FIRE 2023
Forum for Information Retrieval Evaluation

December 15–18, 2023, Panjim, India

A Comparative Analysis of Retrievability and PageRank Measures

Aman Sinha, Priyanshu Raj Mall, Dwaipayan Roy, Kripabandhu Ghosh

Overview

- Introduction
- Brief summary of Pagerank and Retrievability
- Measures used to quantify associated bias
- Experiment and Dataset
- Results
- Conclusion

Introduction

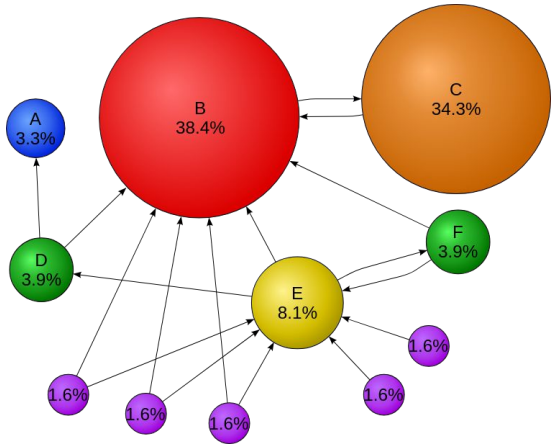
- Accessibility - ease of locating content within collection
- Critical role of document accessibility in IR.
- Two primary avenues:
  - Retrieval models (retrievability scores) and
  - Navigability metrics (PageRank, Hub, Authority).
- Retrievability scores focus on efficient document retrieval within a collection.
- Navigability metrics emphasize the web of connections between documents.

PageRank - A measure of Importance

- PageRank is a link analysis algorithm.
- Assigns a numerical weighting to each page based on its importance within a set of hyperlinked documents.
- Considers links as 'votes' by all other pages on the Web.
- Used as a ranking criterion in search engines, with higher PageRank values indicating greater importance.

PageRank - A measure of importance

$$PR(A) = (1 - \lambda) + \lambda \cdot \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)}$$



- PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.
- The underlying assumption is that more important websites are likely to receive more links from other websites.

Retrievability - A Measure of Accessibility

- Retrievability measures the ease of retrieving a document in an IR system.
- Introduced by Azzopardi and Vinay, quantified through the retrievability score r(d).
- Retrievability score is computed using formula

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

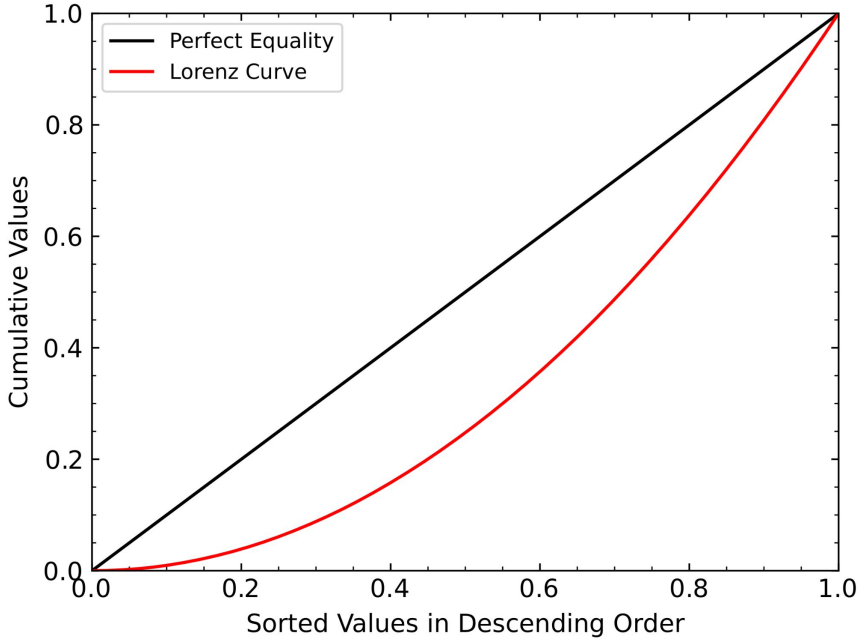
Retrievability - A Measure of Accessibility

- Retrieval relies on a set of queries Q with opportunity weights o_q.
- Query set theoretically encompasses all conceivable queries answerable by the collection D.
- k_dq denotes the retrieval rank of document d for a particular query q.
- Utility function f(k_dq, c) indicates the retrievability of document d within a specified rank cutoff c.

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

Lorenz Curve

- Cumulative score distribution of documents sorted by their retrievability scores in ascending order
- analyze the degree of inequality or bias within the retrieval system.



Linear Lorenz Curve:
- Retrievability scores evenly distributed.
- Indicates a balanced system without significant bias.

Skewed Curve:
- Uneven distribution of retrievability scores.
- Signifies a higher level of inequality or bias within the system.

Gini coefficient - amount of biasness

- To summarize the amount of bias in the Lorenz Curve, the Gini coefficient G is commonly employed

$$G = \frac{\sum_{i=1}^{N} (2i - N - 1) \cdot r(d_i)}{N \sum_{j=1}^{N} r(d_j)}$$

- Expressed as a number between 0 and 1
  - 0 represent perfect equality
  - 1 represents perfect inequality

N represents total number of documents in the collection

"Are there correlations between the PageRank scores and Retrievability scores of the collection documents? Is there a rank correlation between these two metrics?"

