Google's Personalized PageRank Algorithm: Optimizing Search Results

Madeleine Hansen, Elizabeth Huang, Chloe Shim

5 May 2024

1 Introduction

With Google being the primary search engine used across the globe, it has become the most used and powerful tool for search queries. An estimate from Hubspot reports that Google processes approximately 99,000 searches per second, which means 8.5 trillion searches per day and an average of 3-4 daily searches per person.[1] Having grown up in an environment surrounded in digital media, each member in this team is a frequent user of Google search. With such a heavy reliance on this technology, we decided to closely observe the process Google uses to produce personalized search rank results, known as Google's PageRank algorithm. A high-level overview of this process reveals an algorithm that assigns a rank to the importance of a web page and displays highly ranked web pages near the top of your search results.[2] Throughout this paper, we aim to investigate how Google's PageRank algorithm evaluates the importance of websites given a search query. After exploring the history and development of the algorithm, we will describe the complexity of Google PageRank's key algorithmic components, the benefits and harms of this development, and recent progress made to enhance or improve this algorithm.

2 History and Development

The development of the PageRank algorithm began with Alta Vista at DEC's Western Research Lab, founded by Stanford graduate students Larry Page and Sergey Brin in 1998 originally for a research project hoping to create a new search engine.[3] The initial PageRank algorithm in 1998 heavily relied on the number of incoming and outgoing links on a webpage to determine its importance. The paper also touches on the importance of efficient methods for crawling, indexing, and storage as the scale of the research's search operations began with 24 million pages in less than a week, expecting to expand to over 100 million pages in less than a month. As the number of web pages has only grown since then, more efficient strategies for ranking user search queries were needed. More specifically, there needed to be a way to "personalize" search results based on the most relevant content rather than the most popular content. Thus in 2006 (initial patent filing date, approved in 2015), the 1998 version of the PageRank algorithm was updated to personalize search results. This updated version of the algorithm, which is an extension of the original PageRank algorithm uses the shortest-path distances of all web pages to "seed" sites, which is a set of high-quality web pages that Google evaluates as trustworthy or reliable and well-connected to other pages, which restrains the scope of the search and augments the initial ranking of a web page's importance. [4] Hence, even with this update in 2015, the PageRank algorithm from 1998 still makes up a significant portion of how search results are shown to users. Beyond these fundamental algorithms used for search queries, we find that on Google's official overview of their search algorithm, in general, Google evaluates the following information to retrieve search results: keywords in your query, page relevance, reliability or trustworthiness of site, your geographic location, and the intent of your search, or context recognition. [5][6]

3 PageRank Components

To generalize, PageRank is a link analysis algorithm that assigns numerical scores to a website's content within a hyperlinked set of webpages, which measures the relative importance of the page for any given search query. In the following sections, we describe in more detail our modeling of the web including its structure as a webgraph, importance of edge weights, and seed sites.

3.1 Webgraph Representation

The first extensive study of representing a graphical representation of the hyperlink structure of the World Wide Web was introduced in a 1999 paper "The Web as a graph: measurements, models, and methods" which addresses a graph whose nodes are the static html pages, and directed edges are the hyperlinks that connect between them. [7][8] We can model this relationship with a directed graph as web pages have links directed towards themselves, or in-links, in addition to links directed outwards towards other

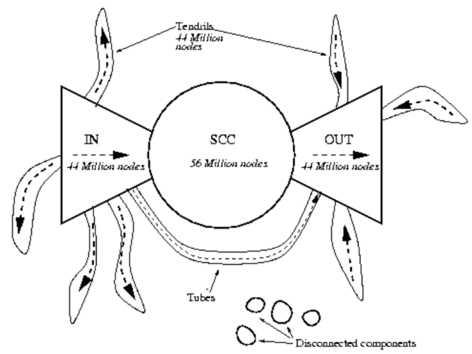

Figure 1: The larger topological structure of the Web Graph is mainly a Bowtie graph with a few extra parts. [12]

pages, or out-links. More generally, out-links are directed in the way that web surfers, or people who "surf" and explore the web, follow them. [9] Using this idea, we can attempt to crawl, repeatedly follow out-links, to explore the entire webgraph and uncover its larger topological structure (Figure 1). This representation of the web as a graph has allowed us to use webgraphs in the following applications:

