

A Comparative Analysis of Retrievability and PageRank Measures

Aman Sinha, Priyanshu Raj Mall, Dwaipayan Roy
 Indian Institute of Science Education and Research, Kolkata, India
 as18ms065@iiserkol.ac.in, prm18ms118@iiserkol.ac.in, dwaipayan.roy@iiserkol.ac.in

ABSTRACT

The accessibility of documents within a collection holds a pivotal role in Information Retrieval, signifying the ease of locating specific content in a collection of documents. This accessibility can be achieved via two distinct avenues. The first is through some retrieval model using a keyword or other feature-based search, and the other is where a document can be navigated using links associated with them, if available. Metrics such as PageRank, Hub, and Authority illuminate the pathways through which documents can be discovered within the network of content while the concept of Retrievability is used to quantify the ease with which a document can be found by a retrieval model. In this paper, we compare these two perspectives, PageRank and retrievability, as they quantify the importance and discoverability of content in a corpus. Through empirical experimentation on benchmark datasets, we demonstrate a subtle similarity between retrievability and PageRank particularly distinguishable for larger datasets.

CCS CONCEPTS

• **Information systems** → **Retrieval effectiveness.**

KEYWORDS

Accessibility, Retrievability, PageRank, Bias, Analysis, Empirical Study

ACM Reference Format:

Aman Sinha, Priyanshu Raj Mall, Dwaipayan Roy. 2023. A Comparative Analysis of Retrievability and PageRank Measures. In *Forum for Information Retrieval Evaluation (FIRE 2023)*, December 15–18, 2023, Panjim, India. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3632754.3632760>

1 INTRODUCTION

Accessibility of documents within a collection serves as a critical facet of information retrieval, specifying the ease with which documents can be located amidst an extensive corpus. Essentially, there exist two primary avenues through which this accessibility is assessed. The first path navigates the terrain of retrieval models, ushering us into the realm of retrievability scores [4]. Here, the primary concern is to ascertain whether a particular document can be retrieved within a rank cut off from the vast expanse of a collection by some retrieval model. Informally, the retrievability scores quantify how efficiently a document can be retrieved within

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FIRE 2023, December 15–18, 2023, Panjim, India

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
 ACM ISBN 979-8-4007-1632-4/23/12...\$15.00
<https://doi.org/10.1145/3632754.3632760>

the collection. It essentially answers the question: how quickly can you pinpoint the proverbial needle in the haystack of documents? On a parallel course, the accessibility of a document can also be viewed from the point of view of navigability. In this context, the focus is not directed at individual documents but rather towards the intricate web of connections between documents and relationships that interlink them. Navigability, characterized by metrics such as PageRank [9], Hub, and Authority [15], illuminates the pathways through which documents can be found. In this scenario, the focus is not merely on retrieval but on traversing the internal network of documents. Navigability metrics, such as PageRank, emphasize not just the inherent content of a document, but also its position and importance within the broader context of the document network. This metric, distinct from retrievability scores, offers insights into how discoverable a document is through journeys across links and connections.

Very few works have been done in the field to compare retrievability and PageRank. To the best of our knowledge, the only systematic study was done in [23] where only 2K documents are considered for the study in a closed set of webpages from a university website. Considering both retrievability and PageRank are designed to quantify the discoverability or accessibility (in terms of importance) of contents in a corpus of documents, in this paper, we investigate their alignment through a comparative analysis.

The rest of the paper is organized as follows. We present the related work in the next section highlighting the concept of retrievability and PageRank together with some of their applications in the domain of information retrieval before representing the motivation for this work. We report the empirical results on two benchmark datasets in Section 3 accompanied by a comprehensive analysis of the results. The paper is concluded in Section 4 mentioning the overall finding and mentioning some future work.

