HITS vs. PageRank: A Comparative analysis of Web Search Algorithms

Richard Osei Adu
Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development (AAMUSTED)

Klinsman Kwaku Boateng
Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development (AAMUSTED)

William Asiedu, PhD
Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development (AAMUSTED)

ABSTRACT
This paper presents a comparative analysis of two prominent web search algorithms, Hyperlink-Induced Topic Search (HITS) and PageRank, which are widely used for ranking web pages in information retrieval systems. The study explores the theoretical foundations, algorithmic structures, performance metrics, and practical applications of both algorithms, highlighting their unique approaches to evaluating the importance of web pages. Using the Google Web Graph dataset and Cit-HepPh citation network, an empirical evaluation was conducted to assess the efficiency and effectiveness of HITS and PageRank in identifying key nodes within a network. The study evaluates their performance in ranking nodes, considering structural properties, correlation analysis, and score distributions. Results indicate that while PageRank ensures a balanced representation of node importance, HITS uniquely identifies key hubs and authorities. The findings reveal that while PageRank offers a more balanced distribution of page importance across a network, HITS effectively distinguishes between hubs and authorities, making it valuable for specific contexts like academic research and topic-specific searches. The low correlation between the scores of the two algorithms underscores their distinct methodologies and implications for search engine optimization. The paper concludes by recommending the use of each algorithm based on specific use cases and the nature of the web environment being analyzed.

General Terms
Information Retrieval, Algorithms, Web Search, Network Analysis, Ranking Metrics, Graph Theory, Search Engine Optimization

Keywords
HITS, PageRank, Web Search Algorithms, Information Retrieval, Link Analysis, Search Engine Optimization

1. INTRODUCTION
Web searching has become a critical function in our current modern information ecosystem. As a fundamental tool for navigating the amount of data on the web, search engines rely heavily on some sophisticated algorithms to help in ranking and retrieving of relevant web pages to users. Notably among these algorithms are the Hyperlink- Induced topic search (HITS) and PageRank algorithms [4]. Although these algorithms were designed to address the same fundamental purpose of ranking web pages, they both adopt different approaches and have different application methods. The PageRank algorithm which was developed late 1990 by Larry page and Sergey Brin is has its foundation in link analysis, where outmost priority of a web page is determined by the number (quantity) and also quality of links pointing to the web page [9]. This has become the base algorithm in Google's search engine, providing efficient and scalable strategy for ranking web pages. On the contrary, the HITS algorithm introduced by Kleinberg in 1991, centers on identification of two types of web pages; which is hubs and authorities. The hubs are pages that are linked to many other pages, while authorities are pages that are linked to many hub pages. HITS assign two scores to the pages, that is hub and authority score. This algorithm is useful in effectively identifying communities of related information, making it valuable in contexts like academic research information retrieval and topic-specific searches. This paper aims to provide a comparative analysis of these algorithms by exploring their practical applications, theoretical justifications and effectiveness in web searching. By examining the strengths and weakness of both algorithms, the study seeks to contribute to current development of search engine technologies and the optimization of web searching.

2. RELATED WORKS
The HITS (Hyperlink-Induced Topic Search) and PageRank algorithms are foundational techniques in web page ranking and network analysis, each with distinct methodologies and implications. This literature review combines key findings from recent studies, highlighting their applications, strengths, and limitations.

2.1 **Algorithmic Foundations:** PageRank operates on the principle of link analysis, assigning scores based on the quantity and quality of incoming links, effectively mirroring the degree distribution in networks. It can enhance minority representation in rankings [7]. HITS, in contrast, differentiates between authority and hub scores, but has been shown to amplify biases in homophilic networks, particularly affecting minority groups negatively [7].

2.2 **Applications and Optimizations:** [10] highlighted that both algorithms are integral to web page ranking strategies, demonstrating fast computation suitable for large networks. Recent research emphasizes their optimization for improved ranking efficiency in complex web environments. The HITS algorithm has been adapted for Graph Neural Networks (GNNs), enhancing performance in semi-supervised learning by incorporating authority and hub scores in its propagation mechanism [1].

