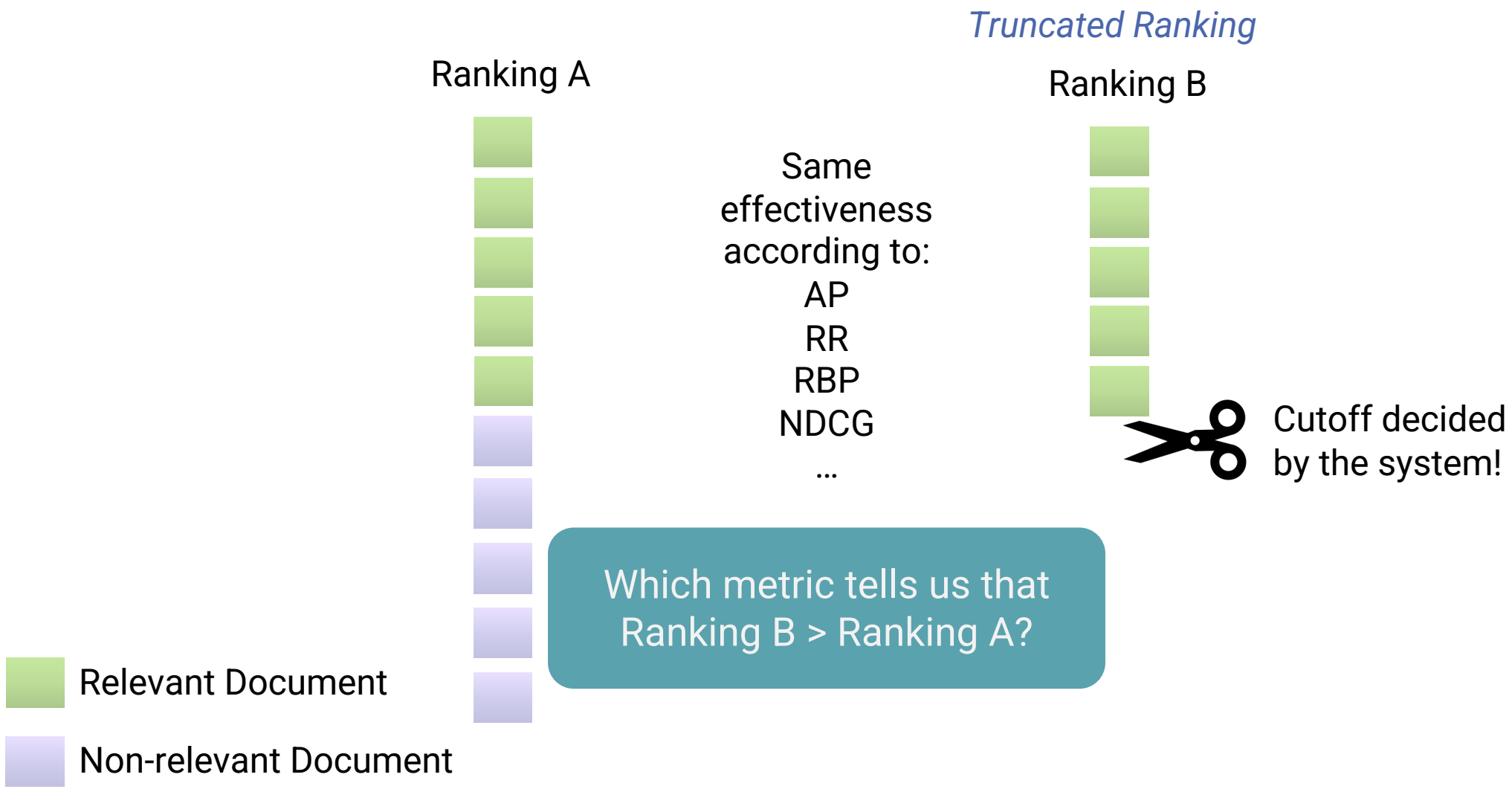
Ranking Interruptus: When Truncated Rankings Are Better and How to Measure That

Enrique Amigó, Stefano Mizzaro, Damiano Spina

UNIVERSITÀ DEGLI STUDI DI UDINE

RMIT UNIVERSITY

Which Ranking is Better?



Why Truncated Rankings?

