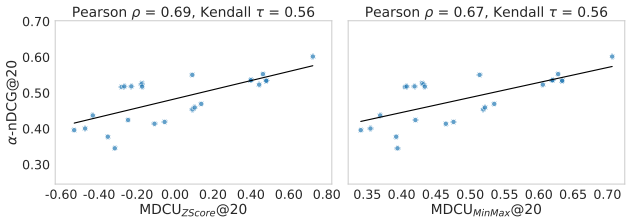

Figure 2: Comparison between the average α-nDCG@20 and Normalized MDCU@20 on the WebDiversity 2012 collection.

"directional agreement". The first dimension categorizes system pair comparisons as *Active (A)* if both evaluation measures detect statistically significant differences between the systems, *Mixed (M)* if only one measure identifies significance, and *Passive (P)* if neither measure finds a significant difference between systems. The second axis assesses whether the measure yields consistent rankings, classifying them as *Agreements (A)* when both measures consider the same system to be better and *Disagreements (D)* when rankings conflict. Combining these dimensions results in six concordance measurements: **Active Agreement (AA)**, **Mixed Agreement (MA)**, **Passive Agreement (PA)**, **Active Disagreement (AD)**, **Mixed Disagreement (MD)**, and **Passive Disagreement (PD)**, each capturing different relationships between pairs of systems. Moreover, we employ the *Conclusion Bias*[4] to quantify the proportion of conflicting outcomes where the two evaluation measures lead to opposite findings, i.e., assessing instances where one measure identifies a model as the statistical best while the other suggests the reverse ranking.

4.2 Empirical Results

4.2.1 MDCU Stability Analysis. Figure 1 shows the results of the different runs considering the average α-nDCG and normalized MDCU on the Web Diversity'12 cat-B collection at $k = 5$. The correlation analysis results indicate a Pearson's correlation of 0.96 for both normalization methods concerning the α-nDCG. On the other hand, Kendall's τ is 0.81 for $MDCU_{ZScore}$, increasing to 0.83 in the MinMax normalization. However, the correlation between the measures remains within a non-pathological range. This ensures that while the measures are related, they still capture different aspects of retrieval performance. Since only the top 5 retrieved documents are

[4]Ferro and Sanderson [8] call this score *Publication Bias*, as it would lead to different outcomes being published.

Table 1: Statistical Concordance Analysis between α-nDCG@5 and Normalized MDCU@5 for the Web Diversity catb systems.

Collection	Pairs	MDCU Stat.Diff.	α-nDCG Stat.Diff.	Agreements			Disagreements			Agreements Ratio $\frac{AA+PA}{Pairs}$	Mixed Ratio $\frac{MA+MD}{Pairs}$	Disagreements Ratio $\frac{AD+PD}{Pairs}$	Conclusion Bias $1-\frac{AA}{AA+AD+\frac{MA+MD}{2}}$
				AA	MA	PA	AD	MD	PD				
α-nDCG@5 vs MDCU$_{MinMax}$@5													
WebDiversity'09	231	72	47	15	37	119	7	38	15	0.580	0.325	0.095	0.748
WebDiversity'10	45	28	10	6	14	13	3	6	3	0.423	0.444	0.133	0.684
WebDiversity'11	378	145	32	29	85	211	0	34	19	0.635	0.315	0.050	0.672
WebDiversity'12	300	39	8	8	15	236	0	16	25	0.813	0.103	0.084	0.660
α-nDCG@5 vs MDCU$_{ZScore}$@5													
WebDiversity'09	231	73	47	16	37	118	7	37	16	0.580	0.320	0.100	0.733
WebDiversity'10	45	31	10	6	16	9	3	7	4	0.333	0.511	0.156	0.707
WebDiversity'11	378	162	32	31	94	198	0	38	17	0.606	0.349	0.045	0.680
WebDiversity'12	300	64	8	8	27	208	0	29	28	0.720	0.187	0.093	0.778

Table 2: Statistical Concordance Analysis between α-nDCG@20 and Normalized MDCU@20 for the Web Diversity catb systems.

