Birmingham 2023 CIKM

Findability: A Novel Measure of Information Accessibility

Aman Sinha, Priyanshu Raj Mall, Dwaipayan Roy

IISER KOLKATA

Motivation

Findability of a document - its capacity to be located *solely* for queries whose intent is satisfied by that particular document.

How to quantify findability of documents in a collection?

Findability

$$f(d) = \frac{1}{|Q_d|} \sum_{q \in Q_d} \xi(p_{dq}, c)$$

- Set of queries generated from d.
- Convenience function.

Generating query set Q_d

A document is considered found when the user is looking for that particular document and it appears in search results for their query.

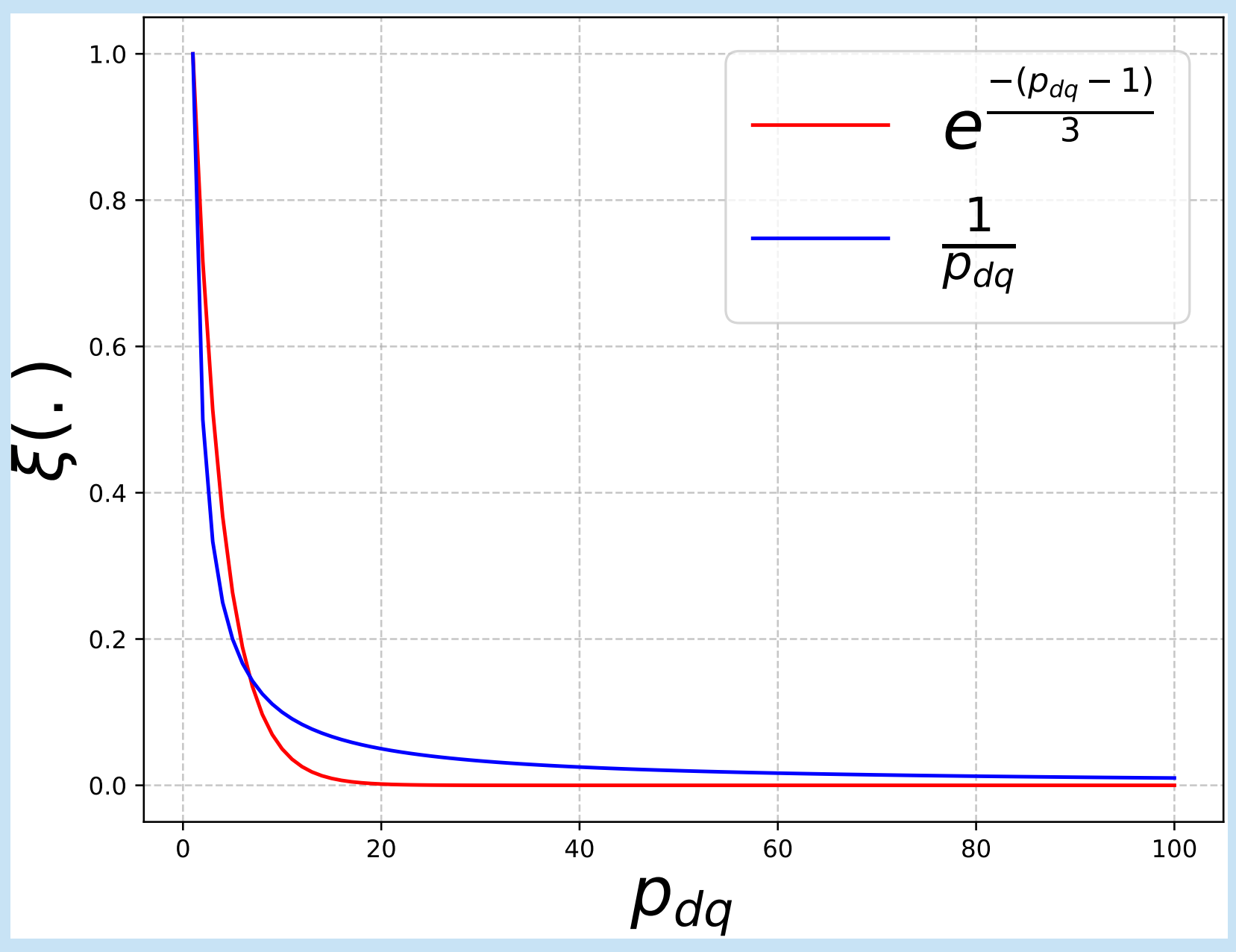
Need a set of known-item search query.

- Applied *Popular+Discrimination strategy* by Azzopardi et al. SIGIR 2007.
- Selects most frequent and discriminative terms from known-item documents.

$\xi(p_{dq}, c)$ - Convenience function

$$\xi(p_{dq}, c) = \begin{cases} 1, & \text{when } p_{dq} = 1. \\ 0, & \text{when } p_{dq} > c. \end{cases}$$

- $\xi(.)$ bounded within the range $[0, 1]$.
- $\xi(.) \approx$ Click Through Rate (CTR).
- CTR on a search engine $\approx$ effort to investigate a rank list.