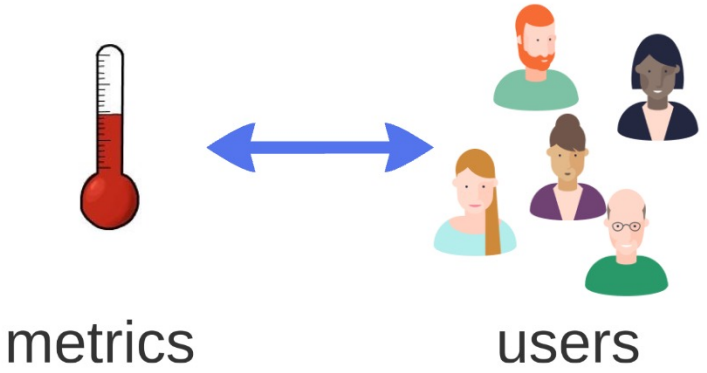
Conversational Search

Small Screens

How to Pick the Right Metric?

”My system is the best one!”

Axiomatic Analysis



System Effectiveness vs. User Satisfaction

Axiomatic Analysis of Effectiveness Metrics

A general evaluation measure for document organization tasks

Full Text: PDF

Authors: Enrique Amigó UNED, Madrid, Spain
Julio Gonzalo UNED, Madrid, Spain
Felisa Verdejo UNED, Madrid, Spain

Published in:

· Proceeding
SIGIR '13 Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval
Pages 643-652

Axiomatic analysis for ad-hoc retrieval

An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric

Enrique Amigó
NLP & IR Group at UNED
Madrid, Spain
enrique@lsi.uned.es

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Jorge Carrillo-de-Albornoz
NLP & IR Group at UNED
Madrid, Spain
jcalbornoz@lsi.uned.es

Extended to diversity metrics + Rank-Biased Utility (RBU)

Information Retrieval Journal (2020) 23:318–386
https://doi.org/10.1007/s10791-020-09374-0

AXIOMATIC THINKING FOR INFORMATION RETRIEVAL

On the nature of information access evaluation metrics: a unifying framework

Enrique Amigó[1] · Stefano Mizzaro[2]

Theoretical validation of the framework

Ranking Interruptus: When Truncated Rankings Are Better and How to Measure That

Enrique Amigó
UNED NLP & IR Group
Madrid, Spain
enrique@lsi.uned.es

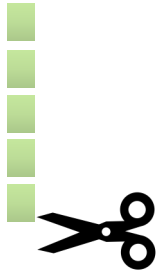
Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

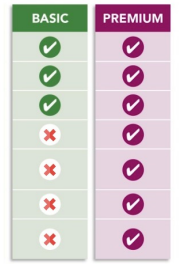
Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Truncated Rankings

Contributions

1. Formal properties for truncated rankings
2. Extension of traditional metrics (ERR, NDCG, and RBP) by adding a user effort factor
3. Theoretical analysis of effectiveness metrics:
   - De-facto standard metrics do not satisfy desirable properties to evaluate truncated rankings
   - Observational Information Effectiveness (OIE) satisfies them all
4. Empirical evidence using 9 TREC test collections
5. Guidelines for metric selection in different ranking scenarios





Desirable Properties

Metrics

	Metric	Non-Truncated: Priority	Non-Truncated: Top-W.	Non-Truncated: Deep. Th.	Non-Truncated: Shall. Th.	Truncated: Confidence	Truncated: Recall	Truncated: Redundancy
FULL RANKING METRICS	Spearman	✓			✓	✓		
	Kendall	✓			✓	✓		
	AUC-ROC	✓			✓	✓		
BINARY RELEVANCE	P@N			✓	✓			
	R@N			✓	✓		✓	
	R-p						✓	
	RR							✓
	F-measure					✓	✓	✓
	AP	✓	✓				✓	
GRADED RELEVANCE	NDCG	✓	✓		✓		✓	
	Q-measure	✓	✓		✓		✓	
	BPref	✓	✓		✓		✓	
PROBABILISTIC USER MODEL BASED	ERR	✓	✓	✓				✓
	RBP	✓	✓	✓	✓			
	iRBU	✓	✓	✓	✓			✓
TERMINAL DOCUMENT BASED	NDCGT	✓	✓		✓	✓	✓	
	ERRT	✓	✓	✓		✓		✓
	RBPT	✓	✓	✓	✓	✓		
UTILITY BASED	Flat Utility	✓			✓	✓		
	RBU	✓	✓	✓	✓	✓		✓
	DCGU	✓	✓		✓	✓		
	ERRU	✓	✓	✓		✓		✓
	RBPU	✓	✓	✓	✓	✓		
INFORMATION BASED	OIE	✓	✓	✓	✓	✓	✓	✓

