Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

ECIR 2023, 2–6 April 2023, Dublin, Ireland




Maik Fröbe[1] Lukas Gienapp[2] Martin Potthast[2] Matthias Hagen[1]

[1] FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

[2] UNIVERSITÄT LEIPZIG

webis.de

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Motivation: One of the main goals of TREC

[Voorhees'19]

Evolve the Cranfield paradigm from complete to essentially complete judgments.

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Motivation: One of the main goals of TREC

[Voorhees'19]

Evolve the Cranfield paradigm from complete to essentially complete judgments.

Complete judgments:

- All query document pairs are judged
- Cranfield experiments: 1,400 documents + 225 topics
  [Cleverdon'67]

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Motivation: One of the main goals of TREC

[Voorhees'19]

Evolve the Cranfield paradigm from complete to essentially complete judgments.

Complete judgments:

- All query document pairs are judged
- Cranfield experiments: 1,400 documents + 225 topics
  [Cleverdon'67]

Essentially complete judgments:
[Voorhees'21]

- Pool submitted runs
- Assume unjudged documents are non-relevant

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Motivation: One of the main goals of TREC

[Voorhees'19]

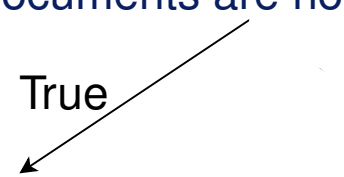
Evolve the Cranfield paradigm from complete to essentially complete judgments.

Complete judgments:

- All query document pairs are judged
- Cranfield experiments: 1,400 documents + 225 topics
  [Cleverdon'67]

Essentially complete judgments:
[Voorhees'21]

- Pool submitted runs
- Assume unjudged documents are non-relevant



TREC collections with robust pools can reliably evaluate modern neural retrieval models
[Voorhees'22]

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Motivation: One of the main goals of TREC

[Voorhees'19]

Evolve the Cranfield paradigm from complete to essentially complete judgments.

Complete judgments:

- All query document pairs are judged
- Cranfield experiments: 1,400 documents + 225 topics
  [Cleverdon'67]

Essentially complete judgments:
[Voorhees'21]

- Pool submitted runs
- Assume unjudged documents are non-relevant

TREC collections with robust pools can reliably evaluate modern neural retrieval models
[Voorhees'22]

ANCE at TREC-COVID [Thakur'21]

Original Qrels		Post-Judgment
Unjudged@10	nDCG@10	nDCG@10
22.4%	0.652	0.735

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Motivation: One of the main goals of TREC

[Voorhees'19]

Evolve the Cranfield paradigm from complete to essentially complete judgments.

Complete judgments:

- All query document pairs are judged
- Cranfield experiments: 1,400 documents + 225 topics
  [Cleverdon'67]

Essentially complete judgments:
[Voorhees'21]

- Pool submitted runs
- Assume unjudged documents are non-relevant

True

TREC collections with robust pools can reliably evaluate modern neural retrieval models
[Voorhees'22]

False

ANCE at TREC-COVID [Thakur'21]

Original Qrels		Post-Judgment
Unjudged@10	nDCG@10	nDCG@10
22.4%	0.652	0.735

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

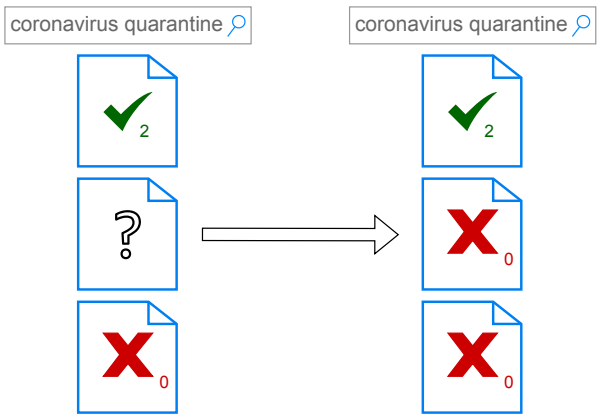
Related Work

- Specialized effectiveness measures
  [Buckley'04, Moffat'08]
- Predict relevance of a document to a query (e.g., based on the content)
  [Aslam'06, Aslam'07, Buttcher07, Carterette'07]
- Stratified Pooling (e.g., infAP, infNDCG)
  [Yilmaz'06, Yilmaz'08, Voorhees'14]
- Bootstrapping in IR
  [Savoy'97, Smucker'07, Sakai'06, Sakai'07, Zobel'20, Cormack'06, Ferro'22]