2 BACKGROUND AND RELATED WORK

2.1 PageRank - a measure of importance

PageRank is a link analysis algorithm developed by Brin and Page [9]. Given a set of hyperlinked documents (such as the World Wide Web), the algorithm assigns a numerical weighting to each page of the set. Based on this weight, the relative importance of the pages is measured within the set. Informally, PageRank considers links to be like ‘votes’ by all the other pages on the Web, about how important a page is. A link to a page counts as a vote of support. In addition, it considers that some votes are more important than others. When utilized as a ranking criterion (such as in Google), documents with greater PageRank values are ranked higher in the ranked list.

Formally, the PageRank algorithm is presented in Equation 1.

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

- $PR(T_i)$: the self-importance of the webpage T_i ;
- $C(T_i)$: the number of outgoing links from webpage T_i ;
- $\frac{PR(T_i)}{C(T_i)}$: if our page (page A) has a backlink from page i , the share of the vote webpage A will get;
- d : the damping factor in PageRank helps balance the influence of following links on the current page with the randomness of jumping to other pages, making the PageRank algorithm more realistic and reflective of how web users navigate the internet; traditionally it is set to 0.85.

PageRank does not consider the content or size of a document, the language of the document, or the surrounding text used as the anchor to a link. It only captures the authoritative feature of linked documents which is proven useful for different tasks from text matching [18] to word sense disambiguation [21] although it was first introduced to rank web pages in the Google search engine. Further, researchers have used it diverse sub-field of research to improve various downstream tasks. PageRank has been used as a factor in ranking in [6]. It is also employed in [18] as a hierarchical noise filtering approach for the long-form text matching problem to filter out noisy information. The authors plug the PageRank algorithm into the Transformer, to identify and filter both sentence and word-level noisy information in the matching process.

In [14], the authors focused on the problem of the deviations in PageRank values caused by restricted crawling. Some further variation of traditional PageRank is proposed in [24] replacing the original transition matrix is replaced with one whose entries are based on the number of a node's N -step neighbours. PageRank has been utilized in [17] to extract and score keywords from text documents based on their co-occurrence and position. It has also been employed for sentiment analysis to extract and rank opinion words and phrases from online reviews in [16]. Gleich shows how PageRank can be applied to any graph or network in any domain, such as bibliometrics, social and information network analysis, and link prediction and recommendation.

A comprehensive survey on the applications of PageRank algorithms in various domains can be found in [10, 19].

2.2 Retrievability - a measure of accessibility

Retrievability, as a metric, gauges the ease with which a document can be retrieved within a specific configuration of an information retrieval (IR) system. The concept of retrievability, formally introduced by Azzopardi and Vinay [1], is quantified through the retrievability score, denoted as $r(d)$, for a document d within a collection D concerning a particular IR system. Mathematically, the retrievability score $r(d)$ for a document d ($d \in D$) within the context of an IR system is computed using the formula depicted in Equation 2.

$$r(d) = \sum_{q \in Q} o_q \cdot f(k_{dq}, c) \quad (2)$$

As illustrated in Equation 2, the computation of a document's retrievability relies on an extensive set of queries denoted as Q . This set theoretically encompasses all conceivable queries that could be answered by the collection D . Each query q is associated with an opportunity weight o_q , which quantifies the likelihood of selecting query q from the query set Q . The retrieval rank of document d for a particular query q is denoted as k_{dq} , and the utility function

$f(k_{dq}, c)$ serves as an indicator of document d 's retrievability within a specified rank cutoff c .

The conventional approach for assessing retrievability relies on a cumulative-based approximation, where the utility function $f(k_{dq}, c)$ is designed to yield a value of 1 if document d is retrieved within the top c documents for query q , and 0 otherwise. This utility function offers a straightforward interpretation of the retrievability score for each document. Essentially, it quantifies how frequently the document appears within the top c rankings of various queries. Documents that fall beyond the top c positions are excluded from consideration, replicating a user's behavior when examining only the first c search results. Consequently, a higher retrievability score indicates that the document is retrieved within the top ranks for a larger number of queries.

