The Impact of Negative Relevance Judgements on nDCG

Lukas Gienapp Maik Fröbe Matthias Hagen Martin Potthast

Leipzig University

Martin-Luther-Universität Halle-Wittenberg

webis.de

29th ACM International Conference on Information and Knowledge Management (CIKM 2020)

Introduction

“nDCG produces scores between 0 and 1.”

(iff gain values are positive)

Introduction

"nDCG produces scores between 0 and 1."

(iff gain values are positive)

Negative gain values (qrels) are prevalent:

- Commonly used at TREC, other venues
- Denote spam, inappropriate documents
- Same amount as "key documents"

TREC Web Track	Qrels Negative
2010	5%
2011	6%
2012	5%
2013	2%
2014	6%

Introduction

"nDCG produces scores between 0 and 1."

(iff gain values are positive)

Negative gain values (qrels) are prevalent:

- Commonly used at TREC, other venues
- Denote spam, inappropriate documents
- Same amount as "key documents"

TREC Web Track	Qrels Negative
2010	5%
2011	6%
2012	5%
2013	2%
2014	6%

Boundedness is necessary:

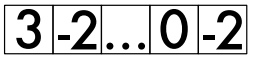
- Ensures nDCG's statistical properties
  nDCG is convergent, top-weighted, realizable, monotonous, localized, complete, scale invariant

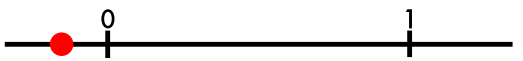
➜ How to handle negative gain values?

Strategies

(1) Original nDCG

- Use orig. nDCG formula on neg. gain values
- **Problem:** boundedness not guaranteed

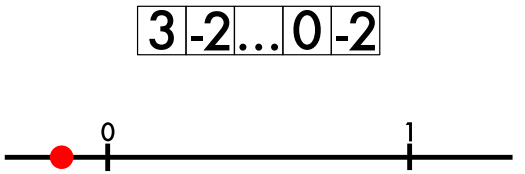




Strategies

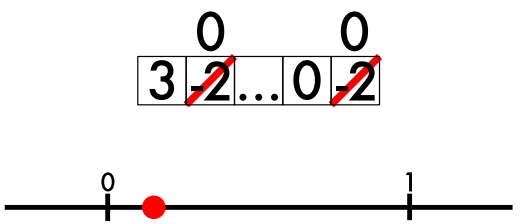
(1) Original nDCG

- Use orig. nDCG formula on neg. gain values
- **Problem:** boundedness not guaranteed



(2) Ignoring negative values

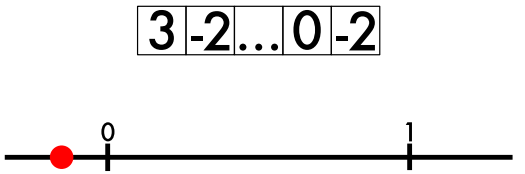
- Negative relevance values are treated as 0
- This is current practice of most eval tools
- **Problem:** loss of information



Strategies

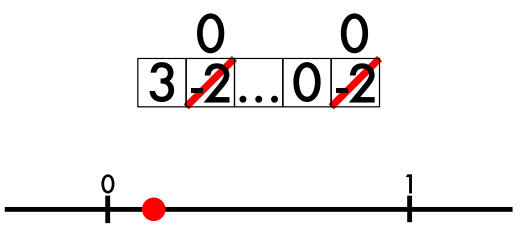
(1) Original nDCG

- Use orig. nDCG formula on neg. gain values
- **Problem:** boundedness not guaranteed



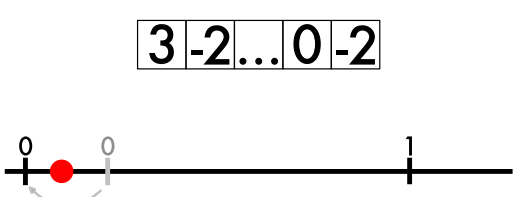
(2) Ignoring negative values

- Negative relevance values are treated as 0
- This is current practice of most eval tools
- **Problem:** loss of information



(3) Min-Max normalization

- Adopt full min-max-normalization by also including worst possible ranking
- **Problem:** unknown properties



Frequency and Impact

TREC Web Tracks 2010–2014

How often is boundedness violated?

