MDCU: A Multi-Dimensional Cumulative Utility Metric for Information Retrieval.*

Francesco Luigi, De Faveri[1], Guglielmo, Faggioli[1], Nicola, Ferro[1] and Kalervo Järvelin[2]

[1] *Department of Information Engineering, University of Padova, Padova, Italy*
[2] *Tampere University, Tampere, Finland*

Abstract

Information Retrieval (IR) effectiveness metrics typically assume that a relevant document fully satisfies the user's information need. However, this assumption becomes inadequate when such information need are faceted or comprise multiple subtopics, addressing only a subset of them. Additionally, ranked lists may contain multiple documents focusing on the same subtopics, leading to content redundancy while neglecting other aspects of the information. Search results that present top-ranked documents covering diverse subtopics are generally more desirable than those offering overlapping content on limited facets. The Multi-Dimensional Cumulated Utility (MDCU) metric, introduced by Järvelin and Sormunen, addresses this issue by incorporating content overlap into the evaluation of novelty and diversity. While their work demonstrated MDCU's conceptual foundation and illustrated it with a simplified example, its empirical application has yet to be explored. In this study, we empirically evaluate the practical stability of MDCU using publicly available TREC test collections. Moreover, we examine its relationship with the well-established α-nDCG metric and present a Python implementation of MDCU to support its adoption in future evaluation settings. Our experimental results reveal a strong positive correlation between MDCU and α-nDCG, indicating that both metrics consistently capture similar performance trends across IR systems. Moreover, compared to α-nDCG, MDCU demonstrates greater statistical power, identifying up to nine times more statistically significant differences between system pairs.

Keywords

Evaluation Metrics, Multidimensional Cumulated Utility, Information Retrieval

1. Introduction

Traditional Information Retrieval (IR) evaluation metrics, such as those in the Cumulated Gain family [2], are based on mono-dimensional relevance judgments that assume queries are mono-thematic and focus solely on topical relevance. This abstraction supports efficient offline evaluation but fails to capture the complexity of real-world information needs, which often span multiple subtopics and involve factors such as novelty, redundancy, and document accessibility [3, 4, 5]. To address these limitations, multi-dimensional evaluation measures have been proposed [6, 7].

A recent development in this area is the Multi-Dimensional Cumulated Utility (MDCU) [7], which accounts for thematic diversity, content overlap, and document-level attributes (e.g., language, complexity, recency). 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 [8]. While the theoretical foundations of MDCU have been

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

[*] This is an extended abstract of [1].

francescoluigi.defaveri@phd.unipd.it (F. L. De Faveri); guglielmo.faggioli@unipd.it (G. Faggioli); nicola.ferro@unipd.it (N. Ferro); kalervo.jarvelin@tuni.fi (K. Järvelin)
https://www.dei.unipd.it/~defaverifr/ (F. L. De Faveri); https://www.dei.unipd.it/~faggioli/ (G. Faggioli); https://www.dei.unipd.it/~ferro/ (N. Ferro); https://www.tuni.fi/en/people/kalervo-jarvelin (K. Järvelin)
0009-0005-8968-9485 (F. L. De Faveri); 0000-0002-5070-2049 (G. Faggioli); 0000-0001-9219-6239 (N. Ferro); 0000-0001-7655-8930 (K. Järvelin)

established, its empirical performance remains unexplored. In this work, we operationalise MDCU as a practical evaluation metric, apply it to TREC Web Diversity Track collections, compare it with the established α-nDCG [6], and release a public implementation[1].

2. Background on Multi-Dimensional Evaluation Measures

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. The $r_{i,t}$ value can be binary or graded.

The Multi-Dimensional Cumulated Utility (MDCU) framework introduced in [7] 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 $(a_{i,1}, \ldots, a_{i,m})$, $(a_{i,j})_{1 \leq j \leq m} \in [0, 1]$. The algorithm takes as input the ranked list of search results, denoted as $\mathcal{D}$, and a discounting parameter b, accounting for the overlap. 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. Challenges and Solutions of the MDCU

Collections for Computing MDCU. Identifying a suitable test collection to validate empirically MDCU represents a preliminary experimental challenge. As noted by Järvelin and Sormunen [7], no collection contains document annotations for multi-theme relevance and usability. We stress that, at the moment of the experiments, the literature lacks a suitable benchmark test collection that assesses both multi-dimensional relevance themes and usability attributes. Consequently, we focus on the multi-theme evaluation, leaving the investigation of the role of the usability attributes, e.g., score that simulate the credibility, virality or sensitivity analysis of retrieved documents, as future work. We remark that, as noted in the seminal MDCU paper by Järvelin and Sormunen [7], not considering the usability attributes in the MDCU analysis means that the measure only accounts for the impact of the themes' relevance, and each document is used equally by the user. 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 [9, 10, 11, 12]. 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.