In order to examine the retrievability bias present in a collection, we can calculate retrievability scores for each document using equation (2). By utilizing the Lorenz Curve, which represents the cumulative score distribution of documents sorted by their retrievability scores in ascending order, we can analyze the degree of inequality or bias within the retrieval system. If retrievability scores are evenly distributed, the Lorenz Curve will be linear. However, a skewed curve indicates a greater level of inequality or bias. To summarize the amount of bias in the Lorenz Curve, the Gini coefficient G is commonly employed [4, 7, 8], which is computed as shown:

$$G = \frac{\sum_{i=1}^N (2i - N - 1) \cdot r(d_i)}{N \sum_{j=1}^N r(d_j)} \quad (3)$$

Here, N represents total number of documents in the collection.

The Gini coefficient is a measure of inequality within a population [11]. A Gini coefficient of zero denotes perfect equality, indicating that all documents in the collection have an equal retrievability score according to $r(d)$. Conversely, a Gini coefficient of one indicates total inequality, with only one document having $r(d) = |Q|$ while all other documents have $r(d) = 0$. In most cases, retrievability scores exhibit varying degrees of inequality, resulting in a Gini coefficient between zero and one. Consequently, the Gini coefficient provides valuable insights into the level of inequality among documents in terms of their retrievability using a specific retrieval system and configuration. By comparing the Gini coefficients obtained from different retrieval methods, we can analyze the retrievability bias imposed by the underlying retrieval system on the document collection.

Retrievability, and the underlying theory of retrievability, has found applications in various domains. For instance, it has been used in the development of inverted indexes to enhance the efficiency and performance of retrieval systems by capitalizing on terms that contribute to a document's retrievability [20]. Additionally, retrievability has been leveraged to investigate bias in search engines and retrieval systems on the web [3] and within patent collections [8], leading to improvements in system efficiency during pruning processes [25].

2.3 Motivation

Retrievability scores offer insights into the accessibility of documents within a collection, reflecting their ease of retrieval by an

information retrieval system. On the other hand, PageRank, a fundamental algorithm in web search, assesses the importance and influence of web pages based on their incoming links. While both metrics aim to measure the significance of documents, they do so from distinct perspectives. Retrieval primarily considers how easily a document can be retrieved, while PageRank evaluates *navigability* of documents in terms of their popularity and how connected they are within a network. Our motivation, in this study, is to compare these two metrics to gain insights into the dynamics of information accessibility and navigability, providing a subtle view of document importance. This analysis can be useful in various domains, such as information retrieval, search engine optimization, content ranking etc.

A study has been conducted in [23] where Wilkie and Azzopardi compares the correlation between retrieval and navigability measures such as Hub, PageRank and Authority. Experiments conducted on three websites with slightly above 2,000 web pages in total reveal a negligible correlation between PageRank and Retrieval with the highest positive correlation reported to be 0.09. However, their study was conducted on a tiny set of institution webpages and the results are not reproducible due to the unavailability of the data. In this paper, we try to perform a similar study on two sizeable and publicly available datasets.

3 EMPIRICAL COMPARISON OF RETRIEVABILITY AND PAGERANK

3.1 Datasets and experimental setup

To conduct an empirical investigation comparing retrieval scores and PageRank values, it is essential that the dataset employed possesses a crucial characteristic - the presence of intra-links connecting the documents within the collection. This interconnection among documents is a prerequisite for the computation of the PageRank values. Without such links, the assessment and comparative study of these important metrics becomes unfeasible and impractical. For our study, we choose datasets that meet this requirement. We employ the English Wikipedia article dump from February 2023¹, an extensive dataset famous for its exhaustive coverage as well as intra-linking structure among articles. Additionally, we utilize the WT10g collection [13], which not only provides textual content but also includes valuable link information for web pages. Overall statistics of the datasets are presented in Table 1.