- Detecting peculiar regions or clusters of the Webgraph: local subsets and hubs might share similar characteristics or have different statistical properties with respect to the whole structure
- Predicting new phenomena in the Web: easy to use known algorithms for crawling and searching for new or unusual activity
- Large scale computations: graph compression [Adler and Mitzenmacher; Boldi and Vigna 2004] [10][11]

3.2 Edge Weights

The core of calculating PageRank (PR) scores for web pages is how these importance scores propagate to influence the PR scores of the pages they link to, known as their out-link pages. Each web page distributes its PR score among its out-link pages. The basic PageRank algorithm assumes an equal distribution of a web page's PR score among its out-link pages, overlooking the different degrees of influence that a page's PR score can have on the PR scores of its out-link pages. To address this limitation, we consider the Weighted PageRank ("WPR") algorithm, an advanced version of the basic PR algorithm, described in further detail in Section 4, which more closely measures the relative importance of any given site.

3.3 Seed Sites

Seed sites are the pages that search engines start upon. They're Google's curated collection of high-quality web sites that Google deems to be reliable, diverse in topic and well-connected to other web pages. [13] Selecting a trustworthy and relevant seed site is important as it determines the quality of the search results provided to a user's query. In the case of selecting a "noisy seed," a starting site that does not accurately reflect a user's preferences or search focus, there will be an overestimate in the final PPR. [14]

4 How The Components Work Together: Complexity and Analysis

Google's PR patent provides the following overview in high-level steps: 1) the algorithm first receives a set of seed pages, pages that are highly connected to other links or have lots of outgoing links, to be ranked, 2) the algorithm then assigns weights to each edge based on the properties of the link and page to compute a WPR score, 3) the shortest distance from the set of seed pages to each outgoing page is computed based on the lengths of the links between the pages (greedy algorithm, sorted by shortest to longest lengths), 4) a ranking score is calculated for each page in the set of pages (greedy algorithm, sorted by shortest to longest distance), 5) finally a ranking is produced for the set of pages based on the resulting ranking scores (PPR).

4.1 Weighted PageRank Scores (Edge Weights)

For each Web page u in the Web graph, let $R(u)$ denote the set of pages that contain a link (reference) to u, and let $K(u)$ denote the set of out-link pages of u. Consider a Web page u and v, where v has a link to u. The PR score of page u depends on the rank values of $v \in R(u)$. The WPR distributes a higher rank value of v to an outlink page of v that is deemed more "important". We capture this notion of importance by adding weight on the directed link (v, u), denoted $W(v, u)$, for each $v \in R(u)$. $W(v, u)$ consists of two components: $W_{in}(v, u)$ and $W_{out}(v, u)$, which together measure the relative importance of u. $W_{in}(v, u)$ and $W_{in}(v, u)$

measure the importance of u in terms of the number of in-links of u and the number of out-links of u respectively, and they are computed using the following formulas:

$$W_{in}(v,u) = \frac{I_u}{\sum_{p \in R(v)} I_p}, W_{out}(v,u) = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

where I_u denotes the number of in-links of u, and O_u denotes the number of out-links of u. I_p denotes the number of pages that contain a link to page p, and O_p denotes the number of pages that p contains a link to. $\sum_{p \in R(v)} I_p$ is the sum of I_p values of all reference pages of v, and $\sum_{p \in R(v)} O_p$ is the sum of O_p values of all reference pages of v. Now, PR score of web page u can be represented using the following relationship [15]:

$$PR(u) \propto \sum_{v \in R(u)} PR(v) \cdot W_{in}(v,u) \cdot W_{out}(v,u)$$

Algorithm 1: $PR(P)$

```
                      ▷ P = set of all web pages, continued use in Component 3
scores = { }
for u in P do
    score = 0
    for v in u.getIncomingLinks do
        score += (PR(v)/v.getOutgoingLinks)·W_in · W_out
                  ▷ W_in is the weight of in-links, W_out is the weight of out-links
    scores[u] = score
return scores.sort(reverse = True)      ▷ sort from highest to lowest scores
```

