Appendix of Principled Multi-Aspect Evaluation of Rankings

Maria Maistro
mm@di.ku.dk
University of Copenhagen
Denmark

Lucas Chaves Lima
lcl@di.ku.dk
University of Copenhagen
Denmark

Jakob Grue Simonsen
simonsen@di.ku.dk
University of Copenhagen
Denmark

Christina Lioma
c.lioma@di.ku.dk
University of Copenhagen
Denmark

ABSTRACT

This is the appendix of the paper: "Principled Multi-Aspect Evaluation of Rankings" [3].

CCS CONCEPTS

• **Information systems** → **Test collections**; *Relevance assessment*; Presentation of retrieval results.

KEYWORDS

Evaluation, ranking, multiple aspects, partial orders

A TOTAL ORDER MULTI-ASPECT EVALUATION (TOMA)

In this section we prove that the *Total Order Multi-Aspect (TOMA)* approach satisfies the partial order, as stated at the and of Section 3.3 [3], and then we prove that the ideal ranking built with TOMA can reach the upper bound 1, as stated in Section 4.2 [3].

A.1 Proof 1: TOMA Respects the Partial Order

Given a set of documents D, a set of aspects A, a set of topics T, and a ground-truth map GT a partial order on the set of tuples of labels $L = \bigtimes_{a \in A} L_a$ is defined as follows:

$$\mathrm{GT}(d,t) \sqsubseteq \mathrm{GT}(d',t) \iff l_i \preceq_{a_i} l'_i \quad \forall i \in \{1,\dots,n\} \tag{1}$$

where $\mathrm{GT}(d,t) = (l_1,\dots,l_n)$ and $\mathrm{GT}(d',t) = (l'_1,\dots,l'_n)$.

Let g be an embedding function that maps tuples of labels in Euclidean space $\mathcal{L} = \mathbb{R}^n$: $g(\mathbf{l}) = g(l_1,\dots,l_n) = (g_{a_1}(l_1),\dots,g_{a_n}(l_n))$. We assume that for each $a \in A$, g_a is a non-decreasing map, i.e., for any $l, l' \in L_a$ if $l \preceq_a l'$ than $g_a(l) \leq g_a(l')$. Through the embedding function g, each tuple of labels $\mathbf{l}$ is represented by a point in the Euclidean space $\mathcal{L}$ denoted by $\vec{l} = g(\mathbf{l})$.

We define the *distance order* $\preceq_*$: a weak order on L such that:

$$\mathbf{l} \preceq_* \mathbf{l}' \iff \mathrm{Dist}(\vec{l},\vec{l^\star}) \geq \mathrm{Dist}(\vec{l'},\vec{l^\star}) \tag{2}$$

where $\mathrm{Dist}\colon \mathcal{L}\times\mathcal{L} \to [0,+\infty[$ is any function such that $\mathrm{Dist}(\vec{l^\star},\vec{l^\star}) = 0$. Moreover, since $\preceq_*$ is a weak order, we write:

$$\mathbf{l} =_* \mathbf{l}' \iff \mathrm{Dist}(\vec{l},\vec{l^\star}) = \mathrm{Dist}(\vec{l'},\vec{l^\star}) \tag{3}$$

Theorem A.1. *When TOMA is instantiated with Euclidean, Manhattan or Chebyshev as distance function, the distance order in Equation* (2) *and Equation* (3) *respects the partial order in Equation* (1)*:*

$$\forall\, \mathbf{l}, \mathbf{l}' \in L \text{ we have } \mathbf{l} \sqsubseteq \mathbf{l}' \Rightarrow \mathbf{l} \preceq_* \mathbf{l}' \tag{4}$$

Proof. We denote the distance order instantiated with Manhattan as $\preceq_1$, with Euclidean as $\preceq_2$, and with Chebyshev as $\preceq_\infty$.

For any tuple of labels $\mathbf{l}$ and $\mathbf{l}'$ in L:

$$\mathbf{l} \sqsubseteq \mathbf{l}' \iff l_a \preceq l'_a \quad \forall a \in A. \tag{5}$$

Since g_a is non-decreasing, and by definition of $\mathbf{l}^\star$, we have:

$$g_a(l_a) \leq g_a(l'_a) \leq g_a(l^\star_a) \quad \forall a \in A. \tag{6}$$

This implies:

$$g_a(l_a) - g_a(l^\star_a) \leq g_a(l'_a) - g_a(l^\star_a) \leq 0 \quad \forall a \in A \tag{7}$$

and by considering the absolute value:

$$|g_a(l_a) - g_a(l^\star_a)| \geq |g_a(l'_a) - g_a(l^\star_a)| \quad \forall a \in A \tag{8}$$

which means that $\mathrm{Dist}(\vec{l},\vec{l^\star}) \geq \mathrm{Dist}(\vec{l'},\vec{l^\star})$ with Manhattan distance, i.e. $\mathbf{l} \preceq_1 \mathbf{l}'$.

Analogously, taking the square values in Equation (7), we obtain:

$$\left(g_a(l_a) - g_a(l^\star_a)\right)^2 \geq \left(g_a(l'_a) - g_a(l^\star_a)\right)^2 \quad \forall a \in A \tag{9}$$

which implies that $\mathrm{Dist}(\vec{l},\vec{l^\star}) \geq \mathrm{Dist}(\vec{l'},\vec{l^\star})$ with Euclidean distance, i.e. $\mathbf{l} \preceq_2 \mathbf{l}'$.

To conclude, since $|A| < \infty$ there exists $\bar{a} \in A$ such that:

$$\max_{a\in A} |g_a(l'_a) - g_a(l^\star_a)| = |g_{\bar{a}}(l'_{\bar{a}}) - g_{\bar{a}}(l^\star_{\bar{a}})| \tag{10}$$

by Equation (9) we have:

$$|g_{\bar{a}}(l'_{\bar{a}}) - g_{\bar{a}}(l^\star_{\bar{a}})| \leq |g_a(l_a) - g_a(l^\star_a)| \leq \max_{a\in A} |g_a(l_a) - g_a(l^\star_a)| \tag{11}$$

$\forall a \in A$, which implies that $\mathrm{Dist}(\vec{l},\vec{l^\star}) \geq \mathrm{Dist}(\vec{l'},\vec{l^\star})$ with Chebyshev distance, i.e. $\mathbf{l} \preceq_\infty \mathbf{l}'$.