While performing the retrieval computation, one major component is the employed query set. For this study, we use the simulation method proposed in [4]. In this procedure, the terms undergo a series of steps that involve analysis and refinement including stemming, and the removal of stopwords. Terms that appear more than five times within the collection are considered single-term queries. Further, two-term queries are generated by pairing consecutive terms that each have a collection frequency of at least 20 occurrences. These generated bigrams are then ranked based on their frequency of appearance, with the top two million selected to form the final set of two-term queries. Note that, the queries are generated separately for each of the collections, and the respective query sets are exclusively used for retrieval on the collection

Table 1: Statistics of the datasets utilised for the study.

Dataset	# documents	Collection Type	# terms
WT10G	1,692,096	Web	9,674,707
Wikipedia	6,584,626	Wiki	18,797,260

Table 2: Gini Coefficient values for the population of Retrieval and PageRank scores computed in the two datasets.

	Gini Coefficient	
	Retrieval	PageRank
WT10g	0.5371	0.6618
Wikipedia	0.5380	0.7050

from which they originate. This ensures that the queries remain contextually relevant to their specific collections, maintaining the integrity of the retrieval process.

During retrieval for computing retrieval scores, we employ the Lucene² implementation of the BM25 model, with the default parameter settings. This choice aligns with the recommendations made by Azzopardi and Vinay in their initial as well as follow-up works [1–3, 5] on retrieval, ensuring consistency with established best practices. The only parameter of retrieval c (in Equation 2) is set to 100 while computing the retrieval scores.

In a similar study conducted in [23], a comparison was made between the hub and authority scores as well within a closed set of 2K documents from a university website. In contrast, it is worth noting that Wikipedia articles are structured around topics and categories, differing from the general web graph. As a result, the application of hub and authority concepts may not be directly applicable in this context. Hence, in our current research, we solely focus on comparing PageRank values as a measure of navigability within the Wikipedia dataset.

3.2 Experimental results and analysis

In this section, we present the outcomes of our experiments and provide insights drawn from these results. Our analysis begins by examining the distribution disparities within the retrieval scores and PageRank values across the datasets we utilized. To quantify these disparities, we employ the Gini coefficient, a well-established measure of inequality as discussed in Section 2.2. The specific values are presented in Table 2. One notable observation that emerges from this table is the substantial contrast between PageRank values and retrieval scores across datasets. This difference is most pronounced in the Wikipedia dataset, where we note a significant 31% difference between PageRank and retrieval values. The cumulative distributions of both scores are also graphically presented with Lorenz curve in Figure 1 where the divergence between the PageRank values and the retrieval values becomes specifically apparent in the latter part of the curve.

In Table 3, we provide correlations between retrieval and PageRank. To ensure a comprehensive analysis, we employ various rank-based correlation metrics, including Kendall’s rank correlation