Exponential decay

$$\xi_e(p_{dq}, c) = \begin{cases} e^{-(p_{dq}-1)/3} & \text{if } p_{dq} \le c \\ 0 & \text{if } p_{dq} > c \end{cases}$$

Inverse law

$$\xi_i(p_{dq}, c) = \begin{cases} \frac{1}{p_{dq}} & \text{if } p_{dq} \le c \\ 0 & \text{if } p_{dq} > c \end{cases}$$

Parameters

- c - maximum rank tolerance of user = 100
- $\xi(.)$ - Inverse law $\xi_i(.)$.
- Retrieval models - LM-Dir, BM25, PL2.

Inequality in Distribution of Findability

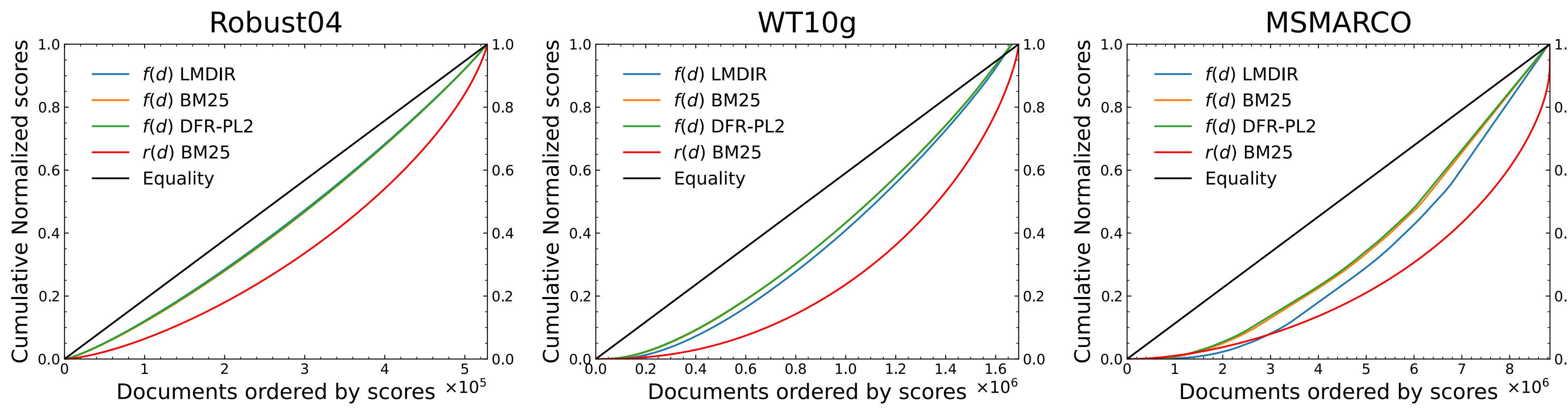
		Robust04	WT10g	MS MARCO
LM-Dir	G	0.1587	0.2847	0.3774
	$\langle f \rangle$	**0.6327**	**0.5209**	**0.5173**
BM25	G	0.1456	0.2503	0.3116
	$\langle f \rangle$	0.6640	0.5985	0.5895
DFR-PL2	G	**0.1424**	**0.2497**	**0.3007**
	$\langle f \rangle$	0.6672	0.6133	0.5888

- Gini coefficient ↑, mean findability ↓.
- Least mean findability → LM-Dir.
- Least Gini coefficient → PL2.
- Collection size ↑ $\implies$ bias ↓.

Datasets

Dataset	Robust	WT10G	MS MARCO
# docs	528,155	1,692,096	8,841,823
Col. type	News	Web	Web excerpts
# terms	1,502,031	9,674,707	1,410,558
# queries	10,230,070	26,041,327	19,839,452

Plots



- Findability distribution remains almost same across all retrieval models.
- LM-Dir yields least findability across all collections.
- Findability bias increases with larger collections.

Retrievability (CIKM 2008)

$$r(d) = \sum_{q \in Q} f(p_{dq}, c)$$

- Set of all possible queries.
- $= 1$ if $p_{dq} \le c$
 $= 0$ otherwise.

Correlation with Retrievability

Query set	Retrievability		Known-item	
	r	τ	r	τ
Robust04	-0.0944	-0.0518	-0.1292	-0.1053
WT10g	-0.0088	0.0084	-0.0256	-0.0287
MS MARCO	0.0115	0.0307	0.0388	0.0269

- Almost negligible association.
- Findability provides a uniform interpretation with a constant range.

Conclusions and Future Work

- Proposed findability, a novel measure for information accessibility.
- Findability provides a uniform interpretation with a constant range.
- To investigate use of findability in fine-tuning retrieval parameters.

Acknowledgement

- The authors would like to thank the anonymous first reviewer without whom this poster would not have been here!
- The travel to CIKM 2023 is supported by IISER-K faculty ARF fund.