4.2 Surfing Traversal

In the second component of the algorithm, we surf the web by making random walks from one site to another; our pseudocode below demonstrates this process. To exemplify, consider yourself a web surfer who starts from some randomly selected site on the internet. When exploring the site, you see that there are several external sites linked to the site you're currently on. You visit any of those sites and keep clicking on sites linked onto the following pages, which includes some revisiting some sites you've already seen and other new sites you have yet to explore. Then, you either get tired of following link after link or you've reached a dead end and you aren't able to find any new sites. Hence, we start this process over from a new randomly selected site, which we denote with the probability $1 - d$, where d stands for the damping factor - the probability that a web surfer will continue browsing by clicking on the linked pages (typically set around 0.85) and $1-d$ represents the probability that a web surfer will no longer continue browsing by clicking on the linked pages. [16]

Algorithm 2: *surf(s = seed.random)*

```
    ▷ select starting website from a seed set; continued use in Component 3
if user.IsTired then
    return surf(seed.random)                  ▷ start again from a new seed
else
    if len(s.getOutgoingLinks) > 0 then
        nextPage = s.getOutgoingLinks.shuffle[0]
        return surf(nextPage)
return surf(seed.random)
```

4.3 k-Shortest Paths

Next, we evaluate the personalization component, which is some score proportional to the distance of a website to a set of seed sites. Seed sites serve as a starting point for our hypothetical web surfer in the second component. The purpose of this is simple: if a web page is just a few clicks away from a reputable and relevant seed site, we can make the heuristic that the search results are likely to be personalized to the user's search query and more trustworthy. As described in the PPR algorithm's patent design, the ranking score of a website u is proportional to $e - D(u)$, wherein $D(u)$ the shortest distance from k of the nearest seed site candidates that to the site u, wherein k is a predetermined positive integer. [17] In other words, we compute distances of u to a fixed number of the nearest seed site candidates (ie: top 3 closest seed sites). The pseudocode below shows an algorithm that computes the k-closest seed sites to a site u.

Algorithm 3: *shortest*$(u, S, d$=[$((u, s)$:∞ for $s \in S$],k=*"optional"*)

```
Q = priorityQ(S)          ▷ initialize a priority queue all seed sites s ∈ S
while len(Q) ≠ 0 do
    s = Q.pop()
    if u.distance(s) < d[(u,s)] then
        d[(u,s)] = u.distance(s)
    return d[:k]
```

Piecing all these components together, we can represent the PPR score of any site u with the following relationship:

$$PPR(u) \propto e^{-D(u)} \sum_{v \in R(u)} PR(v) \cdot W_{in}(v,u) \cdot W_{out}(v,u)$$

Overall, we suggest an upper asymptotic bound of at worst $O(|P||R|)$ for the entire algorithm (all three components), where R is our entire set of references (links) and P is all the pages of our webgraph. In the first component of the algorithm, we simply compute a weight for each edge of the webgraph. In the worst case scenario, we would need to iterate through each web page to converge on a PR score, which can be considered as $O(|P||R|)$. In the second component of the algorithm, we traverse through each of the out-links to as many as all web pages in the webgraph in the worst case scenario. Hence, we can compare this component of the algorithm to the time complexity of BFS, where all neighbors are explored for a time complexity in the magnitude of $O(|R| + |P|)$. Lastly, our third component of the algorithm is a modified Dijkstra using a priority queue, which finds the shortest path distance at a complexity of $O(|R| log |P|)$. [17][18] Altogether, we have can rank the time complexities of these three components from largest to smallest: $O(|P||R|)$, $O(|R| log |P|)$, and $O(|R| + |P|)$ for an overall complexity of $O(|R||P|)$. [19][20]

5 Algorithm Benefits

Personalized web recommendations benefit users by providing them with relevant information based on their previous search history and demographics. A web page ranking algorithm that increases the likelihood that a user will select one of the top-ranked websites is considered efficient and preferable.