¹<https://dumps.wikimedia.org/enwiki>

²<https://lucene.apache.org/>

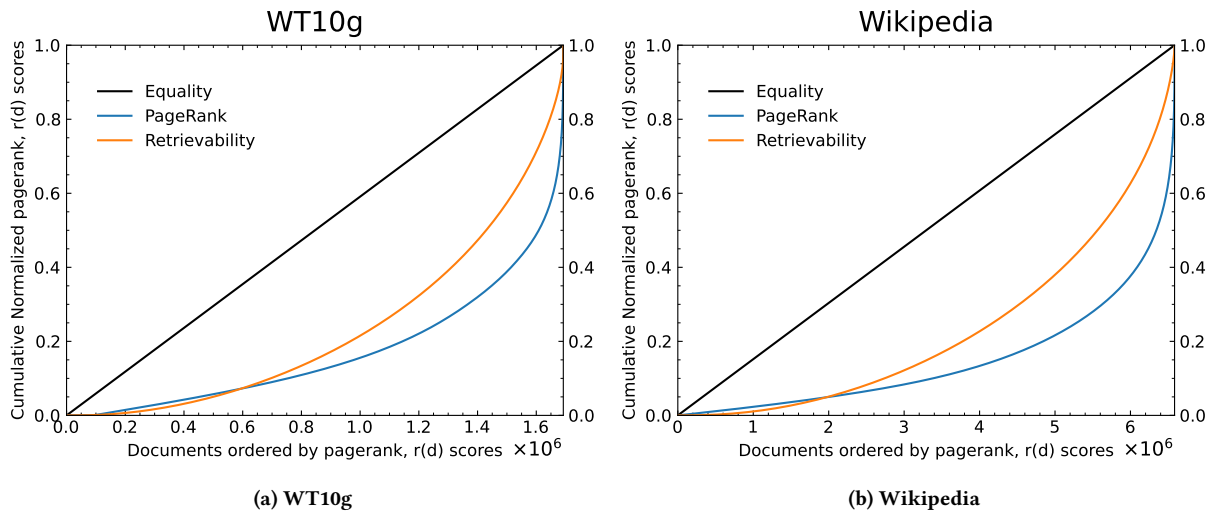


Figure 1: The Lorenz curve with the distribution of PageRank and Retrievability values on the WT10g collection and the Wikipedia English collection.

(τ), Spearman’s ρ , and Ranked Biased Overlap (RBO) [22]. Given the inherent differences in the values of retrievability and PageRank due to the way they are computed, we opt for rank-based correlation measures, excluding Pearson’s correlation coefficient, which would not be suitable in this context.

Our analysis reveals a relatively low correlation between the retrievability and PageRank values indicated by Kendall’s τ of 0.04 in WT10g collection. This observation is consistent with the findings from a previous study [23]. Further, Spearman’s rank correlation coefficient is noted to be 0.07 signifying a similar weak positive correlation between these two metrics. In contrast, we observe a notable increase in correlation in terms of both Kendall’s as well as Spearman’s rank correlation coefficient when we extend our analysis to the substantially larger Wikipedia collection. Specifically, we report correlation coefficients of 0.15 (τ) and 0.22 (ρ) between retrievability scores and PageRank values in the Wikipedia dataset. The significant increase in correlation coefficients for larger dataset suggests that dataset size and content diversity play a substantial role in the relationship between retrievability and PageRank. In other words, retrievability scores and PageRank values tend to exhibit a stronger correlation when working with more extensive and diverse datasets like Wikipedia. This observation implies that the nature of the documents and their interlinking within the dataset can influence how closely retrievability and PageRank align.

The most interesting insight arises from the value of RBO which exceeds 0.5 in both the datasets. This suggests a strong similarity between the rankings of documents when sorted based on their Retrievability and PageRank values. In essence, while lower Kendall’s τ and Spearman’s ρ indicate weak correlations overall, the higher value of RBO reveals a substantial overlap in the top-ranked documents when considering both retrievability and PageRank. This implies that, although the two metrics may not be highly correlated overall, they tend to agree on at least in terms of the top elements of their respective ranked lists (sorted based on the retrievability and PageRank values).

Table 3: Statistical correlation between Retrievability and PageRank when Retrievability computation is done using the original query generation technique [4].

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

4 CONCLUSION AND FUTURE WORK

Given a collection, the accessibility of the documents indicates the ease with which we can find documents which can be dissected based on distinct techniques employed. One can use a retrieval model, leading to the computation of retrievability scores, which gauges how readily a document can be retrieved within the collection. Another avenue involves navigation, where the navigability measures are derived from the interconnections and links between the documents themselves. The navigability metrics, such as PageRank, Hub, Authority provide insights into the discoverability via traversing document network and are distinct from the retrievability. Considering the diverse nature of finding documents based on the two approaches, in this paper, we have done a comparative study of retrievability and PageRank using two web datasets. Experimentation on WT10g collection reveals an almost negligible correlation between the two metrics in terms of Kendall’s and Spearman’s correlation coefficient methods. In contrast, better agreement is observed when Wikipedia, a larger and more extensively linked dataset, is used for the study. The ranked biased overlap measurements for both datasets show a significant similarity in the ranking of documents sorted based on the respective values. As part of a future study, a joint measurement of PageRank and Retrievability based on some fusion techniques will be tried.