Designed for non-truncated rankings

Add one ’relevant’ terminal document at the bottom of the ranking

Add a component to take “effort” into account

Based on Information Theory: quantity of information obtained from ranking and ground truth

Empirical Validation

- Nine TREC test collections
- Optimal Truncation Points
- Truncated vs. Full Ranking Effectiveness
- Validation of the aspects captured by truncated rankings

(Some) Results

- Confirmation of theoretical analysis
  - Observational Information Effectiveness (OIE)
  - RBPU: extension of RBP that incorporates effort penalty

- For most metrics, optimal truncation is rather substantial

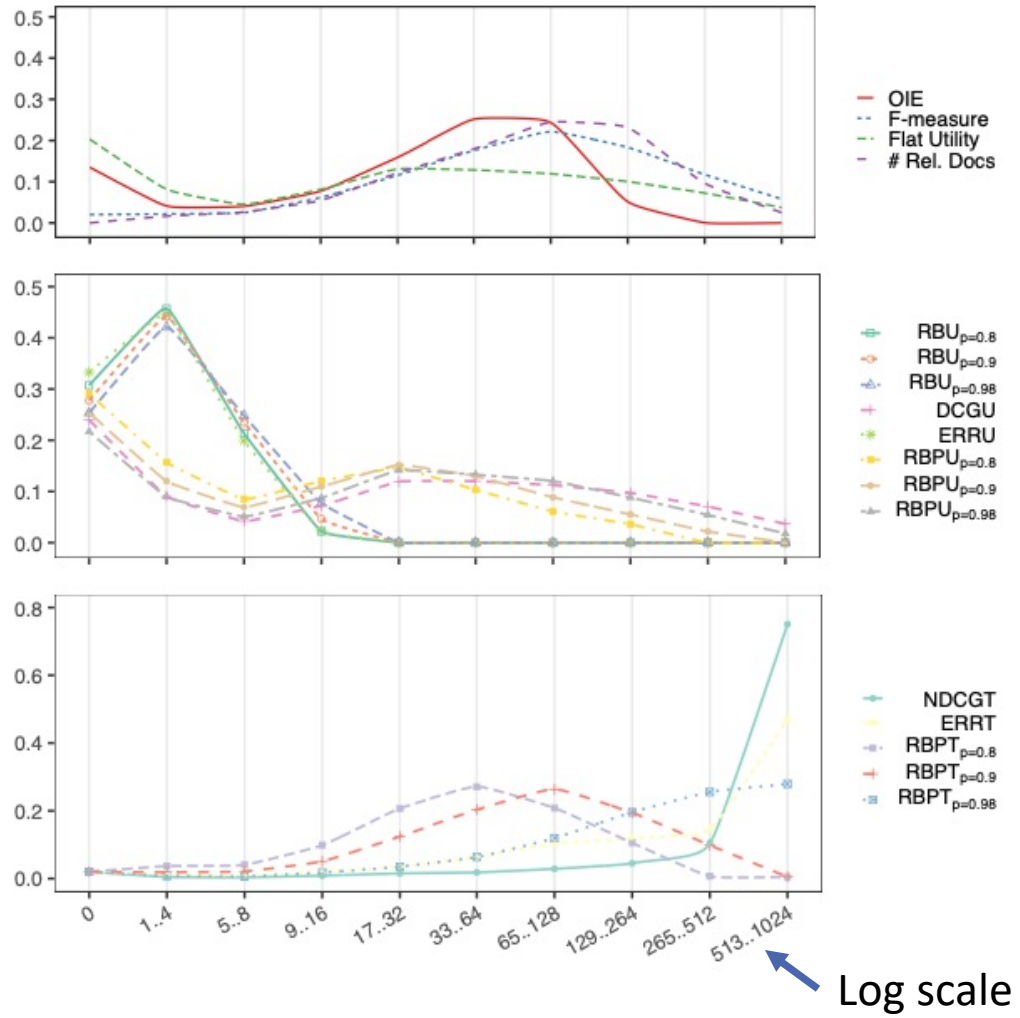