In addition, reaching the top of a user's search results has become a powerful tool for companies to get more exposure to their products and services; due to this reliance on search engine rankings, companies often look for ways to optimize their placement on search engines using SEO (Search Engine Optimization) techniques. According to Aziz Barbar and Anis Ismail, the writers of "Search Engine Optimization (SEO) for Websites," in 2019, SEO strategies can be categorized into on-page SEO and off-page SEO. On-page SEO techniques include website factors like page title, header tags, Meta keywords and description, headers, ALT tags, URL structure and size, internal linking, sitemaps, webpage compression and robot.txt. For example, this article's research demonstrates that Google prioritizes websites with ".com" versus ".tk" and websites with headers, bold, and italics. Off-page

SEO strategies increase website ranking with techniques that are not visible on the web page itself, such as "building backlinks for the websites, submitting the websites to online directories, increasing the number of visitors through blogs and online articles, and increasing the websites' reputation and visitors through forums, communities, and social media." [21] The 2021 article authored by Dirk Lewandowski, Sebastian Sünkler, and Nurce Yagci entitled "The influence of search engine optimization on Google's results: A multi-dimensional approach for detecting SEO" highlights that SEO strategies positively affect usability and accessibility of websites as well as relevance of website information, seen through the on-page SEO strategies. [22]

6 Algorithm Harms

Because search engines like Google are so widely used to look for information on a wide range of subjects, many people assume that search results are fair and unbiased. However, this is not the case and search engine algorithms are negatively impacted by algorithm bias, personalization efforts, and unethical SEO practices. "Evaluation metrics for measuring bias in search engine results" by Gezici1 et. al finds that Google search results have an ideological bias, meaning the algorithm prefers one ideological standing of conservative or liberal over the other in news search engine result pages. Without the possible impact of a user's search history on search results, the researchers determined that the search engines that they looked at, including Google, favored liberal articles possibly because of the input data that's received or because of the algorithm itself. Ideological bias on these algorithms can be powerful because the search result pages of a search engine mainly offer one particular view, limiting a user's comprehensive understanding of a subject. [23]

When personalized results come into play, it becomes even harder to find an unbiased point of view when searching. The user is likely to click on one of the first results, regardless of whether it offers a balanced perspective. "Bubble effect: including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility" by Marko Ćurković and Andro Košec discussed the research bubble effect, which is defined as the "tendency to be selectively exposed to personalized information in a way that influences individual beliefs and attitudes." The authors discuss how personalized search results inhibit the scope and depth of research because researchers see results that align with their previous beliefs and readings, which can prevent outside information. The researchers propose that because internet searching is irreproducible, researchers should disclose their search processes in their journals and use the internet as a secondary source of information. [24]

Finally, while search engine optimizers that follow guidelines are considered "white hat," many optimizers are unethical in their approach to achieve higher search engine rankings, using "black hat" SEO techniques. Since PPR algorithm is directly proportional to the number of links directed towards a page (backlinks), one technique in black hat SEO automates sites to point towards the site a company desires to promote, which inorganically produces a higher PR score. [25] Although these techniques may seem harmless for small businesses, the problem of unethical SEO practices is amplified when coupled with malicious intent. Cyber attackers using black hat SEO techniques, for instance, can use these techniques to manipulate search rankings and lead users to unsafe phishing sites. Additionally, using these techniques can suppress important communication or information from marginalized communities that may already be at a quantitative disadvantage in terms of popularity or importance scores.