- Between **70% and 100%** of topics violate the boundedness property (neg. scores possible) when using Original nDCG
- Between **8% and 68%** of topics may even score below -1

How do the two proposed solutions impact system rankings?

- Ignoring negative labels affects the rankings slightly ($\rho \approx 0.89$)
- Min-Max nearly reproduces rankings given by Original in full ($\rho \approx 0.98$)

Conclusions:

- Unboundedness is a widespread issue and needs to be addressed.
- The current best practice seems unsuitable, as it affects system rankings.

➜ Investigation of reliability, sensitivity, and stability of the three strategies.

Reliability

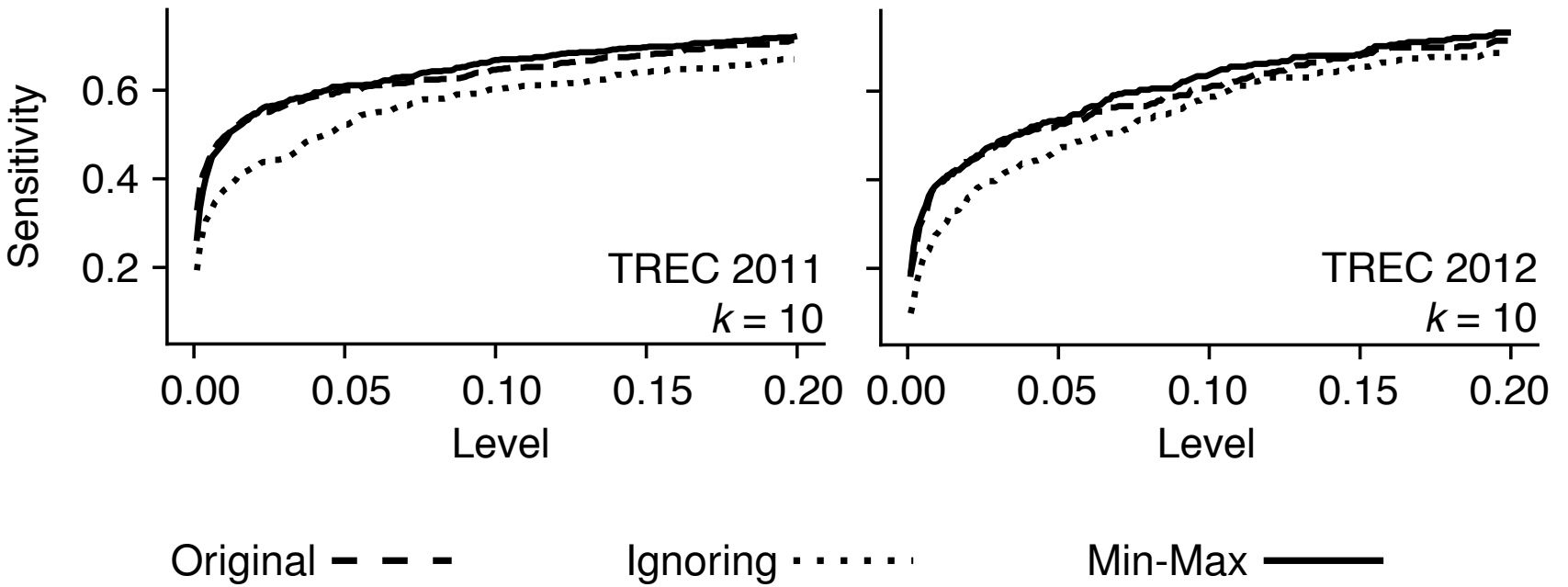
A measure's ability to reflect the actual performance differences of systems.