2.3 Comparative Efficiency: In network attack scenarios, PageRank has been found to outperform HITS in terms of computational and attack efficiency, making it a preferred choice for strategic applications in complex networks [8].

While both algorithms have proven effective in various contexts, their differing impacts on bias and efficiency highlight the need for careful consideration in their application, particularly in sensitive domains.

3. METHODOLOGY

The study utilizes a mixed approach, thus integrating theoretical analysis with empirical assessment. The mixed methods approach is justified in this research due to its ability to enhance validity through data triangulation and effectively handle complex social phenomena by integrating both quantitative and qualitative data [3]. Furthermore, it supports the development of comprehensive frameworks and interdisciplinary insights by merging theoretical and empirical analyses [5]. The theoretical study entails a thorough investigation of the HITS and PageRank algorithms, with a specific emphasis on their mathematical formulations, computational complexity, and fundamental principles. The empirical evaluation is performed utilizing Google Web Graph dataset from the Stanford Large Network Dataset Collection (SNAP). Measures including correlation and computation efficiency were evaluated. The experiment was carried out within a simulated search environment to guarantee controlled and replicable outcomes.

3.1 Theoretical analysis

The theoretical analysis involves a deep dive into the core principles, mathematical formulations, and computational complexities of the HITS and PageRank algorithms.

3.1.1: Mathematical Formulation

HITS Algorithm: The HITS algorithm operates by iteratively assigning two scores to each web page: a hub score and an authority score. These scores are derived from the link structure of the web. The algorithm is mathematically represented by the following equations:

$$\text{Hub(p)} = \sum_{q \in L(p)} \text{Authority(q)}$$

$$\text{Authority(p)} = \sum_{q \in B(p)} \text{Hub(q)}$$

Here, **L(p)** represents the set of pages linked by page **p**, and **B(p)** represents the set of pages that link to page **p**. The algorithm iteratively updates these scores until convergence is achieved.

PageRank Algorithm: PageRank models the web as a directed graph, with web pages as nodes and hyperlinks as edges. The PageRank of a page is calculated using the formula:

$$PR(\text{p}) = \frac{1-d}{N} + d \sum_{q \in B(p)} \frac{PR(q)}{L(q)}$$

In this equation, d is the damping factor (typically set to 0.85), **N** is the total number of pages, **B(p)** is the set of pages that link to page **p**, and **L(q)** is the number of outbound links from page **q**. Like HITS, PageRank is also computed iteratively until the values converge.

3.1.2 Computational Complexity:

• HITS Complexity: HITS requires multiple iterations to compute the hub and authority scores for all pages in the network. The time complexity of HITS depends on the number of iterations required for convergence and the number of links in the web graph. Generally, the complexity is O(k×(n+m))O(k \times (n + m))O(k×(n+m)), where k is the number of iterations, n is the number of pages, and m is the number of links.

• PageRank Complexity: The complexity of PageRank is similarly dependent on the number of iterations and the structure of the web graph. The time complexity of PageRank is typically O(k×n)O(k \times n)O(k×n), where k is the number of iterations and n is the number of pages. The damping factor d plays a role in the speed of convergence.

3.1.3 Core Principles:

• HITS Principles: HITS is designed to identify two types of important web pages: hubs and authorities. A good hub points to many authoritative pages, and a good authority is pointed to by many hubs. This bipartite relationship between hubs and authorities is a key distinguishing feature of HITS.

• PageRank Principles: PageRank is based on the idea that a page is important if it is linked to by many other important pages. The algorithm models the random surfing behavior of a user, with the damping factor representing the probability that the user will continue following links. PageRank's core strength lies in its ability to rank pages based on their overall link structure rather than just their immediate connections.

3.2 Empirical Evaluation

An empirical evaluation of the Hyperlink-Induced Topic Search (HITS) and PageRank algorithms on a large-scale web graph dataset, the Google Web Graph from the Stanford Large Network Dataset Collection (SNAP), which consists of **875,713 nodes and 5,105,039 edges**. The objective was to compare the performance and effectiveness of these algorithms in identifying the most important nodes within the network. By analyzing the top-ranked nodes and calculating the correlations between the different scores, this analysis aims to understand how these algorithms evaluate node importance and identify key differences in their approach. To extend the evaluation, the DBLP citation network dataset was utilized, featuring nodes as authors and edges representing citation relationships. This dataset, smaller and more community-driven, enabled analysis under contrasting structural conditions.