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

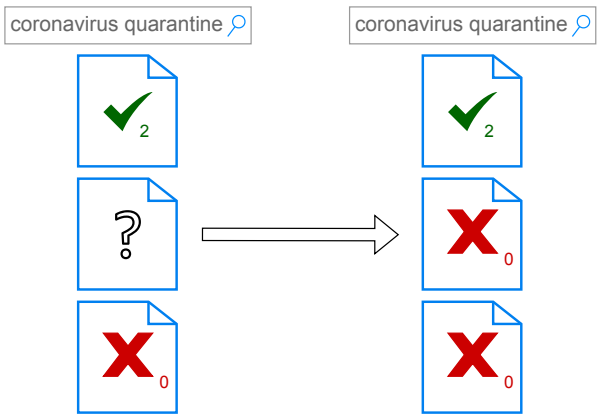
- Standard: Assume Not Relevant



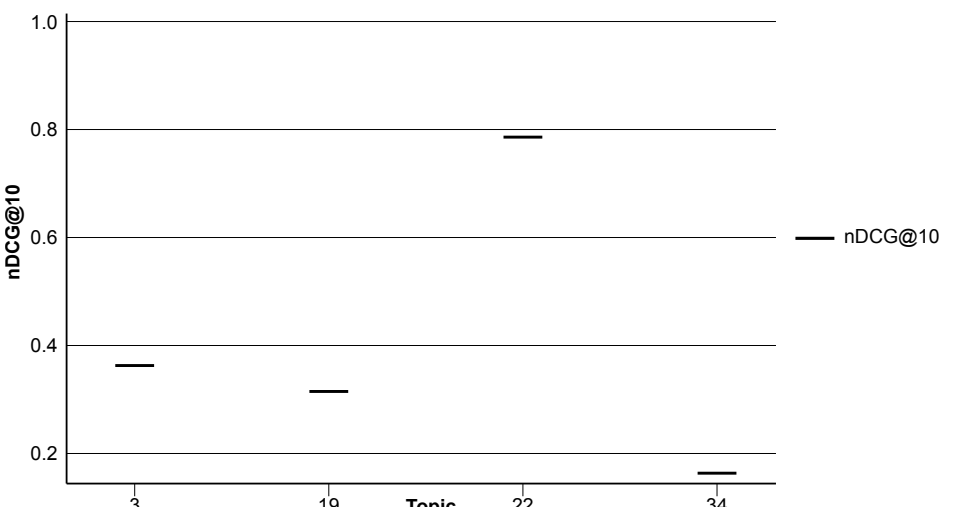
Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

- Standard: Assume Not Relevant



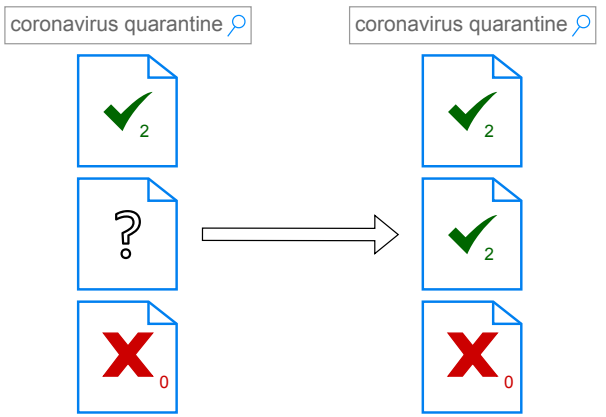
ANCE on TREC-COVID:



Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

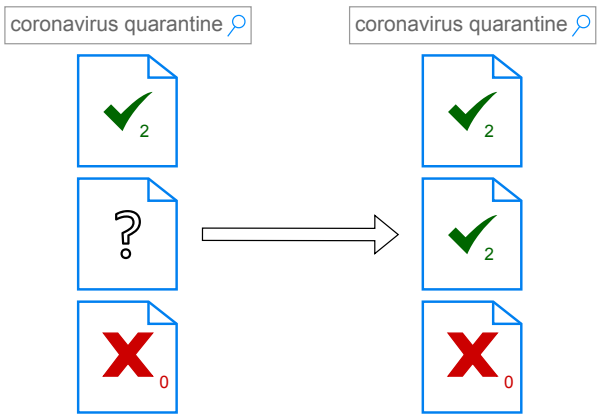
- Upper bounds: Assume unjudged documents are highly relevant
  [Lu'16]