Strategy	TREC 2011	TREC 2012
Original	0.937	0.930
Ignoring	0.973	0.975
Min-Max	**0.993**	**0.995**

- Min-Max is most reliable, followed by ignoring negative labels, and Original
- Unboundedness increases the measurements' variance for Original

Sensitivity

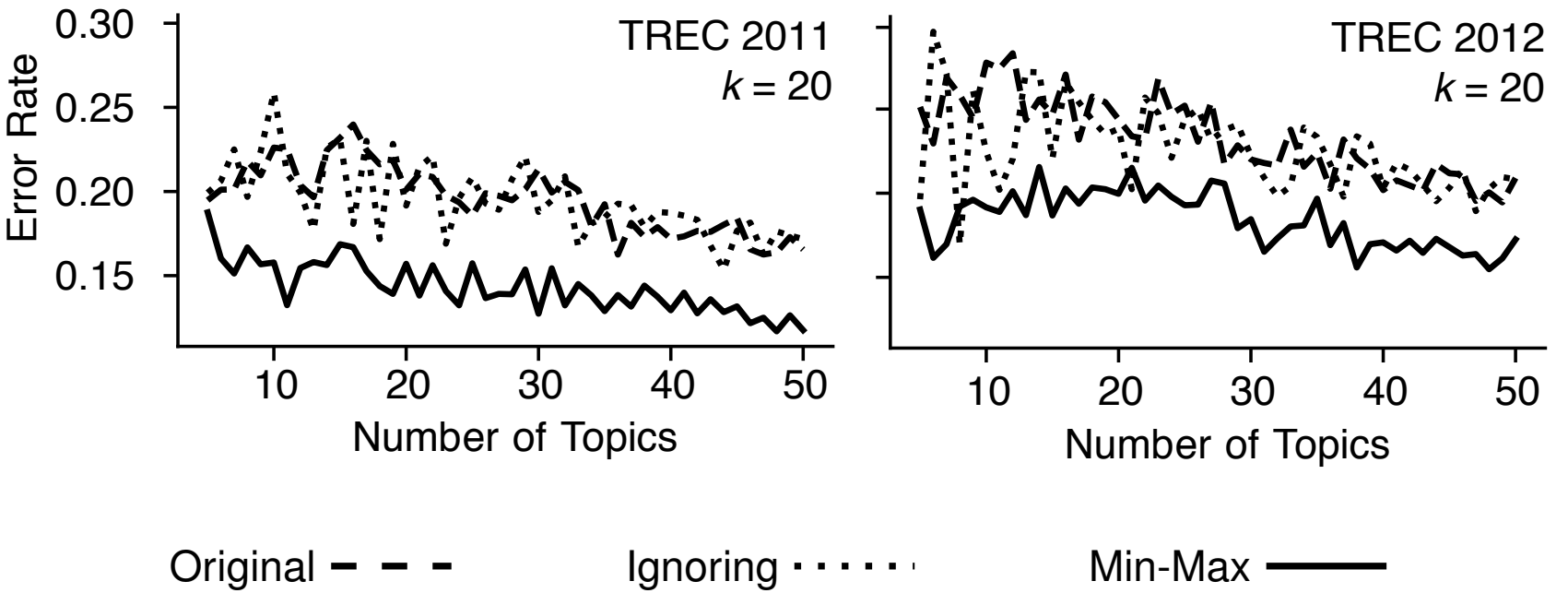
A measure's ability to successfully tell two systems apart at significance level.



- Min-Max performs best, followed by Original
- Ignoring negative values is disfavorable, as it negatively impacts sensitivity.

Stability

A measure's dependence on number of topics.



- Min-Max performs much better, likely due to reduced cross-topic variance.
- Even with more topics, other strategies can't match the improved error rate.

Conclusion

Identified Problem:

- Negative gain values can lead to boundedness violation for nDCG.
- Many evaluation experiments use negative relevance judgments.
- Current strategy is not equipped to adequately address these issues.

Proposed Solution:

- Adopting full min-max normalization.
- Restores boundedness while preserving system rankings.
- Yields additional benefits with increased stability, reliablity, and sensitivity.

Conclusion

Identified Problem:

- Negative gain values can lead to boundedness violation for nDCG.
- Many evaluation experiments use negative relevance judgments.
- Current strategy is not equipped to adequately address these issues.

Proposed Solution:

- Adopting full min-max normalization.
- Restores boundedness while preserving system rankings.
- Yields additional benefits with increased stability, reliablity, and sensitivity.

Thank you!