□

Note that TOMA is defined for any type and number of aspects, with the hypothesis that the aspects and their labels can be embedded in an Euclidean space. Indeed, $\preceq_2$, $\preceq_1$, $\preceq_\infty$, and the corresponding distance functions are always defined for any dimension of the

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11.
https://doi.org/10.1145/3459637.3482287

Euclidean space (number of aspects) and any point in the Euclidean space (embedded labels).

A.2 Proof 2: TOMA Upper Bound

Let D be a set of items, A a set of aspects, T a set of topics, and μ any single-aspect IR evaluation measure such that μ scores range in $[0, 1]$.

Theorem A.2. *With TOMA approach for all ranked lists $r_t \in D^*$ we have:*

$$M(r_t) \leq 1 \qquad \textit{and} \qquad \exists i \in D^* : M(i_t) = 1 \tag{12}$$

where M is a multi-aspect evaluation measure defined as in step 3 *of TOMA in Section* 3.3 *[3]:*

$$M = \mu \circ W : M(r_t) = \mu(W(GT(d_1, t)), \ldots, W(GT(d_N, t))) \tag{13}$$

Proof. Let $\hat{r}_t$ be the assessed run for a given topic t, i.e. $\hat{r}_t = (\mathrm{GT}(d_1, t)), \ldots, W(\mathrm{GT}(d_N, t)))$.

It is trivial to prove that the left hand side of Equation (12) holds. Indeed, TOMA first aggregates the tuples of labels, and then computes the measure score as $\mu(W(\hat{r}_t))$. Thus by definition of μ, $\mu(W(\hat{r}_t)) \leq 1$.

The right hand side of Equation (12) is a consequence of W being non decreasing with respect to $\preceq_*$, combined with the behaviour of single-aspect IR evaluation measures. For any single-aspect evaluation measure, we can define an *ideal ranking* $i \in D^*$, which is obtained by sorting the documents in D by decreasing label: e.g., in the case of only relevance, i is built by ranking documents by decreasing relevance label [1].

Intuitively, the ideal ranking is the best way in which one can rank the documents in D. Thus, any single-aspect ranking evaluation measure assigns the maximum achievable score to the ideal ranking. For example, *Average Precision (AP)* returns a score equal to 1 to any ranking where relevant documents are placed before non-relevant ones. Similarly, *Normalized Discounted Cumulated Gain (NDCG)* returns 1 if we sort the documents by their relevance label: first all relevant documents, followed by moderately relevant, partially relevant and non-relevant.

We define the ideal ranking in the multi-aspect setting with the order relation induced by $\preceq_*$ in D:

$$i = (d_1, d_2, \ldots, d_N) \quad \text{such that} \quad \mathrm{GT}(d_i, t) \preceq_* \mathrm{GT}(d_{i-1}, t) \tag{14}$$

$\forall i \in \{1, \ldots, N\}$. By definition of the weight function W:

$$\mathrm{GT}(d, t) \preceq_* \mathrm{GT}(d', t) \implies W(\mathrm{GT}(d, t)) \leq W(\mathrm{GT}(d', t)) \tag{15}$$

thus when W maps the documents in the ideal ranking i to the their weights, it returns a list of decreasing weights, thus an ideal ranking in the single-aspect case. This means that

$$M(i_t) = \mu(W(\mathrm{GT}(d_1, t)), \ldots, W(\mathrm{GT}(d_N, t))) = 1 \tag{16}$$

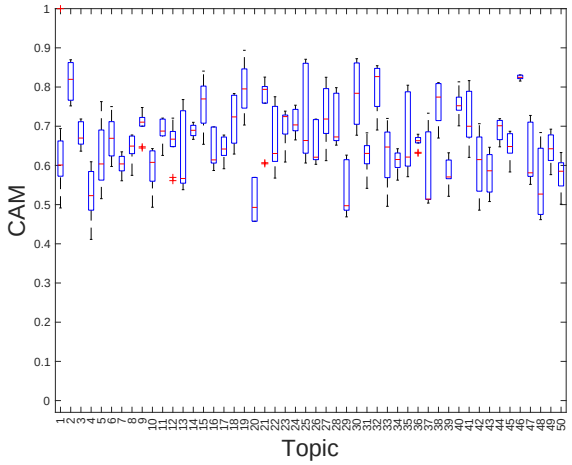
□

B EXPERIMENTAL EVALUATION

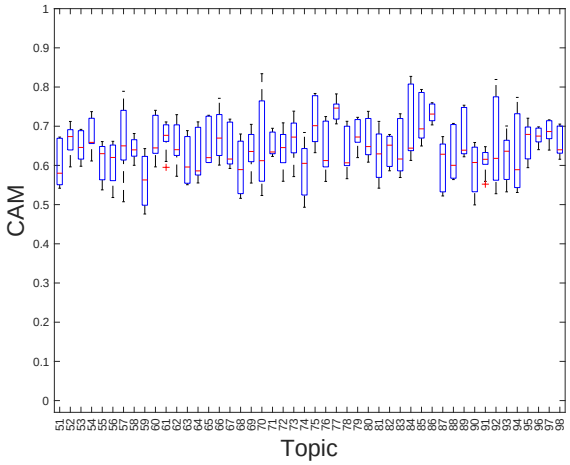
In this section we report the boxplots for the distribution of *Convex Aggregating Measure (CAM)* [2] and *Multidimensional Measure (MM)* [4] scores. These corresponds to Figure 2 in [3], specifically:

- Figure 1: CAM instantiated with AP;
- Figure 2: CAM instantiated with NDCG;
- Figure 3: MM instantiated with AP;
- Figure 4: MM instantiated with NDCG.