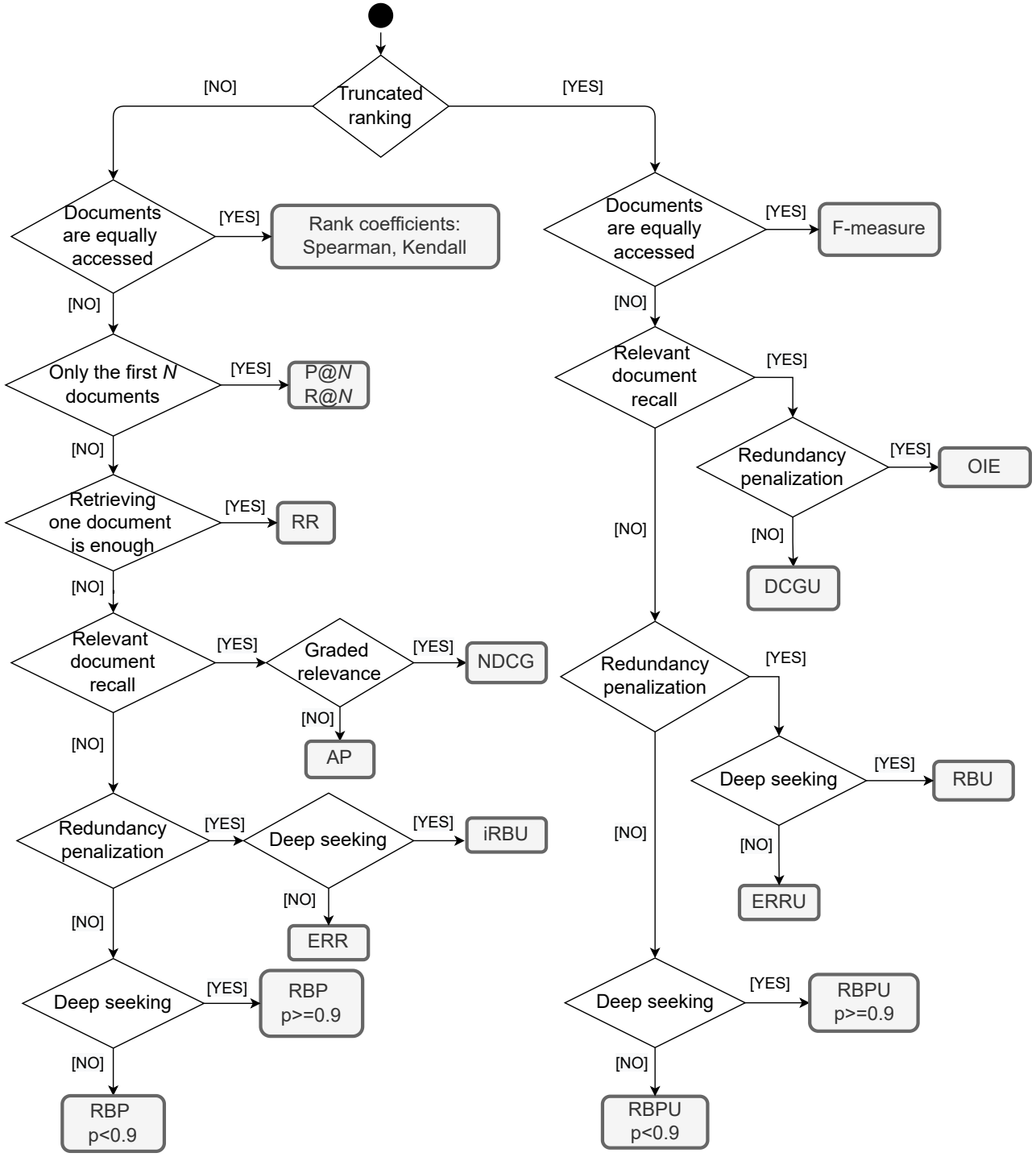


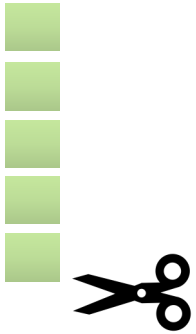
Figure 2: The distributions of optimal ranking lengths according to different truncation metrics. Results are averaged across the nine datasets.

Guidelines for Metric selection

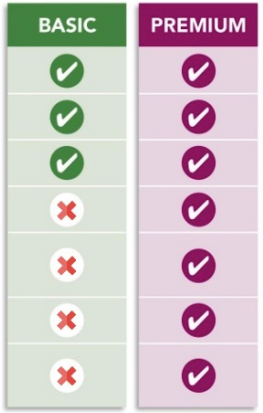
To help practitioners identifying which effectiveness metric should be used for different ranking tasks / conditions



Summary



Evaluation of Truncated Rankings



Axiomatic Analysis +
Empirical Validation

Guidelines to
Choose the
Appropriate Metric

Future work (long road ahead!)

- Experiments with systems which incorporate a stopping criterion
- User studies to calibrate penalty effort
- …