7 Summary and Recent Inquiries

Google, as the world's primary search engine, processes a massive amount of user queries daily. In our exploration of this topic, we analyzed the inner workings of personalized search result rankings, most notably Google's PageRank algorithm and the ranking scores assigned to various web pages across the internet. Overall, we evaluate the algorithm as effective and working as intended; search results are relevant to our queries and personalized based on sites we frequently visit (more significant edge weights). Beyond the algorithm's current construction, we also wanted to provide some recent developments and open research problems in providing more efficient, flexible (to changes in the webgraph), scalable (distributed), and diverse search results to user search queries. Ongoing research to improve search engine systems include the following. Aimed at reducing the expense of complex matrix operations, recent improvements focus on approximation techniques and parallelization or distributed computing to handle massive graphs more efficiently. [26] Also, recent research focuses on balancing personalization with diverse content by improving on graph partitioning and data structure representations. [27] After examining the bias of current SEO techniques, we've also explored current ongoing research to equalize reliable web pages.

Specifically, improvements to the current PPR algorithm to strengthen the trustworthiness of resulting sites such as Google's E-E-A-T, which stands for Experience, Expertise, Authoritativeness, and Trustworthiness that extracts key features of a site's content to evaluate its quality.[28] Finally, another area of ongoing research is an adaptation to user behavior to provide further personal privacy while generating personalized results. [29] As search engine technology continues to evolve with updates in user query behavior and web page systems, algorithms such as PageRank are already beginning to fade in relevance. However, even after decades of research in the field, the core of these search engine algorithms to provide the most relevant and important results to a user's queries remains fundamental to understanding the latest adaptations of the algorithm and its future.

References

[1] Prater, M. (2023, August 30). 31 Google Search Statistics to Bookmark ASAP. HubSpot Blog. https://blog.hubspot.com/marketing/google-search-statistics

[2] Roberts, E. (2016). The Google PageRank Algorithm. https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf

[3] Brin, S., Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30(1-7), 107-117. https://doi.org/10.1016/S0169-7552(98)00110-X

[4] Hajaj, Nissan. "Producing a Ranking for Pages Using Distances in a Web-Link Graph." Google Patents, 20 Oct. 2015, patents.google.com/patent/US9165040B1/en

[5] Ranking Results – How Google Search Works. (n.d.). Google. Retrieved April 30, 2024, from https://www.google.com/search/howsearchworks/how-search-works/ranking-results/

[6] A Guide to Google Search Ranking Systems — Google Search Central — Documentation. (n.d.). Google for Developers. Retrieved from https://developers.google.com/search/docs/appearance/ranking-systems-guide

[7] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (2003). The Web As a Graph: Measurements, Models, and Methods. In T. Asano, H. Imai, D.T. Lee, S. Nakano, & T. Tokuyama (Eds.), Computing and Combinatorics: 5th Annual International Conference, COCOON'99, Tokyo, Japan, July 26-28, 1999, Proceedings (pp. 1-17). Springer Berlin Heidelberg.

[8] Donato, D., Laura, L., Leonardi, S., & Millozzi, S. (2007, February 1). The Web as a graph: How far we are. ACM Transactions on Internet Technology, 7(1), 1-25. https://doi-org.ezproxy.wellesley.edu/10.1145/1189740.1189744

[9] Randall, K. H., Stata, R., Wickremesinghe, R., & Wiener, J. L. (2001, November 16). The Link Database: Fast Access to Graphs of the Web. Search Research Report. Retrieved from https://www.eecs.harvard.edu/~michaelm/CS222/linkdb.pdf

[10] Boldi, P., & Vigna, S. (2004, May). The Webgraph Framework I: Compression Techniques. WWW '04: Proceedings of the 13th international conference on World Wide Web, 595–602. https://doi-org.ezproxy.wellesley.edu/10.1145/988672.988752

[11] Donato, D., Laura, L., Leonardi, S., & Millozzi, S. (2007, February 1). The Web as a graph: How far we are. ACM Transactions on Internet Technology, 7(1), 1-25. https://doi-org.ezproxy.wellesley.edu/10.1145/1189740.1189744

[12] Broder, Andrei, et al. "Graph Structure in the Web." Computer Networks, vol. 33, no. 1-6, June 2000, pp. 309–320, https://doi.org/10.1016/s1389-1286(00)00083-9