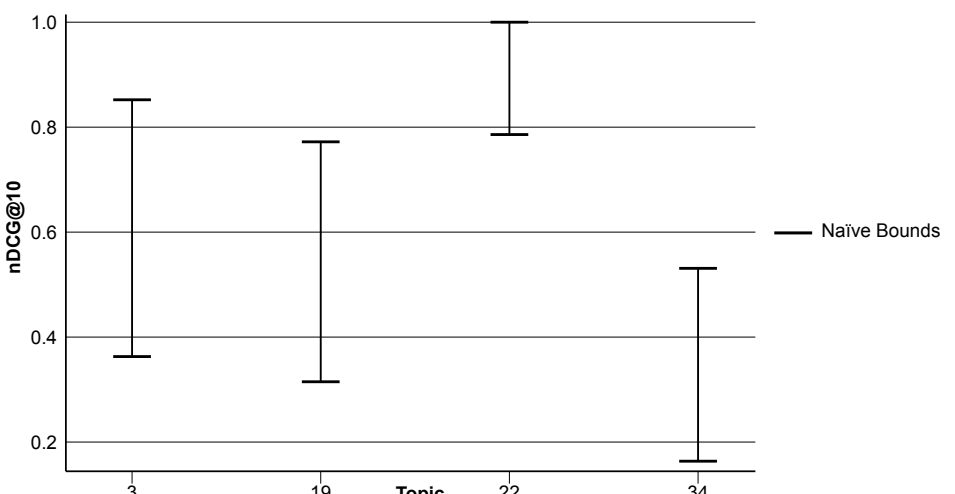
Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

- Upper bounds: Assume unjudged documents are highly relevant
  [Lu'16]



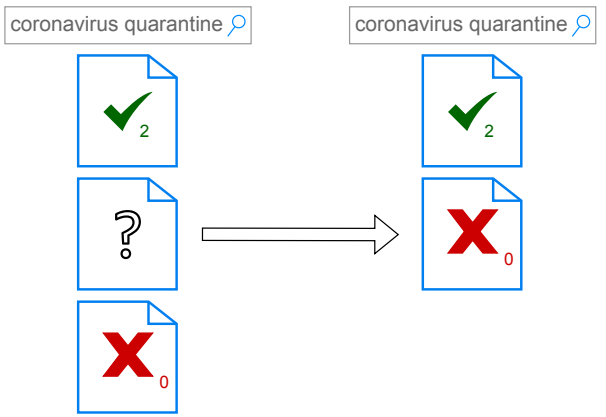
ANCE on TREC-COVID:



Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

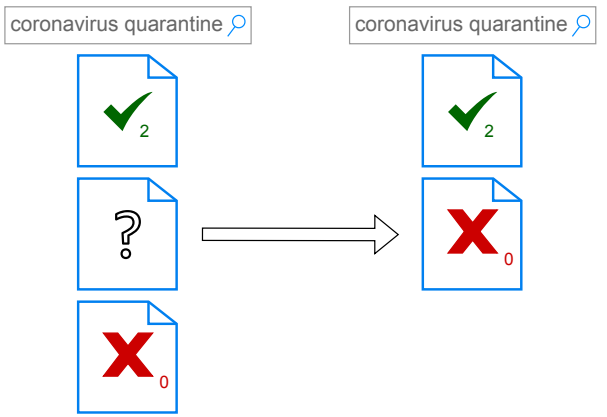
- Condensed lists: Remove unjudged documents
  [Sakai'06]



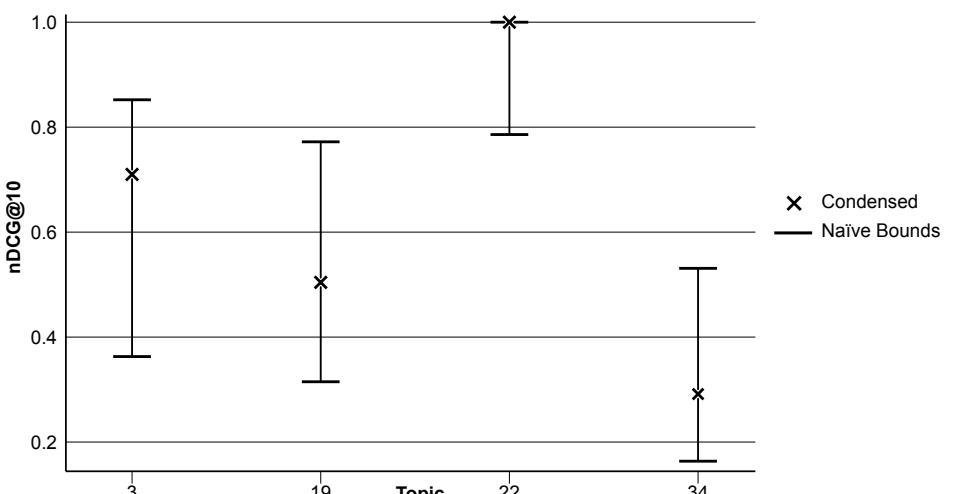
Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