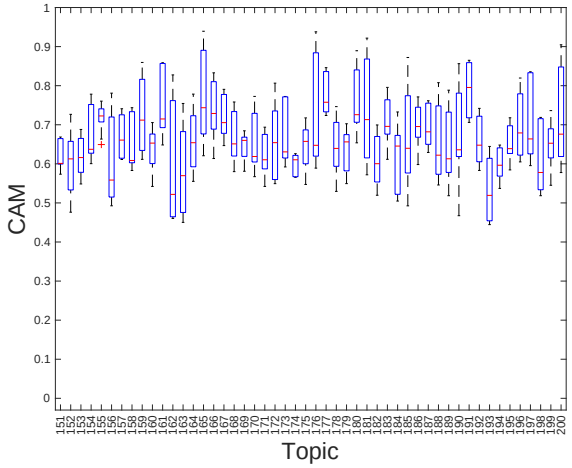
We can see that the upper bound X achieved by CAM and MM depends on the dataset, the set of aspects and the topics. For the Web tracks, CAM and MM can not achieve an upper bound equal to 1 for any of the topics. The Misinformation tracks is the only exception, where CAM and MM can achieve the upper bound 1 for the majority of the topics.



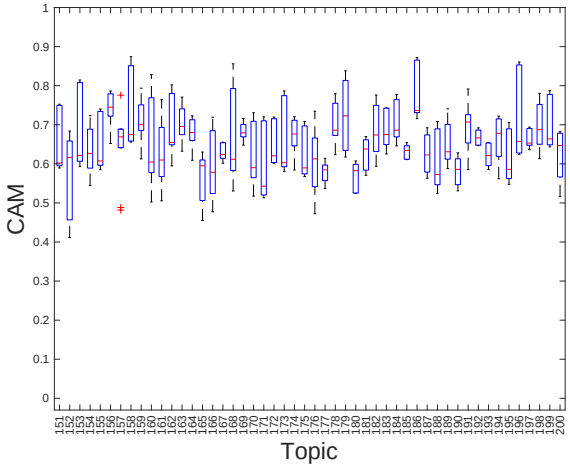
(a) Web Track 2009



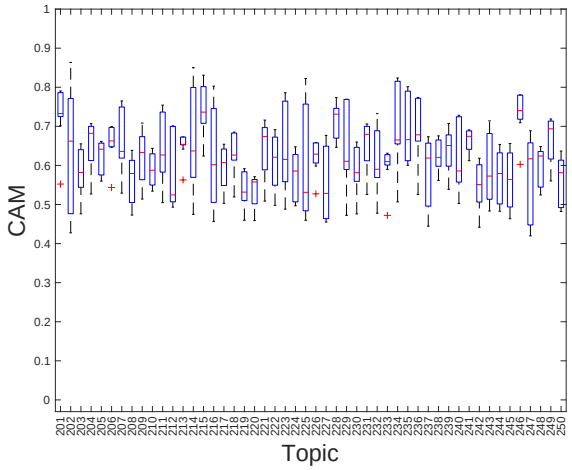
(b) Web Track 2010



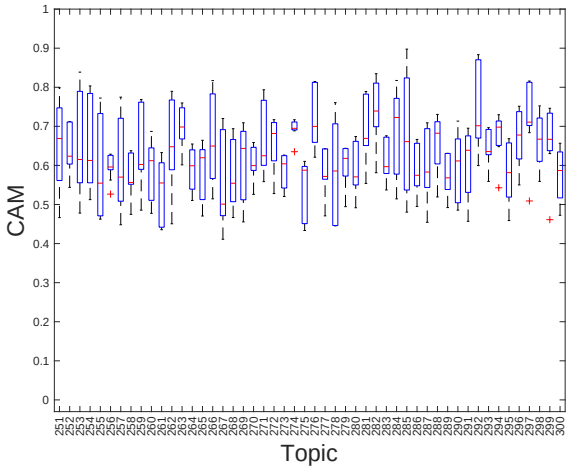
(c) Web Track 2011



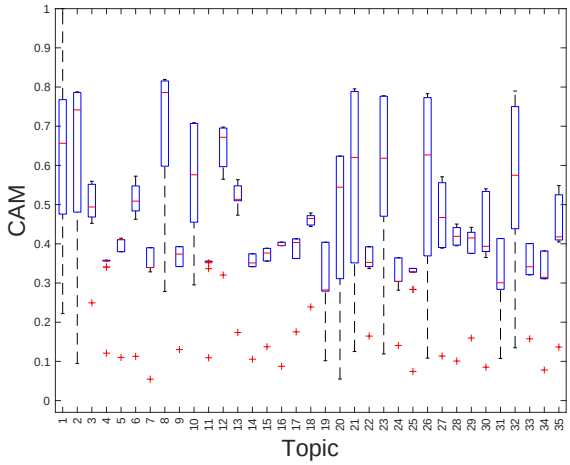