3.3 Experimental Setup

The simulation was conducted in Google Colab using Python and the NetworkX library. Dataset utilized was a directed graph representation of the Google web graph dataset, with nodes representing webpages and edges representing hyperlinks between them. The HITS algorithm was applied to compute Hub and Authority scores, and the PageRank algorithm was used to determine node importance. Key steps in the implementation included loading the graph data, applying the HITS algorithm using the nx.hits() function, and calculating PageRank scores with nx.pagerank(). To further expand the evaluation, the Cit-HepPh dataset was also simulated. This is a citation network derived from the High Energy Physics - Phenomenology category on arXiv. Nodes in this dataset represent papers, while directed edges represent citation relationships, where one paper cites another.

4. Results

The tables in this section presents a comparative analysis of the top 5 nodes based on Hub, Authority, and PageRank scores derived from the Google web graph and Cit-HepPh dataset after the simulation.

Table 1: Network properties

Dataset	Nodes	Edges	Average Degree	Density
web-Google	875,713	5,105,039	11.66	6.66×10^{-6}
Cit-HepPh	34,546	421,578	24.41	3.53×10^{-4}

Table 2: Top Nodes Based on HITS (Hub and Authority) and PageRank Scores

Metric	Top Node	Score
web-Google		
Hub Score	707772	0.000384
Authority Score	819223	0.0741
PageRank Score	163075	0.000952
Cit-HepPh		
Hub Score	9909232	0.001809
Authority Score	9803315	0.013204
PageRank Score	9303255	0.003655

Table 3: Correlation between Scores

Dataset	Correlation (Hub vs. PageRank)	Correlation (Authority vs. PageRank)
web-Google	0.002	0.144
Cit-HepPh	0.791	0.908

4.1 Interpretation of Top Scores

Web-Google Dataset:
Hub Scores: Nodes with the highest hub scores (e.g., Node 707772 with a score of 0.000384) link to many other nodes identified as authorities. However, the narrow range of scores indicates that these hubs are not overwhelmingly dominant, suggesting a relatively distributed connectivity pattern among hubs.

Authority Scores: The top authority score (Node 819223 with 0.0741) stands out significantly compared to other nodes. This node is heavily linked to by high-quality hubs, making it an important resource within the network.

PageRank Scores: The PageRank scores are tightly clustered, with the highest score (Node 163075, 0.000952) slightly exceeding others. This indicates a balanced distribution of importance among the nodes, with no single node disproportionately central.

Cit-HepPh Dataset:

Hub Scores: The top hub scores (e.g., Node 9909232 with 0.001809) are more varied compared to the web-Google dataset, highlighting nodes with stronger roles in linking to authoritative sources within the citation network.

Authority Scores: Node 9803315, with a score of 0.0132, is notably more authoritative than others, indicating a highly cited work or author that serves as a significant reference point. **PageRank Scores:** The highest PageRank score (Node 9303255, 0.003655) reflects nodes that are well-connected and central, aligning with their citation importance.

4.2 Correlation between HITS and PageRank

.Web-Google Dataset: The correlation between hub scores and PageRank scores is extremely low (0.002), reflecting the fundamental differences in how these metrics define importance—HITS focuses on directional connectivity, while PageRank emphasizes global link structure. The correlation between authority scores and PageRank scores is modest (0.144), indicating some overlap in the measurement of node importance but with distinct methodologies.