- Condensed lists: Remove unjudged documents
  [Sakai'06]



ANCE on TREC-COVID:



Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Bootstrapping

[Efron'94]

Bootstrapping is a statistical technique in which repeated samples are drawn from data to obtain a distribution for subsequent statistical analyses

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Bootstrapping

[Efron'94]

> Bootstrapping is a statistical technique in which repeated samples are drawn from data to obtain a distribution for subsequent statistical analyses

We use Bootstrapping obtain distribution of possible nDCG scores

- Sample relevance judgment from remaining pool without replacement
  - Preserves iDCG
  - Ensures comparability to runs from the pool

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Bootstrapping

[Efron'94]

Bootstrapping is a statistical technique in which repeated samples are drawn from data to obtain a distribution for subsequent statistical analyses

We use Bootstrapping obtain distribution of possible nDCG scores

- Sample relevance judgment from remaining pool without replacement
  - Preserves iDCG
  - Ensures comparability to runs from the pool

Sampling priors

- Pool-based priors

$$P(\textit{rel} = r \mid J) = \frac{|\{d \in J : rel(d,q) = r\}|}{|J|}$$

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Bootstrapping

[Efron'94]

Bootstrapping is a statistical technique in which repeated samples are drawn from data to obtain a distribution for subsequent statistical analyses

We use Bootstrapping obtain distribution of possible nDCG scores

- Sample relevance judgment from remaining pool without replacement
  - Preserves iDCG
  - Ensures comparability to runs from the pool

Sampling priors

- Pool-based priors

$$P(\textit{rel} = r \mid J) = \frac{|\{d \in J : rel(d,q) = r\}|}{|J|}$$

- Run-based priors

$$P(\textit{rel} = r \mid R) = \frac{|\{d \in R : rel(d,q) = r\}|}{|\{d \in R : d \text{ is judged}\}|}$$

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Bootstrapping

[Efron'94]

Bootstrapping is a statistical technique in which repeated samples are drawn from data to obtain a distribution for subsequent statistical analyses

We use Bootstrapping obtain distribution of possible nDCG scores

- Sample relevance judgment from remaining pool without replacement
  - Preserves iDCG
  - Ensures comparability to runs from the pool

Sampling priors

- Pool-based priors

$$P(\textit{rel} = r \mid J) = \frac{|\{d \in J : rel(d,q) = r\}|}{|J|}$$

- Run-based priors

$$P(\textit{rel} = r \mid R) = \frac{|\{d \in R : rel(d,q) = r\}|}{|\{d \in R : d \text{ is judged}\}|}$$

- Pool+run-based priors

$$P(\textit{rel} = r \mid J, R) = \frac{P(r \mid J) + P(r \mid R)}{2}$$

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Bootstrapping

[Efron'94]

Bootstrapping is a statistical technique in which repeated samples are drawn from data to obtain a distribution for subsequent statistical analyses

We use Bootstrapping obtain distribution of possible nDCG scores

- Sample relevance judgment from remaining pool without replacement
  - Preserves iDCG
  - Ensures comparability to runs from the pool

Sampling priors

- Pool-based priors

$$P(\textit{rel} = r \mid J) = \frac{|\{d \in J : rel(d,q) = r\}|}{|J|}$$

- Run-based priors

$$P(\textit{rel} = r \mid R) = \frac{|\{d \in R : rel(d,q) = r\}|}{|\{d \in R : d \text{ is judged}\}|}$$