(d) Web Track 2012



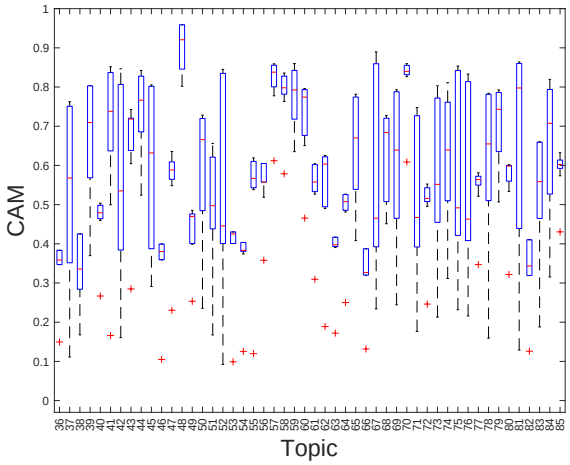
(e) Web Track 2013



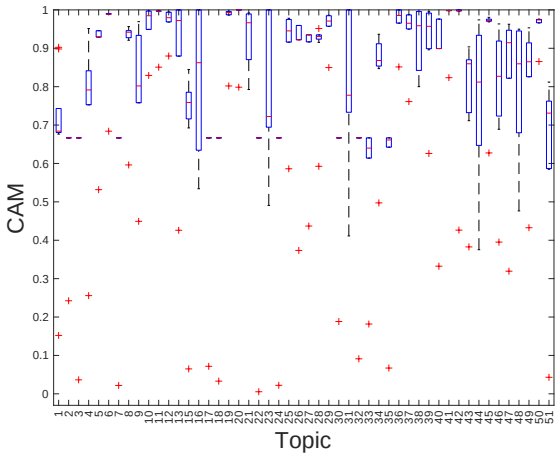
(f) Web Track 2014



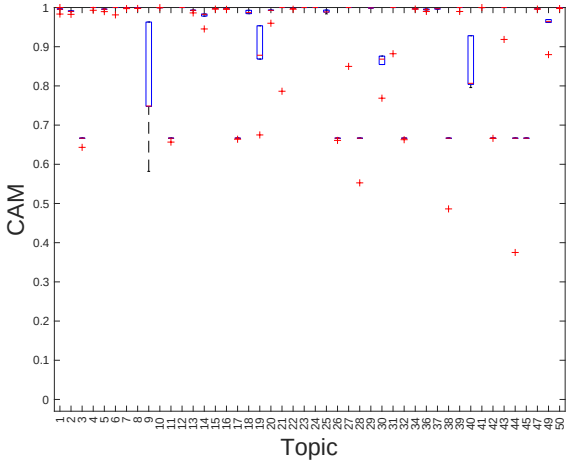
(g) Task Track 2015



(h) Task Track 2016

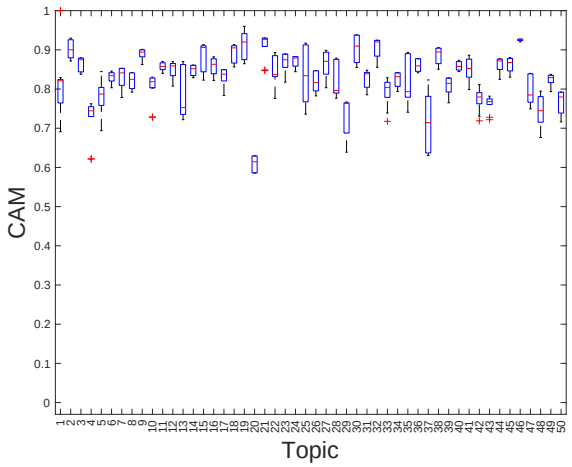


(i) Decision Track 2019

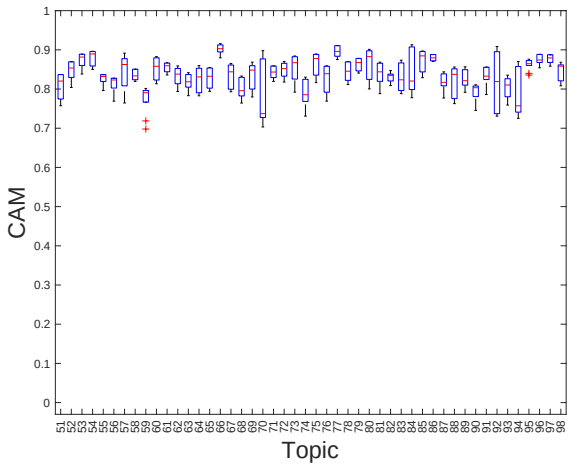


(j) Misinformation Track 2020

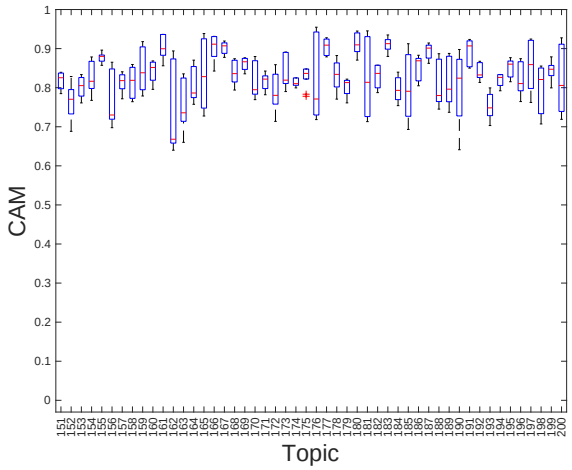
Figure 1: Box-plots for CAM instantiated with AP. The x-axis reports the query number, the y-axis reports CAM scores. The maximum achievable value for CAM depends on the query, the aspects, and the dataset.