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 the MDCU computation. 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 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 [13].

[1] https://github.com/Kekkodf/MDCUEval

We follow Webber et al. [14] 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:

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

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:

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

4. Experimental Evaluation

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 in the online repository.

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 correspond to the measurement standard adopted in the original challenges of the Web Diversity track [9, 10, 11, 12]. For space limitations, we discuss here only the results for $k = 5$, but a more comprehensive discussion is provided in the paper [1] and in the repository. We analyze the correlation and agreement between different runs by computing Pearson's ρ and Kendall's τ correlation coefficients [15, 16] 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.

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 [19] to assess the concordance between α-nDCG and MDCU, as proposed in [20, 21, 22]. In detail, the concordance measures proposed by [20, 21, 22] 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 "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 the following 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 analysed. Moreover, we employ the *Conclusion Bias*[4] to

[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 [12]. The results on the other collections are reported in the repository.

[4] Ferro and Sanderson [22] 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 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

quantify the proportion of conflicting outcomes discovered where the two evaluation measures lead to opposite findings, i.e., assessing instances where one measure identifies a model as the statistical best.

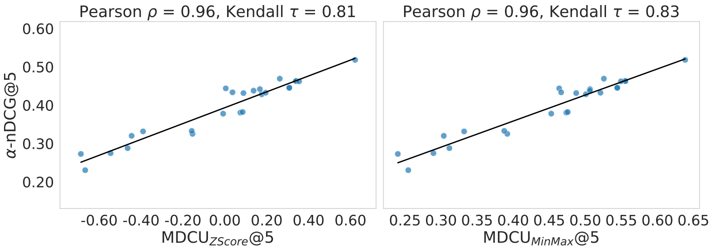

Figure 1: Correlation on the WebDiversity 2012 collection.

MDCU Stability Analysis. Figure 4 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 considered in the evaluation, there is a limited opportunity for the documents to overlap themes, leading to a higher correlation.

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. Finally, the *Conclusion Bias* highlights that the use of MDCU or α-nDCG emphasises 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. One study limitation is the absence of an analysis of the documents' usability attributes. Since the literature lacks a suitable benchmark, we intend to develop *ad hoc* collections. The usability attributes may be derived by employing LLMs to generate the usability aspects of the systems' retrieved documents, thus simulating, for example, the virality and credibility of the texts.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for Readability and Spelling checks. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] F. L. De Faveri, G. Faggioli, N. Ferro, K. Järvelin, Evaluating multi-dimensional cumulated utility in information retrieval, in: N. Ferro, M. Maistro, G. Pasi, O. Alonso, A. Trotman, S. Verberne (Eds.), Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025, ACM, 2025, pp. 2622–2626. URL: https://doi.org/10.1145/3726302.3730191. doi:10.1145/3726302.3730191.

[2] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: http://doi.acm.org/10.1145/582415.582418. doi:10.1145/582415.582418.

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

[4] C. L. A. Clarke, N. Craswell, I. Soboroff, A. Ashkan, A comparative analysis of cascade measures for novelty and diversity, in: I. King, W. Nejdl, H. Li (Eds.), Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011, ACM, 2011, pp. 75–84. URL: https://doi.org/10.1145/1935826.1935847. doi:10.1145/1935826.1935847.

[5] H. Wu, Y. Zhang, C. Ma, F. Lyu, B. He, B. Mitra, X. Liu, Result diversification in search and recommendation: A survey, IEEE Trans. Knowl. Data Eng. 36 (2024) 5354–5373. URL: https://doi.org/10.1109/TKDE.2024.3382262. doi:10.1109/TKDE.2024.3382262.