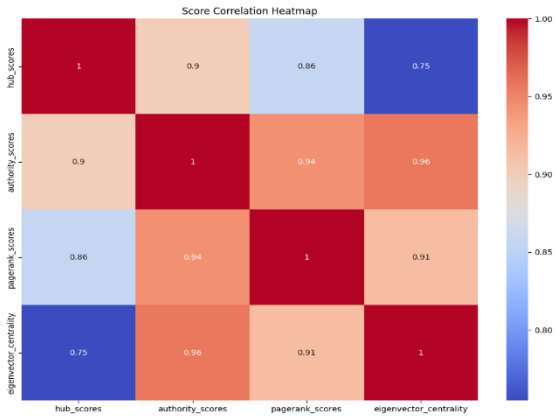

Fig. 1 Web-Google Dataset score correlation heatmap

Cit-HepPh Dataset: The correlation between hub scores and PageRank scores is much higher (0.791), suggesting that nodes functioning as hubs also tend to exhibit centrality as measured by PageRank. The correlation between authority scores and PageRank scores is notably strong (0.908), showing a significant alignment between being an authoritative source and being centrally positioned in the citation network.

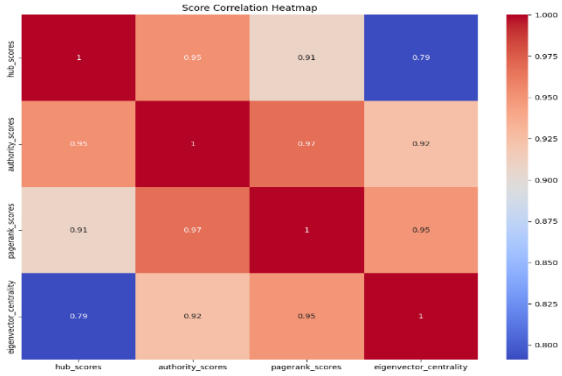

Fig. 2 Web-Google Dataset score correlation heatmap

The differing correlation patterns between the datasets highlight the influence of network structure on the performance and interrelationship of the algorithms. In citation networks, where importance tends to be hierarchical, HITS and PageRank align more closely. In web-based networks, the distributed and diverse nature leads to weaker correlations.

4.3 Distribution Insights

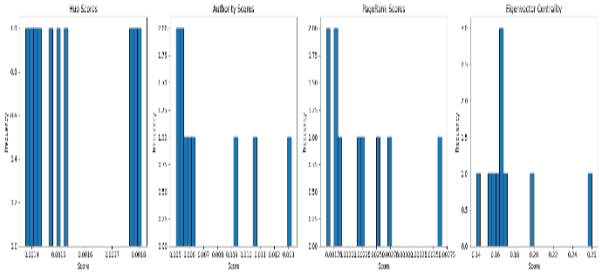

Fig. 3 Web-Google Dataset distribution graph

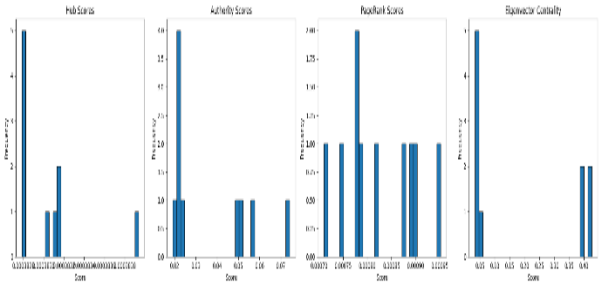

Fig. 3 Web-Google Dataset distribution graph

Cit-HepPh Dataset

Hub Scores Distribution:

The distribution is highly **right-skewed**, with a significant concentration of scores near the lower end. This pattern highlights that only a small number of nodes (papers) serve as key hubs, linking to highly authoritative nodes. The sparse appearance of higher hub scores emphasizes the selective referencing nature within academic citations.

Authority Scores Distribution:

The distribution exhibits more **uniform dispersion**, with a few dominant peaks. This indicates the presence of key authoritative nodes (highly cited papers) alongside other moderately significant nodes. These patterns align with the hierarchical nature of citation networks, where a handful of papers dominate in authority.

PageRank Scores Distribution:

Scores are concentrated, with a few outliers achieving higher values. The clustering suggests that most papers are of moderate importance, but a small group holds significant influence due to direct citations and their connections.

Eigenvector Centrality Distribution:

Conversely, the web-Google dataset demonstrates lower correlations, underscoring the distributed nature of web-based networks and the distinct methodologies of the algorithms.