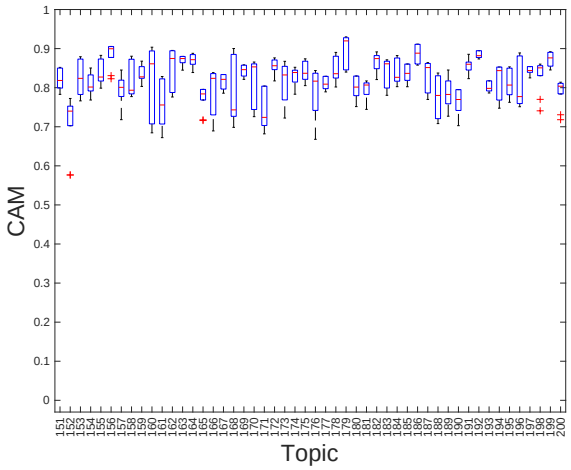
(a) Web Track 2009



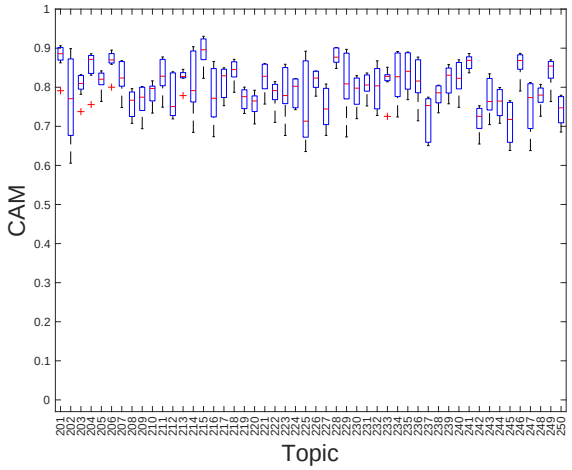
(b) Web Track 2010



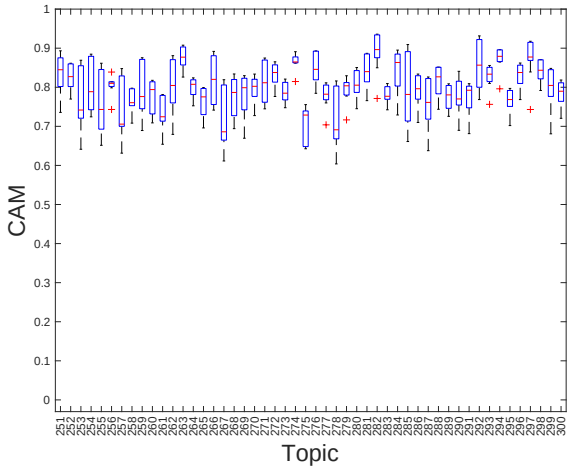
(c) Web Track 2011



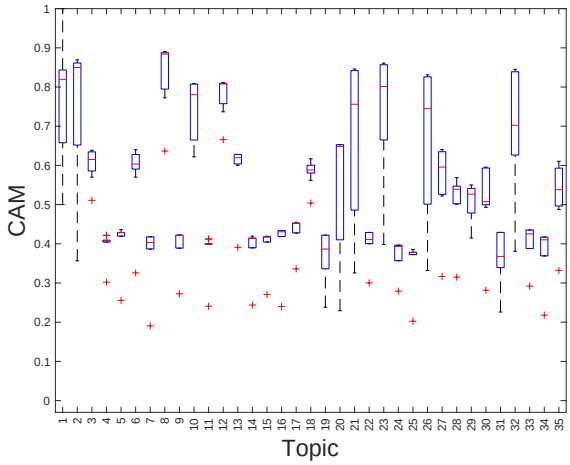
(d) Web Track 2012



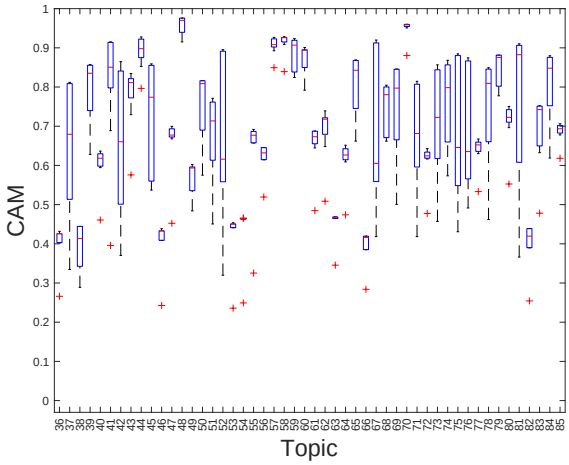
(e) Web Track 2013



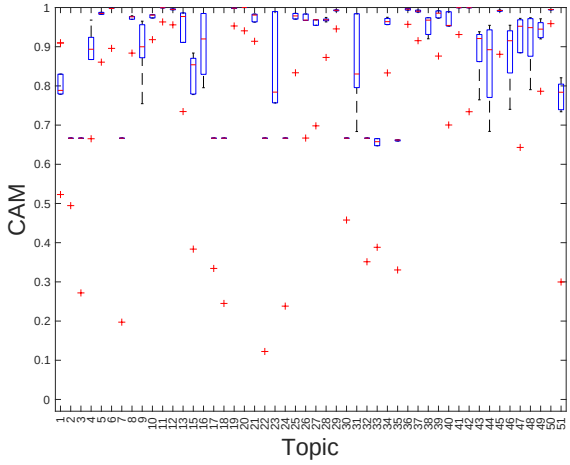
(f) Web Track 2014



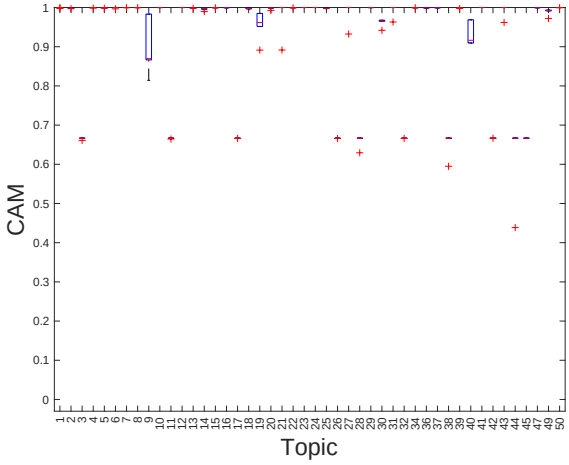