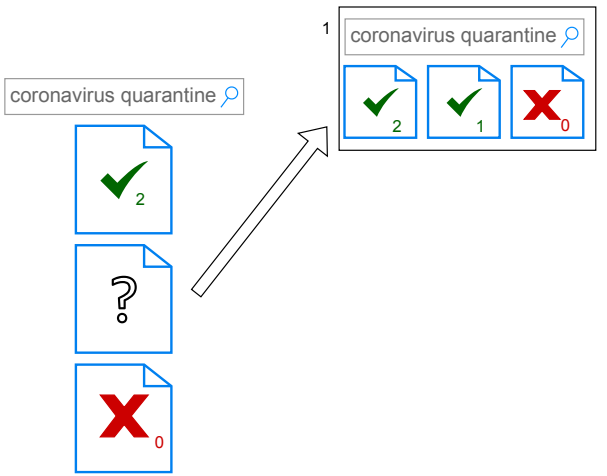
- Pool+run-based priors

$$P(\textit{rel} = r \mid J, R) = \frac{P(r \mid J) + P(r \mid R)}{2}$$

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

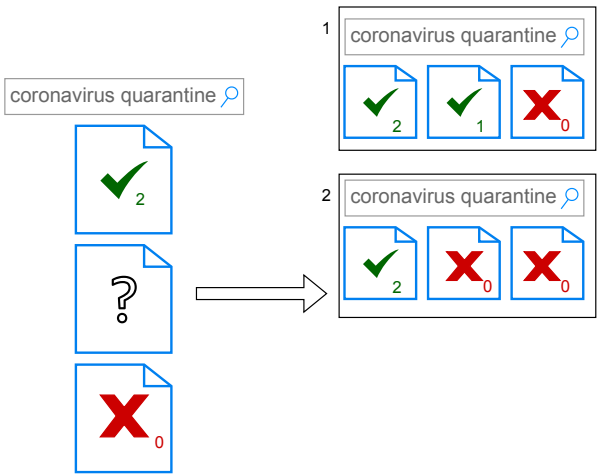
- Bootstrapping



Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

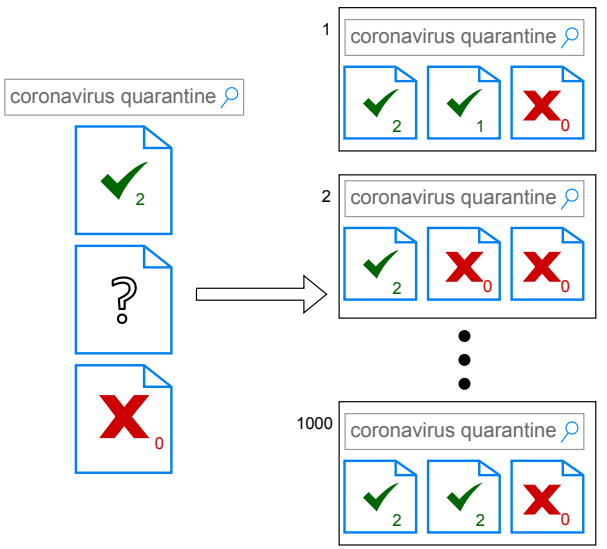
- Bootstrapping



Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

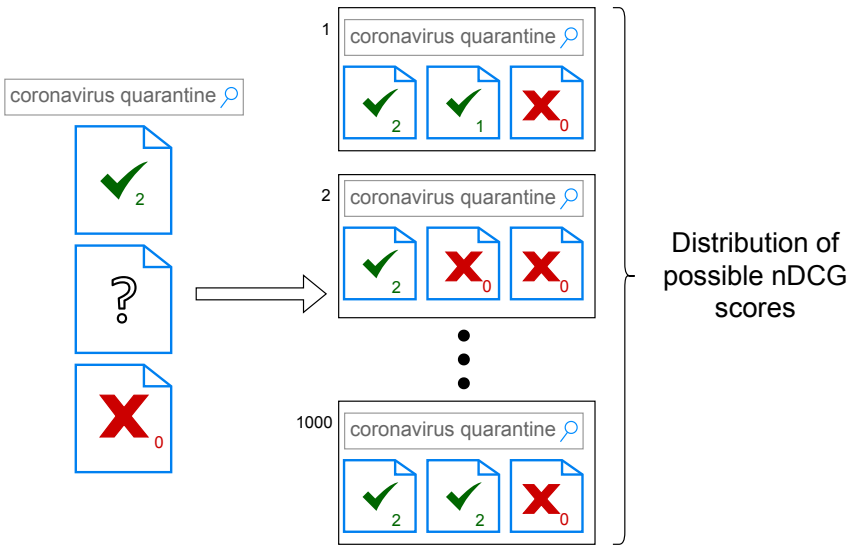
- Bootstrapping



Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

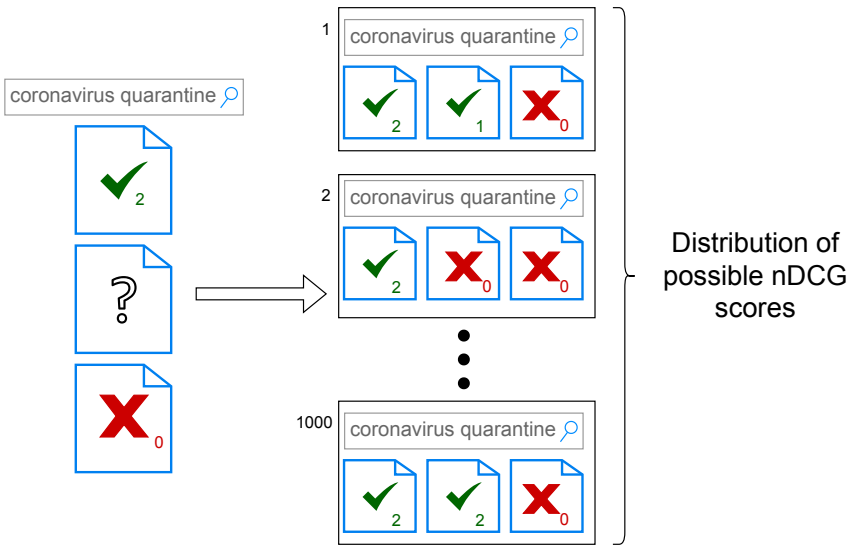
- Bootstrapping



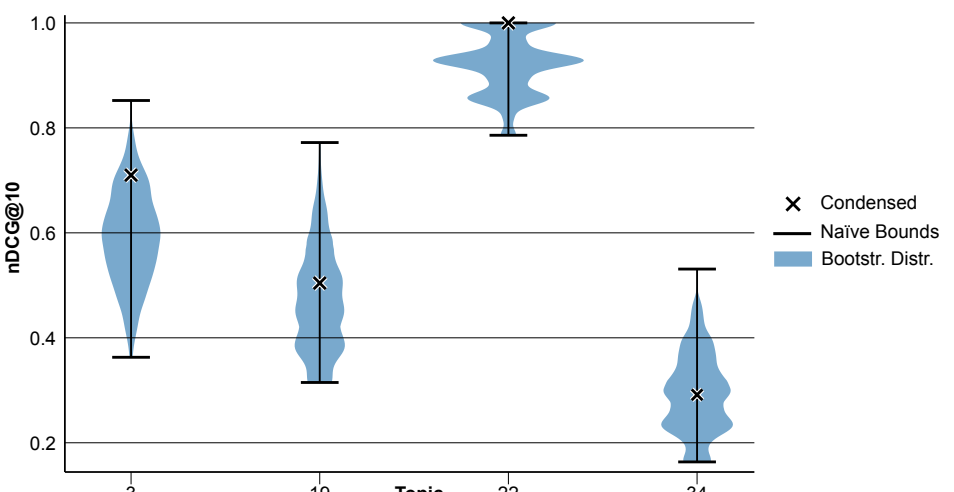
Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

- Bootstrapping



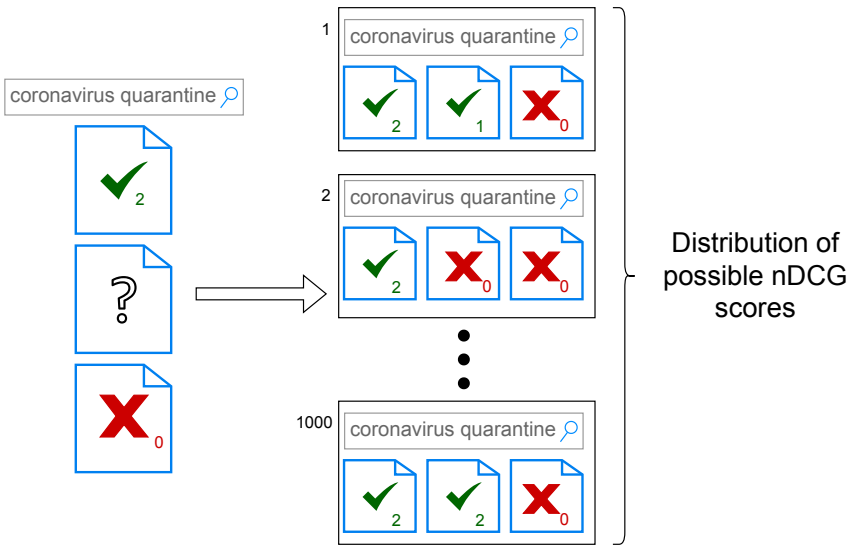
ANCE on TREC-COVID:



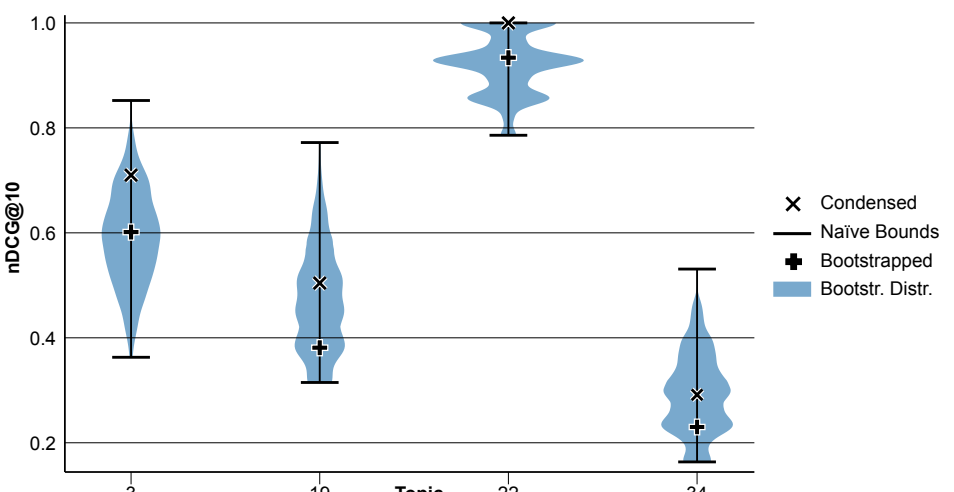
Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

- Bootstrapping



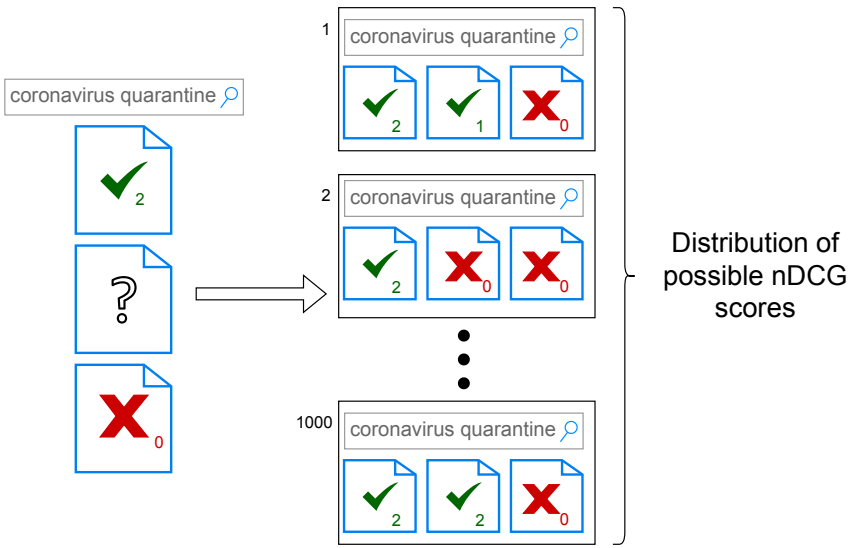
ANCE on TREC-COVID:



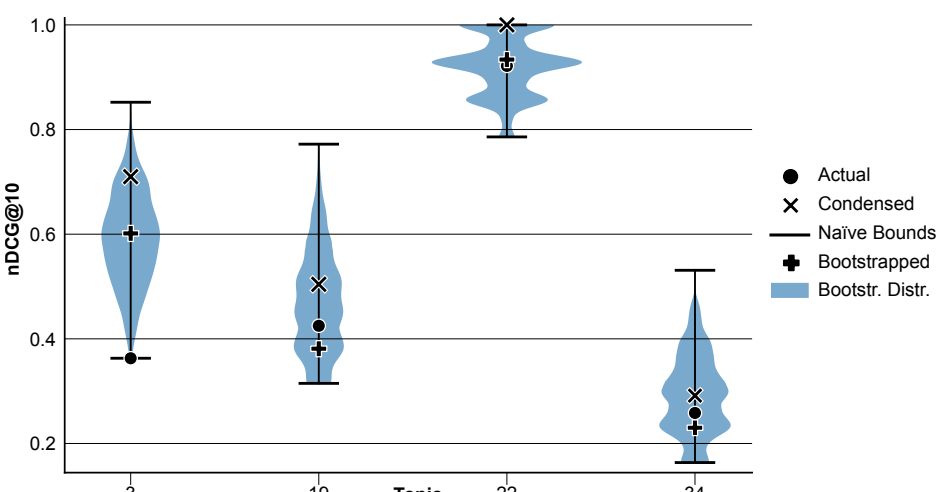
Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Approaches to Handle Unjudged Documents

- Bootstrapping



ANCE on TREC-COVID:



Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Experiments: nDCG score Prediction

Simulated Incompleteness with leave one group out

Benchmarks

	Robust04	ClueWeb09	ClueWeb12
Non-Relevant among Unjudged	96 %	80 %	67 %
Essentially Complete	✓	?	✗