The differences in score distributions further emphasize their complementary roles: HITS excels in identifying standout nodes with high authority or hub values, while PageRank ensures equitable importance distribution across nodes. These insights are critical for tailoring algorithm use to specific network characteristics, whether in search engine optimization, academic impact analysis, or social network exploration. Future research could build on these findings by developing

The scores are **bimodal**, with a dominant cluster near the lower end and a smaller group in the higher range. This suggests a dichotomy where some nodes exhibit limited influence while others are critical for bridging sub-networks.

Web-Google Dataset

Hub Scores Distribution:

The scores are more **uniformly distributed**, indicating a broader spread of nodes acting as hubs. Unlike Cit-HepPh, web networks tend to have more even participation in terms of linking to authorities. The presence of higher scores showcases the importance of well-connected nodes, such as directory or content aggregation pages.

Authority Scores Distribution:

The scores are moderately skewed, with the majority of nodes clustering at lower values and a few achieving significant authority. This reflects the web's natural structure, where a few authoritative pages dominate search rankings.

PageRank Scores Distribution:

A pronounced cluster near the lower range is visible, with a gradual rise in scores toward the higher range. This reflects the hierarchical structure of the web, where a small subset of nodes (such as homepages or popular websites) achieve dominance.

Eigenvector Centrality Distribution:

The scores are **highly concentrated**, with a few significant outliers. This reflects the web's power-law distribution, where a small number of nodes are highly influential in connecting disparate parts of the network.

5. CONCLUSION

This study provides a comparative analysis of the HITS and PageRank algorithms across two distinct datasets: the web-Google graph and the Cit-HepPh citation network. By examining the algorithms' performance in ranking nodes, their correlation, and score distributions, we highlight their unique strengths and contextual applicability.

The findings reveal that HITS effectively distinguishes between hubs and authorities, making it suitable for applications requiring dual-role analysis, such as identifying critical nodes in topic-specific searches or academic networks. PageRank, on the other hand, demonstrates a balanced distribution of importance across networks, emphasizing global connectivity and centrality.

The inclusion of the Cit-HepPh dataset enriches the study, showcasing how algorithmic behaviors adapt to network structures. In the citation network, the stronger correlations between PageRank and HITS scores reflect a hierarchical structure where authoritative nodes align with central ones.

hybrid models that combine the strengths of HITS and PageRank to optimize ranking accuracy. Additionally, extending evaluations to include dynamic and evolving networks could provide insights into algorithmic adaptability over time.

6. REFERENCES

[1] Khan, M., Mello, G. B. M., Engelstad, P., Habib, L., & Yazidi, A. (2022). HITS-GNN: A Simplified Propagation Scheme for Graph Neural Networks. 2022 IEEE International Conference on Big Data (Big Data).

[2] Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (2008). Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. arXiv (Cornell University).

[3] Mireia, Bolíbar. (2016). Macro, meso, micro: broadening the 'social' of social network analysis with a mixed methods approach. Quality & Quantity, 50(5):2217-2236.

[4] S. Deshmukh and K. Vishwakarma, (2021). "A Survey on Crawlers used in developing Search Engine," 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1446-1452.

[5] Sharlene, Nagy, Hesse-Biber. (2010). Mixed Methods Research: Merging Theory with Practice.

[6] SNAP: Network datasets: Google web graph. (n.d.). https://snap.stanford.edu/data/web-Google.html

[7] Stoica, A. A., Litvak, N., & Chaintreau, A. (2024). Fairness Rising from the Ranks: HITS and PageRank on Homophilic Networks.

[8] Su, Y., Yi, Y., & Qin, J. (2019). The attack efficiency of PageRank and HITS algorithms on complex networks. International Journal of Embedded Systems, 11(3), 306.

[9] Weiss, G. M., Nguyen, N., Dominguez, K., & Leeds, D. D. (2021). Identifying Hubs in Undergraduate Course Networks Based on Scaled Co-Enrollments: Extended Version. arXiv (Cornell University).

[10] Zhang, X., & Wu, H. (2021). PageRank Algorithm and HITS Algorithm in Web Page Ranking. In Advances in intelligent systems and computing (pp. 389–395).