(g) Task Track 2015



(h) Task Track 2016

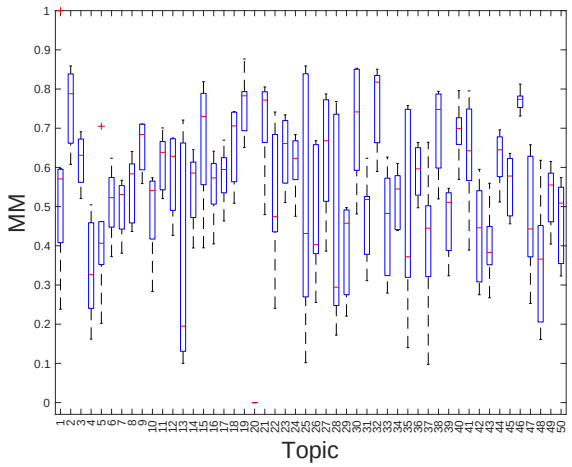


(i) Decision Track 2019

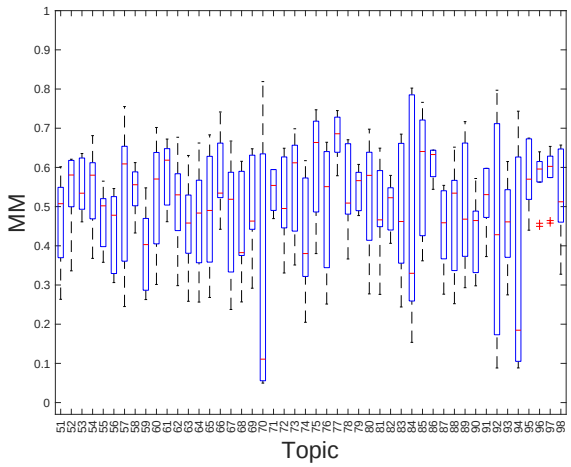


(j) Misinformation Track 2020

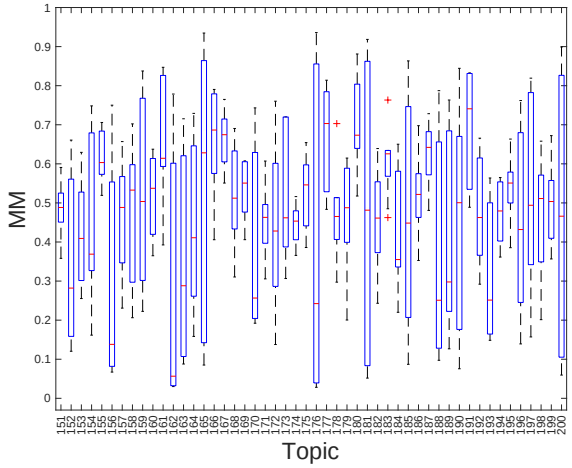
Figure 2: Box-plots for CAM instantiated with NDCG. The x-axis reports the query number, the y-axis reports CAM scores. The maximum achievable value for CAM depends on the query, the aspects, and the dataset.