[13] "More Guidance on Building High-Quality Sites." Google Developers,developers.google.com/search/blog/2011/05/more-guidance-on-building-high-quality

[14] S. Huang, X. Li, K. S. Candan and M. L. Sapino, ""Can you really trust that seed?": Reducing the impact of seed noise in personalized PageRank," 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, 2014, pp. 216-223, doi:10.1109/ASONAM.2014.6921586

[15] Xing, W., & Ghorbani, A. (2004). Weighted PageRank Algorithm. https://doi.org/10.1109/dnsr.2004.1344743

[16] Dekhtyar, Alexander.CSC 466: Knowledge Discovery from Data Alexander Dekhtyar. 2009.

[17] Lacki, J. Mitrovic, S., Onack, K., Sankowski, P. (2019). Walking Randomly, Massively, and Efficiently. arXiv. http://arxiv.org/abs/1907.05391

[18] Yen, Jin Y. "An Algorithm for Finding Shortest Routes from All Source Nodes to a given Destination in General Networks." Quarterly of Applied Mathematics, vol. 27, no. 4, 1 Jan. 1970, pp. 526–530, https://doi.org/10.1090/qam/253822

[19] Lofgren, Peter. Efficient Algorithms For Personalized Pagerank A Dissertation Submitted To The Department Of Computer Science And The Committee On Graduate Studies Of Stanford University In Partial Fulfillment Of The Requirements For The Degree Of Doctor Of Philosophy. 2015.

[20] Liu, Jiaying, Et Al. "Random Walks: A Review Of Algorithms And Applications." IEEE Transactions On Emerging Topics In Computational Intelligence, Vol. 0, No. 0, 2019, Arxiv.org/pdf/2008.03639.pdf

[21] Barbar, Aziz, and Anis Ismail. "Search Engine Optimization (SEO) for Websites." Proceedings of the 2019 5th International Conference on Computer and Technology Applications - ICCTA 2019, 2019, https://doi.org/10.1145/3323933.3324072

[22] Lewandowski, Dirk, et al. "The Influence of Search Engine Optimization on Google's Results." 13th ACM Web Science Conference 2021, 21 June 2021, pp. 12–20, https://doi.org/10.1145/3447535.3462479

[23] Gezici, Gizem, et al. "Evaluation Metrics for Measuring Bias in Search Engine Results." Information Retrieval Journal, vol. 24, no. 2, 27 Jan. 2021, pp. 85–113, https://doi.org/10.1007/s10791-020-09386-w

[24] Ćurković, Marko, and Andro Košec. "Bubble Effect: Including Internet Search Engines in Systematic Reviews Introduces Selection Bias and Impedes Scientific Reproducibility." BMC Medical Research Methodology, vol. 18, no. 1, 13 Nov. 2018, https://doi.org/10.1186/s12874-018-0599-2

[25] Malaga, Ross A. "Chapter 1 - Search Engine Optimization—Black and White Hat Approaches." ScienceDirect, Elsevier, 1 Jan. 2010, www.sciencedirect.com/science/article/abs/pii/S0065245810780013

[26] Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., & Smola, A. J. (2013). Distributed large-scale natural graph factorization. Proceedings of the 22nd International Conference on World Wide Web - WWW '13. https://doi.org/10.1145/2488388.2488393

[27] P. Mohapatra and Huamin Chen, "WebGraph: a framework for managing and improving performance of dynamic Web content," in IEEE Journal on Selected Areas in Communications, vol. 20, no. 7, pp. 1414-1425, Sept. 2002, doi:10.1109/JSAC.2002.802072

[28] Tucker, Elizabeth. "Our latest update to the quality rater guidelines: E-A-T gets an extra E for Experience", Google, 15 December 2022,https://developers.google.com/search/blog/2022/12/google-raters-guidelines-e-e-a-t

[29] Xu, Yabo, et al. "Privacy-Enhancing Personalized Web Search." Proceedings of the 16th International Conference on World Wide Web, 8 May 2007, https://doi.org/10.1145/1242572.1242652