Datasets

Dataset	# documents	Collection Type
WT10G	1,692,096	Web
Wikipedia	6,584,626	Wiki

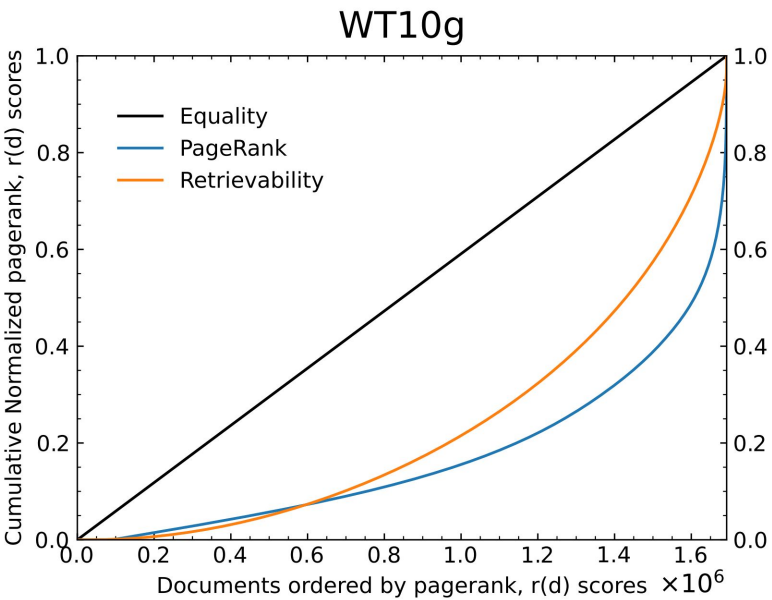
English Wikipedia article dump from February 2023

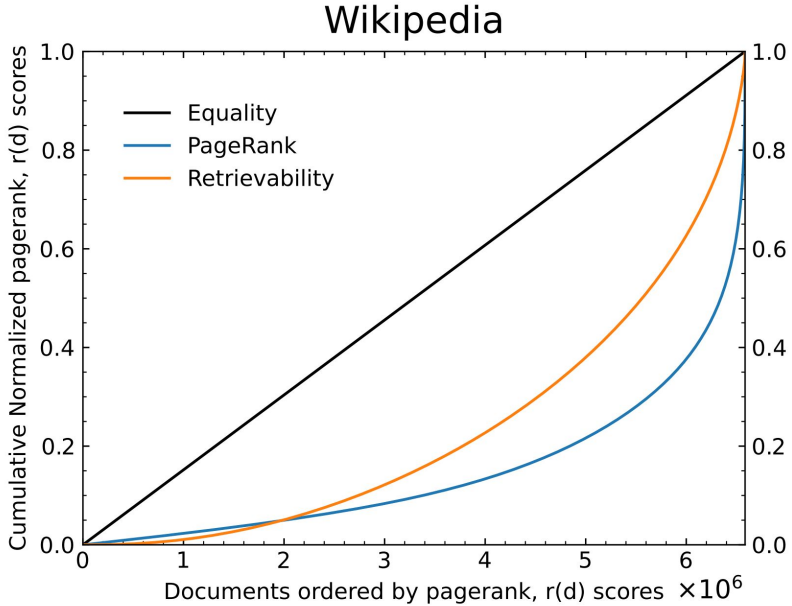
WT10g: TREC Web Corpus

1. https://dumps.wikimedia.org/enwiki

Lorenz curve

Examining the distribution disparities within the retrievability scores and PageRank





Experiment Results

Gini Coefficient quantifies the disparities between retrievability and pagerank scores

	Gini Coefficient	
	Retrievability	**PageRank**
WT10g	0.5371	0.6618
Wikipedia	0.5380	0.7050

	Kendall's τ	**Spearman's ρ**	**RBO**
WT10g	0.0487	0.0730	0.5173
Wikipedia	0.1532	0.2247	0.5633

A weak correlation exists between Ranks of documents based on scores of PageRank and Retrievability

Conclusion

- Document Accessibility:
  - Retrieval models and navigation assess document accessibility.
  - Retrievability scores measure retrieval ease; navigation metrics indicate discoverability via document links.
- Retrievability vs. PageRank:
  - Study on WT10g and Wikipedia datasets.
  - WT10g shows minimal correlation; Wikipedia demonstrates better agreement due to size, interconnectivity and diversity.
- Correlation Analysis:
  - Used Kendall's and Spearman's coefficients.
  - More correlation in the larger, linked Wikipedia dataset.
- Ranked Biased Overlap Measurements
  - Top ranked documents are substantially similar for both datasets.

References

Leif Azzopardi and Richard Bache. 2010. On the relationship between effectiveness and accessibility. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 889–890.

Colin Wilkie and Leif Azzopardi. 2013. An initial investigation on the relationship between usage and findability. In Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35. Springer, 808–811.

Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Comput. Networks 30, 1-7 (1998), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Corrado Gini. 1936. On the measure of concentration with special reference to income and statistics. Colorado College Publication, General Series 208, 1 (1936), 73–79.

Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (Napa Valley, California, USA) (CIKM '08). Association for Computing Machinery, New York, NY, USA, 561–570. https://doi.org/10.1145/1458082.1458157

Thank You!