REFERENCES

- [1] 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.
- [2] Leif Azzopardi, Rosanne English, Colin Wilkie, and David Maxwell. 2014. Page retrievability calculator. In *Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings 36*. Springer, 737–741.
- [3] Leif Azzopardi and Ciaran Owens. 2009. Search engine predilection towards news media providers. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 774–775.
- [4] 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>
- [5] Leif Azzopardi, Colin Wilkie, and Tony Russell-Rose. 2013. Towards Measures and Models of Findability. In *SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013)*. Citeseer, 3.
- [6] Ricardo A. Baeza-Yates, Paolo Boldi, and Carlos Castillo. 2006. Generalizing PageRank: damping functions for link-based ranking algorithms. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*. ACM, 308–315. <https://doi.org/10.1145/1148170.1148225>
- [7] Shariq Bashir and Andreas Rauber. 2009. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 1863–1866.
- [8] Shariq Bashir and Andreas Rauber. 2010. Improving retrievability of patents in prior-art search. In *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings 32*. Springer, 457–470.
- [9] 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](https://doi.org/10.1016/S0169-7552(98)00110-X)
- [10] Fan Chung. 2014. A Brief Survey of PageRank Algorithms. *IEEE Trans. Netw. Sci. Eng.* 1, 1 (2014), 38–42. <https://doi.org/10.1109/TNSE.2014.2380315>
- [11] 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.
- [12] David F. Gleich. 2015. PageRank Beyond the Web. *SIAM Rev.* 57, 3 (2015), 321–363. <https://doi.org/10.1137/140976649>
- [13] David Hawking. 2000. Overview of the TREC-9 Web Track. In *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000 (NIST Special Publication, Vol. 500-249)*, Ellen M. Voorhees and Donna K. Harman (Eds.). National Institute of Standards and Technology (NIST). <http://trec.nist.gov/pubs/trec9/papers/web9.pdf>
- [14] Helge Holzmann, Avishek Anand, and Megha Khosla. 2019. Estimating PageRank deviations in crawled graphs. *Applied Network Science* 4, 1 (Oct. 2019). <https://doi.org/10.1007/s41109-019-0201-9>
- [15] Jon M. Kleinberg. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM* 46, 5 (sep 1999), 604–632. <https://doi.org/10.1145/324133.324140>
- [16] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, 1065–1074. <https://aclanthology.org/D07-1114>
- [17] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 404–411. <https://aclanthology.org/W04-3252>
- [18] Liang Pang, Yanyan Lan, and Xueqi Cheng. 2021. Match-Ignition: Plugging PageRank into Transformer for Long-form Text Matching. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. ACM, 1396–1405. <https://doi.org/10.1145/3459637.3482450>
- [19] Sungchan Park, Wonseok Lee, Byeongseo Choe, and Sang-Goo Lee. 2019. A Survey on Personalized PageRank Computation Algorithms. *IEEE Access* 7 (2019), 163049–163062. <https://doi.org/10.1109/ACCESS.2019.2952653>
- [20] Jeremy Pickens, Matthew Cooper, and Gene Golovchinsky. 2010. Reverted indexing for feedback and expansion. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1049–1058.
- [21] Ahmed El Sheikh, Michele Bevilacqua, and Roberto Navigli. 2021. Integrating Personalized PageRank into Neural Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 9092–9098. <https://doi.org/10.18653/v1/2021.emnlp-main.715>
- [22] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (nov 2010), 38 pages. <https://doi.org/10.1145/1852102.1852106>
- [23] 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.
- [24] Li Zhang, Tao Qin, Tie-Yan Liu, Ying Bao, and Hang Li. 2007. N-Step PageRank for Web Search. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings (Lecture Notes in Computer Science, Vol. 4425)*. Springer, 653–660. https://doi.org/10.1007/978-3-540-71496-5_63
- [25] Lei Zheng and Ingemar J Cox. 2009. Document-oriented pruning of the inverted index in information retrieval systems. In *2009 International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 697–702.