Collection	Pairs	MDCU Stat.Diff.	α-nDCG Stat.Diff.	Agreements			Disagreements			Agreements Ratio $\frac{AA+PA}{Pairs}$	Mixed Ratio $\frac{MA+MD}{Pairs}$	Disagreements Ratio $\frac{AD+PD}{Pairs}$	Conclusion Bias $1-\frac{AA}{AA+AD+\frac{MA+MD}{2}}$
				AA	MA	PA	AD	MD	PD				
α-nDCG@20 vs MDCU$_{MinMax}$@20													
WebDiversity'09	231	87	46	32	32	110	3	31	23	0.615	0.273	0.112	0.519
WebDiversity'10	45	30	13	7	19	9	4	2	4	0.356	0.467	0.177	0.674
WebDiversity'11	378	184	38	37	113	151	0	35	42	0.497	0.392	0.111	0.667
WebDiversity'12	300	55	6	5	42	178	0	9	66	0.610	0.170	0.220	0.836
α-nDCG@20 vs MDCU$_{ZScore}$@20													
WebDiversity'09	231	95	46	38	29	108	3	30	23	0.632	0.255	0.113	0.461
WebDiversity'10	45	30	13	7	19	8	4	2	5	0.333	0.467	0.200	0.674
WebDiversity'11	378	203	38	38	122	139	0	43	36	0.468	0.437	0.095	0.685
WebDiversity'12	300	61	6	5	47	172	0	10	66	0.590	0.190	0.220	0.851

considered in the evaluation, there is limited opportunity for the documents to overlap themes, thus leading to a higher correlation between the measures.

In Figure 2 are reported the results of the different runs considering the average α-nDCG and MDCU on the Web Diversity cat-B collections with a cut-off point equal to 20, as used in the challenge evaluation [5]. Increasing the cut-off value k to 20 allows the measures to consider a larger set of retrieved documents, thereby increasing the likelihood of thematic overlap. Therefore, α-nDCG and the normalized MDCU capture different aspects of systems effectiveness in the retrieval. In this case, Pearson's correlation decreases to 0.69 and 0.67 for the Z-Score and MinMax normalization strategies, respectively. While the correlation remains positive, Kendall's τ reaches a value of 0.56, indicating that, despite the differences, a significant positive degree of concordance between the two measures persists.

4.2.2 MDCU Statistical Power. Table 1 presents the concordance results for $k = 5$. The normalized MDCU measure demonstrates a stronger ability to identify statistically significant differences between system pairs. Moreover, the number of ADs, representing the most undesirable outcome in the analysis, consistently remains the lowest, indicating that the two measures rarely produce conflicting rankings between systems. Notably, the agreement ratio, i.e., the sum of active and passive agreements, shows strong concordance between the system rankings found in all four collections.

Table 2 shows the concordance results for the cut-off at 20. At a higher cut-off value, the ability of the normalized MDCU to identify statistically significant differences remains. Likewise, the patterns of AA and AD are consistently maintained. In addition, the measure preserves the same system ranking as α-nDCG, as indicated by the consistently low counts of AD and MD, confirming minimal statistically significant different ranking contradictions.

For both cut-off points, the disagreement ratio corresponds to the lowest percentage of cases observed, indicating that regardless of whether the two systems are statistically different, the level of discordant pairs identified by the measures remains lower than the mixed and agreement ratios. Finally, the *Conclusion Bias* highlights that, in most cases, the use of MDCU or α-nDCG leads to opposite findings, emphasizing their differing evaluation perspectives.

5 CONCLUSION

In this paper, we propose the implementation of the MDCU framework and evaluate its effectiveness using the α-nDCG measure as the baseline to assess novelty and diversity with overlapping themes in system search results. We also performed a statistical analysis of the results obtained in four TREC collections, investigating the concordance between these two measures. Our findings indicate strong positive correlations among these metrics at a cut-off of 5, which exhibits a slightly lower positive correlation at a cut-off of 20 due to the higher number of distinct relevance aspects assessed considering a larger pool of documents.

In future work, we intend to develop *ad hoc* collections to assess and examine the impact of usability attributes on the final MDCU score and its correlation with the α-nDCG. The usability attributes may be derived by employing Large Language Models to outline the usability aspects of the systems' retrieved documents.