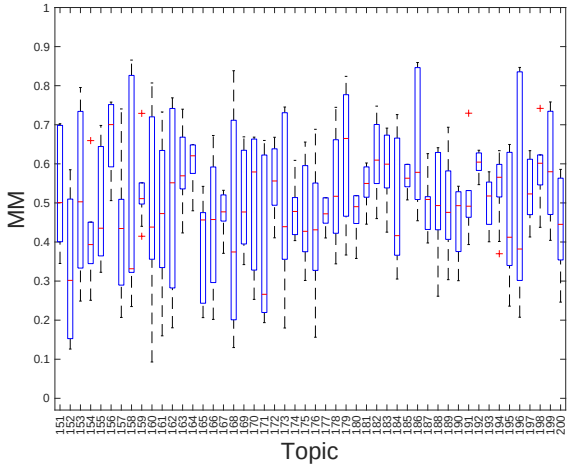
(a) Web Track 2009



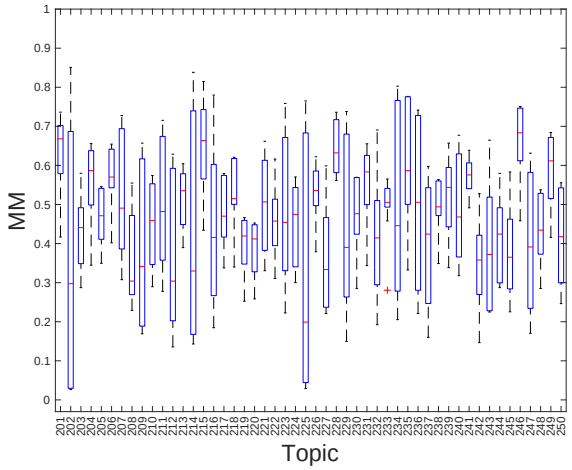
(b) Web Track 2010



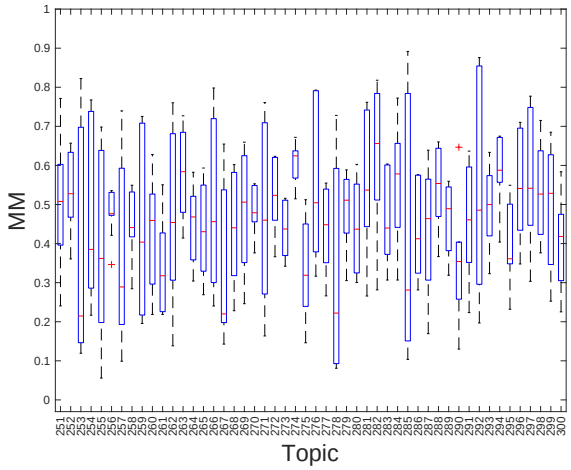
(c) Web Track 2011



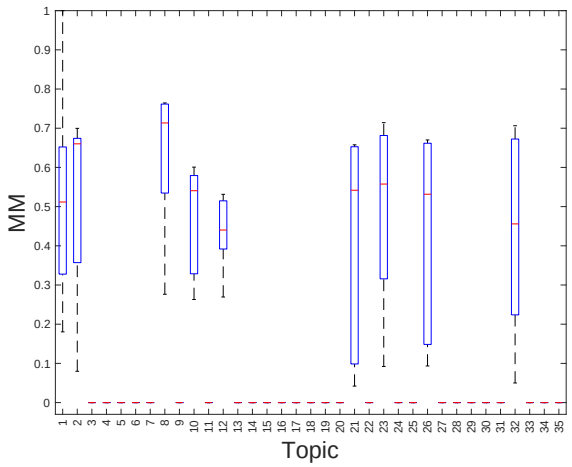
(d) Web Track 2012



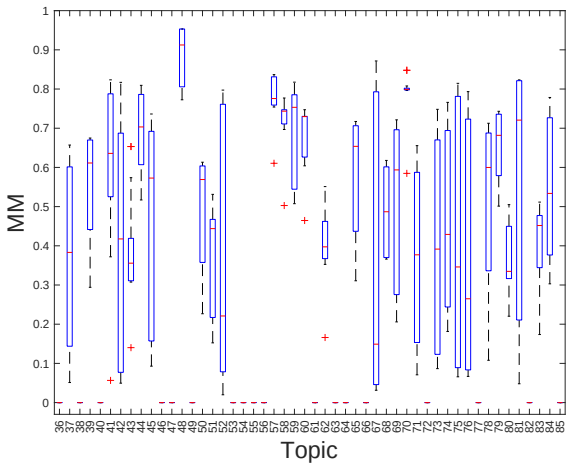
(e) Web Track 2013



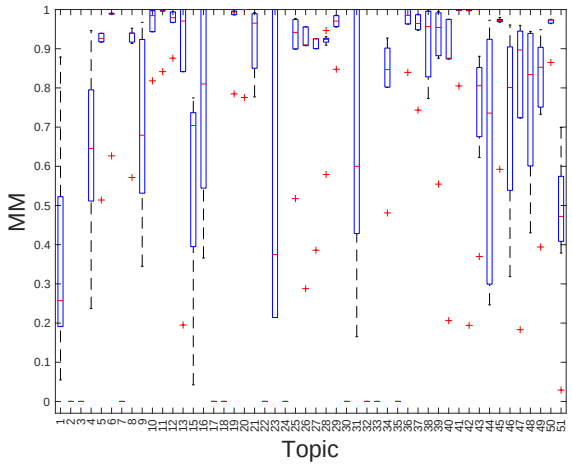
(f) Web Track 2014



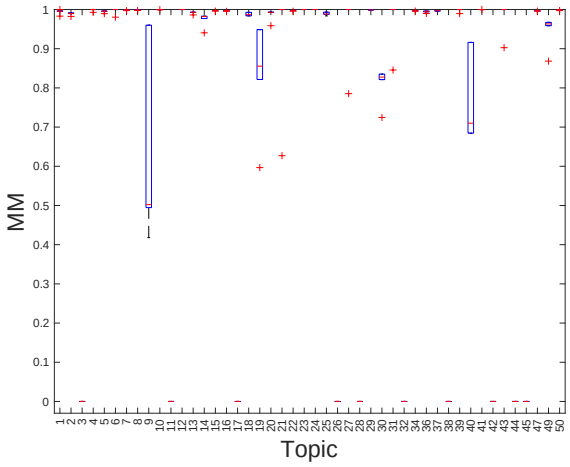
(g) Task Track 2015



(h) Task Track 2016

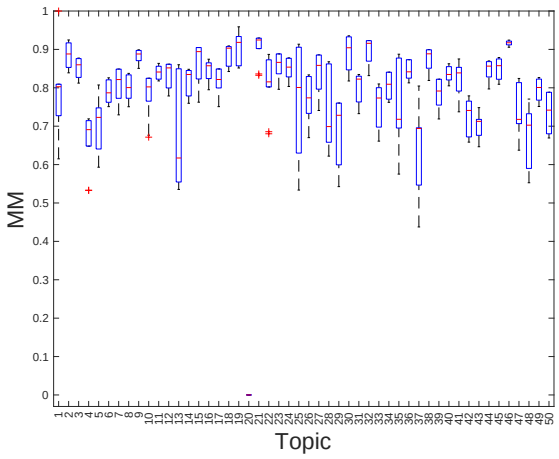


(i) Decision Track 2019

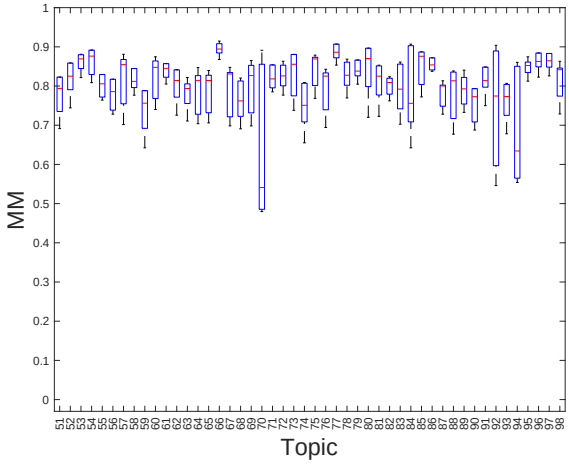


(j) Misinformation Track 2020

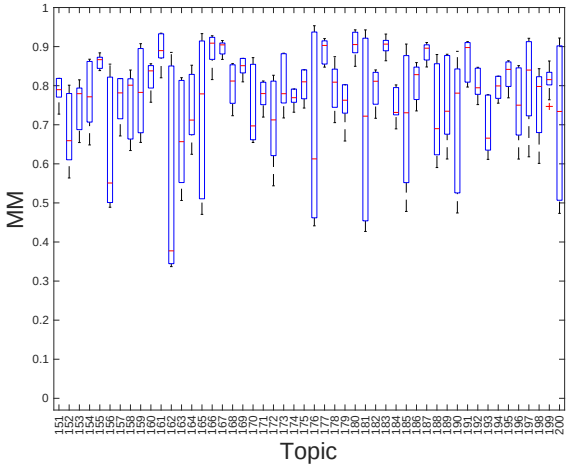
Figure 3: Box-plots for MM instantiated with AP. The x-axis reports the query number, the y-axis reports MM scores. The maximum achievable value for MM depends on the query, the aspects, and the dataset.