Effectiveness

Approach	**RMSE**		
	Robust04	ClueWeb09	ClueWeb12
Lower Bound	$0.058^{*\ddagger}$	$0.076^{*\ddagger}$	$0.113^{*\ddagger}$
Condensed List	$0.068^{\dagger\ddagger}$	$0.087^{\dagger\ddagger}$	$0.092^{\dagger\ddagger}$
Upper Bound	$0.210^{\dagger *}$	$0.338^{\dagger *}$	$0.307^{\dagger *}$
Bootstrapping	$\mathbf{0.056}^{*\ddagger}$	$\mathbf{0.074}^{*\ddagger}$	$\mathbf{0.083}^{\dagger\ddagger}$

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Experiments: nDCG score Prediction

Simulated Incompleteness with leave one group out

Benchmarks

	Robust04	ClueWeb09	ClueWeb12
Non-Relevant among Unjudged	96 %	80 %	67 %
Essentially Complete	✓	?	✗

Effectiveness

Approach	**RMSE**		
	Robust04	ClueWeb09	ClueWeb12
Lower Bound	$0.058^{*‡}$	$0.076^{*‡}$	$0.113^{*‡}$
Condensed List	$0.068^{†‡}$	$0.087^{†‡}$	$0.092^{†‡}$
Upper Bound	$0.210^{†*}$	$0.338^{†*}$	$0.307^{†*}$
Bootstrapping	$\mathbf{0.056}^{*‡}$	$\mathbf{0.074}^{*‡}$	$\mathbf{0.083}^{†‡}$

Experiments on real incompleteness

TREC-COVID experiments confirm observations on simulated incompleteness

Bootstrapped nDCG Estimation in the Presence of Unjudged Documents

Experiments: Correlation of System Rankings for nDCG

Simulated Incompleteness with leave one group out

Effectiveness

Approach	**Kendall's τ**			
	Robust04	ClueWeb09	ClueWeb12	Mean
Lower Bound	0.936	**0.821**	0.646	0.801
Condensed List	0.924	0.610	0.786	0.773
Upper Bound	0.189	-0.411	-.097	-.106
Bootstrapping	**0.966**	0.716	**0.814**	**0.832**

Conclusion

Bootstrapping allows to account for unjudged documents in post-hoc experiments

- More accurate than condensed lists in all cases
- More accurate than assuming non-relevance for small judgment pools

Implementation available as trectools plugin: github.com/webis-de/ECIR-23

Conclusion

Bootstrapping allows to account for unjudged documents in post-hoc experiments

- More accurate than condensed lists in all cases
- More accurate than assuming non-relevance for small judgment pools

Implementation available as trectools plugin: github.com/webis-de/ECIR-23

Expand our bootstrapping to more evaluation measures

- Q-Measure
- MAP
- RBP
- ...

Combination with predictions based on the document content

- could lead to more informed bootstrap priors

Conclusion

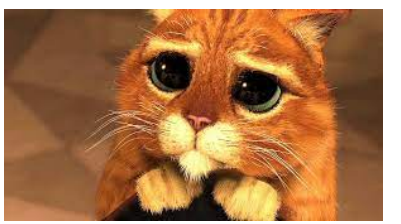
Bootstrapping allows to account for unjudged documents in post-hoc experiments

- More accurate than condensed lists in all cases
- More accurate than assuming non-relevance for small judgment pools

Implementation available as trectools plugin: github.com/webis-de/ECIR-23

Expand our bootstrapping to more evaluation measures

- Q-Measure
- MAP
- RBP
- . . .



Please, no tough questions :)

Combination with predictions based on the document content

- could lead to more informed bootstrap priors

github.com/webis-de/ECIR-23

Conclusion

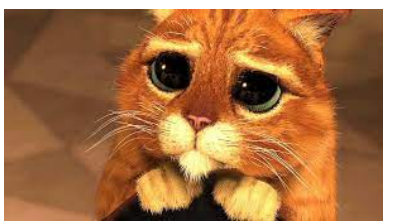
Bootstrapping allows to account for unjudged documents in post-hoc experiments

- More accurate than condensed lists in all cases
- More accurate than assuming non-relevance for small judgment pools

Implementation available as trectools plugin: github.com/webis-de/ECIR-23

Expand our bootstrapping to more evaluation measures

- Q-Measure
- MAP
- RBP
- . . .



Please, no tough questions :)

Combination with predictions based on the document content

- could lead to more informed bootstrap priors

github.com/webis-de/ECIR-23

Thank You!