REFERENCES

[1] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009 (NIST Special Publication, Vol. 500-278)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf

[2] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, Irwin King, Wolfgang Nejdl, and Hang Li (Eds.). ACM, 75–84. https://doi.org/10.1145/1935826.1935847

[3] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. 2010. Overview of the TREC 2010 Web Track. In *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010 (NIST Special Publication, Vol. 500-294)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec19/papers/WEB.OVERVIEW.pdf

[4] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. 2011. Overview of the TREC 2011 Web Track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011 (NIST Special Publication, Vol. 500-296)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec20/papers/WEB.OVERVIEW.pdf

[5] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012 (NIST Special Publication, Vol. 500-298)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec21/papers/WEB12.overview.pdf

[6] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 659–666. https://doi.org/10.1145/1390334.1390446

[7] Guglielmo Faggioli and Nicola Ferro. 2021. System Effect Estimation by Sharding: A Comparison Between ANOVA Approaches to Detect Significant Differences. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 33–46. https://doi.org/10.1007/978-3-030-72240-1_3

[8] Nicola Ferro and Mark Sanderson. 2022. How Do You Test a Test?: A Multifaceted Examination of Significance Tests. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 280–288. https://doi.org/10.1145/3488560.3498406

[9] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[10] Kalervo Järvelin and Eero Sormunen. 2024. A Blueprint of IR Evaluation Integrating Task and User Characteristics. *ACM Trans. Inf. Syst.* 42, 6 (2024), 164:1–164:38. https://doi.org/10.1145/3675162

[11] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1-2 (1938), 81–93.

[12] Zurab Khasidashvili and John R. W. Glauert. 1996. Discrete Normalization and Standardization in Deterministic Residual Structures. In *Algebraic and Logic Programming, 5th International Conference, ALP'96, Aachen, Germany, September 25-27, 1996, Proceedings (Lecture Notes in Computer Science, Vol. 1139)*, Michael Hanus and Mario Rodríguez-Artalejo (Eds.). Springer, 135–149. https://doi.org/10.1007/3-540-61735-3_9

[13] Stefano Mizzaro. 1998. How many relevances in information retrieval? *Interact. Comput.* 10, 3 (1998), 303–320. https://doi.org/10.1016/S0953-5438(98)00012-5

[14] Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and metrics: IR evaluation as a user process. In *The Seventeenth Australasian Document Computing Symposium, ADCS '12, Dunedin, New Zealand, December 5-6, 2012*, Andrew Trotman, Sally Jo Cunningham, and Laurianne Sitbon (Eds.). ACM, 47–54. https://doi.org/10.1145/2407085.2407092

[15] Karl Pearson. 1895. VII. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London* 58, 347-352 (1895), 240–242.

[16] Georgios Peikos and Gabriella Pasi. 2024. A systematic review of multidimensional relevance estimation in information retrieval. *WIREs Data. Mining. Knowl. Discov.* 14, 5 (2024). https://doi.org/10.1002/WIDM.1541

[17] Andrew Rutherford. 2011. *ANOVA and ANCOVA: a GLM approach.* John Wiley & Sons.

[18] Sidney Siegel and John W. Tukey. 1960. A Nonparametric Sum of Ranks Procedure for Relative Spread in Unpaired Samples. *J. Amer. Statist. Assoc.* 55 (1960), 429–445. https://api.semanticscholar.org/CorpusID:121903915

[19] Alkis Simitsis, Akanksha Baid, Yannis Sismanis, and Berthold Reinwald. 2008. Multidimensional content eXploration. *Proc. VLDB Endow.* 1, 1 (2008), 660–671. https://doi.org/10.14778/1453856.1453929

[20] Raphael Vallat. 2018. Pingouin: statistics in Python. *Journal of Open Source Software* 3, 31 (Nov. 2018), 1026. https://doi.org/10.21105/joss.01026

[21] William Webber, Alistair Moffat, and Justin Zobel. 2008. Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 51–58. https://doi.org/10.1145/1390334.1390346

[22] Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Bowei He, Bhaskar Mitra, and Xue Liu. 2024. Result Diversification in Search and Recommendation: A Survey. *IEEE Trans. Knowl. Data Eng.* 36, 10 (2024), 5354–5373. https://doi.org/10.1109/TKDE.2024.3382262