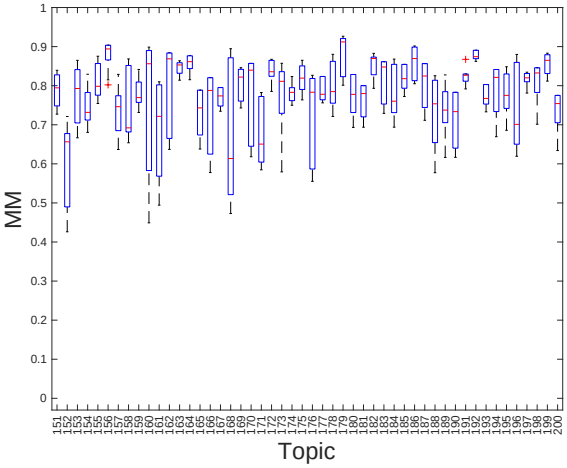
(a) Web Track 2009



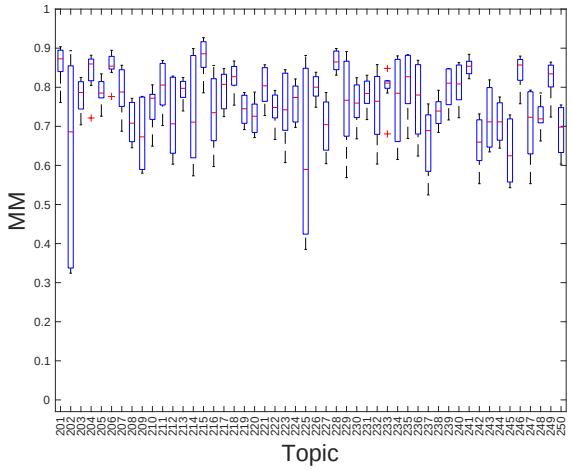
(b) Web Track 2010



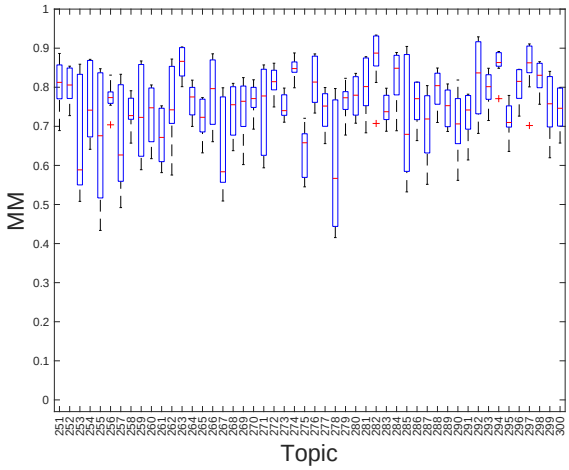
(c) Web Track 2011



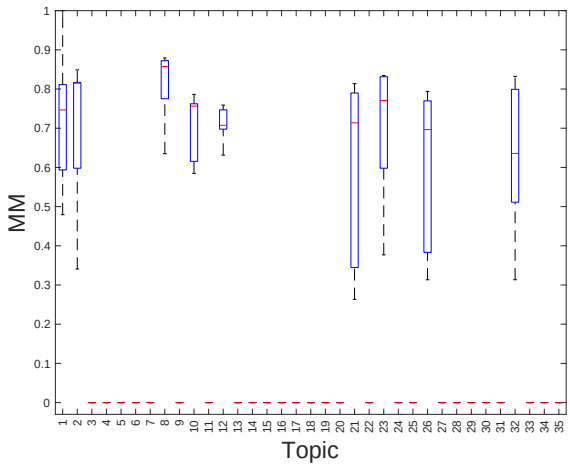
(d) Web Track 2012



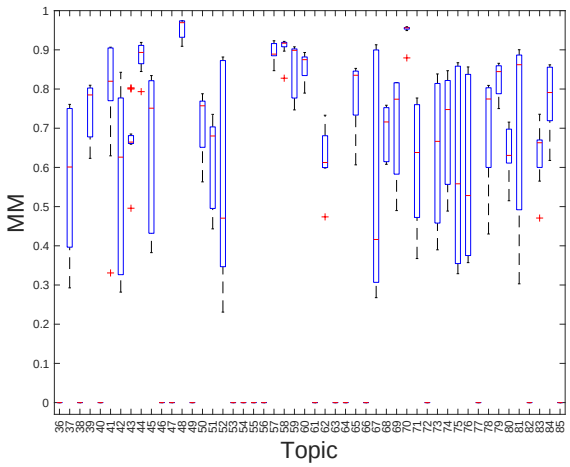
(e) Web Track 2013



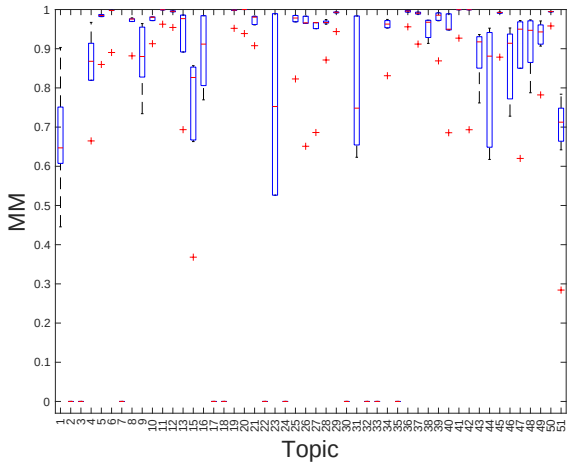
(f) Web Track 2014



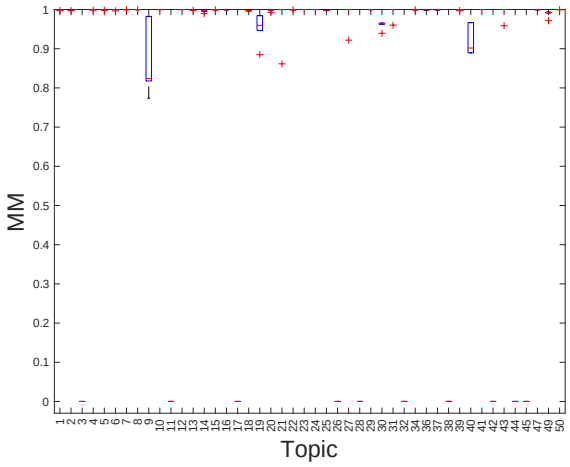
(g) Task Track 2015



(h) Task Track 2016



(i) Decision Track 2019



(j) Misinformation Track 2020

Figure 4: Box-plots for MM instantiated with NDCG. The x-axis reports the query number, the y-axis reports MM scores. The maximum achievable value for MM depends on the query, the aspects, and the dataset.

REFERENCES

[1] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[2] C. Lioma, J. G. Simonsen, and B. Larsen. 2017. Evaluation Measures for Relevance and Credibility in Ranked Lists. In *Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017)*, J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz (Eds.). ACM Press, New York, USA, 91–98.

[3] M. Maistro, L. C. Lima, J. G. Simonsen, and C. Lioma. 2021. Principled Multi-Aspect Evaluation of Rankings. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021)*, G. Demartini, G. Zuccon, C. Culpepper, Z. Huang, and H. Tong (Eds.). ACM Press, New York, USA.

[4] J. Palotti, G. Zuccon, and A. Hanbury. 2018. MM: A New Framework for Multidimensional Evaluation of Search Engines. In *Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018)*, A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Broder, M. J. Zaki, S. Candan, A. Labrinidis, A. Schuster, and H. Wang (Eds.). ACM Press, New York, USA, 1699–1702.