[6] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon, Novelty and diversity in information retrieval evaluation, in: S. Myaeng, D. W. Oard, F. Sebastiani, T. Chua, M. Leong (Eds.), Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008, ACM, 2008, pp. 659–666. URL: https://doi.org/10.1145/1390334.1390446. doi:10.1145/1390334.1390446.

[7] K. Järvelin, E. Sormunen, A blueprint of IR evaluation integrating task and user characteristics, ACM Trans. Inf. Syst. 42 (2024) 164:1–164:38. URL: https://doi.org/10.1145/3675162. doi:10.1145/3675162.

[8] N. Fuhr, A. Giachanou, G. Grefenstette, I. Gurevych, A. Hanselowski, K. Järvelin, R. Jones, Y. Liu, J. Mothe, W. Nejdl, I. Peters, B. Stein, An information nutritional label for online documents, SIGIR Forum 51 (2017) 46–66. URL: https://doi.org/10.1145/3190580.3190588. doi:10.1145/3190580.3190588.

[9] C. L. A. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2009 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009, volume 500-278 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2009. URL: http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf.

[10] C. L. A. Clarke, N. Craswell, I. Soboroff, G. V. Cormack, Overview of the TREC 2010 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010, volume 500-294 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2010. URL: https://trec.nist.gov/pubs/trec19/papers/WEB.OVERVIEW.pdf.

[11] C. L. A. Clarke, N. Craswell, I. Soboroff, E. M. Voorhees, Overview of the TREC 2011 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Twentieth Text REtrieval

Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011, volume 500-296 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2011. URL: http://trec.nist.gov/pubs/trec20/papers/WEB.OVERVIEW.pdf.

[12] C. L. A. Clarke, N. Craswell, E. M. Voorhees, Overview of the TREC 2012 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012, volume 500-298 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2012. URL: http://trec.nist.gov/pubs/trec21/papers/WEB12.overview.pdf.

[13] Z. Khasidashvili, J. R. W. Glauert, Discrete normalization and standardization in deterministic residual structures, in: M. Hanus, M. Rodríguez-Artalejo (Eds.), Algebraic and Logic Programming, 5th International Conference, ALP'96, Aachen, Germany, September 25-27, 1996, Proceedings, volume 1139 of *Lecture Notes in Computer Science*, Springer, 1996, pp. 135–149. URL: https://doi.org/10.1007/3-540-61735-3_9. doi:10.1007/3-540-61735-3_9.

[14] W. Webber, A. Moffat, J. Zobel, Score standardization for inter-collection comparison of retrieval systems, in: S. Myaeng, D. W. Oard, F. Sebastiani, T. Chua, M. Leong (Eds.), Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008, ACM, 2008, pp. 51–58. URL: https://doi.org/10.1145/1390334.1390346. doi:10.1145/1390334.1390346.

[15] K. Pearson, Vii. note on regression and inheritance in the case of two parents, proceedings of the royal society of London 58 (1895) 240–242.

[16] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.

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

[18] S. Siegel, J. W. Tukey, A nonparametric sum of ranks procedure for relative spread in unpaired samples, Journal of the American Statistical Association 55 (1960) 429–445. URL: https://api.semanticscholar.org/CorpusID:121903915.

[19] R. Vallat, Pingouin: statistics in python, Journal of Open Source Software 3 (2018) 1026. doi:10.21105/joss.01026.

[20] A. Moffat, F. Scholer, P. Thomas, Models and metrics: IR evaluation as a user process, in: A. Trotman, S. J. Cunningham, L. Sitbon (Eds.), The Seventeenth Australasian Document Computing Symposium, ADCS '12, Dunedin, New Zealand, December 5-6, 2012, ACM, 2012, pp. 47–54. URL: https://doi.org/10.1145/2407085.2407092. doi:10.1145/2407085.2407092.

[21] G. Faggioli, N. Ferro, System effect estimation by sharding: A comparison between ANOVA approaches to detect significant differences, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, volume 12657 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 33–46. URL: https://doi.org/10.1007/978-3-030-72240-1_3. doi:10.1007/978-3-030-72240-1_3.

[22] N. Ferro, M. Sanderson, How do you test a test?: A multifaceted examination of significance tests, in: K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, J. Tang (Eds.), WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022, ACM, 2022, pp. 280–288. URL: https://doi.org/10.1145/3488560.3498406. doi:10.1145/3488560.3498406.