deti universidade de aveiro
departamento de electrónica,
telecomunicações e informática

PageRank and Link-Based Ranking Algorithms

José Miguel Costa Gameiro, 108840

Professors José Luís Oliveira and Tiago Almeida

Information Retrieval
DETI
Universidade de Aveiro
January, 2025

Table of Contents

1 Introduction

The wide expanse of the World Wide Web presents an important challenge, which is, how to organize and retrieve relevant information efficiently from all the existing content. In this context, ranking algorithms play an important role in information retrieval systems, particularly in search engines. With these algorithms, it is possible to prioritize web pages based on their relevance and authority, ensuring that users are presented with the most useful results for their queries. If these ranking systems didn't exist, then navigating through all the content available online would become overwhelming and inefficient.

Link-based algorithms are computational methods that analyze the structure of links within a network such as the World Wide Web, to make conclusions about the importance of authority nodes,like a web page. These algorithms depend on the natural inter connectivity of networks whereby a hyperlink from one page to another is basically an endorsement, or an indicator of relevance. Link-based algorithms rank pages or nodes based on the number, quality, and context of such links and attempt to return the most authoritative and relevant content for a user. Examples of such algorithms include Google's Page Rank and the HITS algorithm.

An example of a search engine is the one from Google, where it uses ranking algorithms to evaluate billions of web pages and produce a set of results that are connected to the query in context. These algorithms help the user by addressing some factors like the relevance of content to the query, the credibility of sources, and the ability to filter out spam (irrelevant material).

As previously stated before, the development and integration of link-based ranking algorithms in information retrieval systems helps organize and rank massive amounts of data systematically. Unlike keyword-based ranking, which relies only on textual matches, link-based algorithms consider hyperlinks as indicators of trust and relevance, recognizing that links from authoritative websites often reflect valuable content.

The objective of this monograph is to present multiple link-based ranking algorithms, focusing on their methodologies, and provide a comparison between all of the algorithms presented, following some topics that are important to information retrieval systems.

2 Approaches Studied

2.1 Central Methods/Strategies

2.1.1 Page Rank (PR)

Page Rank is a foundational algorithm developed by Larry Page and Sergey Brin that evaluates the importance of web pages based on their link structure[1]. It models a *random surfer* who navigates the web randomly clicking on links, with a probability d (damping factor) of continuing to another page and $(1-d)$ probability of jumping to any page randomly. This damping factor, typically set around 0.85, ensures the model accounts for random jumps, preventing rank sinks.

Mathematically, Page Rank can be represented as

$$PR(i) = \frac{1-d}{N}kd \sum_{j \in \mathrm{In}(i)} \frac{PR(j)}{\mathrm{Out}(j)}$$

where N is the total number of pages, $\mathrm{In}(i)$ denotes the set of pages linking to page i, and $\mathrm{Out}(j)$ is the number of outbound links from page j. This equation is derived from the random surfer model[1].

Expressed in a matrix form, Page Rank becomes:

$$\mathbf{PR} = d\mathbf{P}^T\mathbf{PR} + \frac{1-d}{N}\mathbf{1}$$

where $\mathbf{P}$ is the transition probability matrix, and $\mathbf{1}$ is a vector of ones. This formulation aligns with the iterative computation method described in[2]. The Page Rank vector $\mathbf{PR}$ is the principal eigenvector of the modified transition matrix, computed iteratively until convergence occurs.

Recent advancements have explored the integration of artificial intelligence and machine learning techniques to enhance Page Rank's adaptability and efficiency in processing large-scale web data[3]. Additionally, adaptations like Gene Rank apply Page Rank principles to rank genes in biological networks, demonstrating the algorithm's versatility beyond web search[4]

2.1.2 HITS Algorithm (Hyperlink-Induced Topic Search)

The HITS algorithm, introduced by Jon Kleinberg, identifies two key roles for web pages: *hubs* and *authorities*. Hubs are pages that link to multiple authoritative pages, while authorities are pages extensively linked by hubs. The algorithm operates in two main steps:

1. **Authority Update:** Each page's authority score is updated based on the hub scores of pages linking to it.

2. **Hub Update:** Each page's hub score is updated based on the authority scores of pages it links to.

These updates are performed iteratively until the scores converge. Mathematically, if a_i and h_i represent the authority and hub scores of page i, respectively:

$$a_i = \sum_{j \in \text{In}(i)} h_j, \quad h_i = \sum j \in \text{Out}(i) a_j$$

Recent research has applied the HITS algorithm in various domains, including evaluating news' authenticity by identifying authoritative sources[5], and enhancing graph neural networks by incorporating HITS-based propagation paradigms to improve learning on graph-structured data[6].

2.1.3 SALSA (Stochastic Approach for Link-Structure Analysis)

SALSA is a link analysis algorithm that combines elements of both HITS and Page Rank. It models the web as a bipartite graph, distinguishing between hubs and authorities, and performs random walks within this structure. By analyzing the stationary distribution of these random walks, SALSA assigns authority and hub scores to pages.

The algorithm constructs two Markov chains: one for hubs and one for authorities. Transition probabilities are determined based on the link structure, and the stationary distributions of these chains provide the hub and authority scores. This stochastic approach offers improved stability and resistance to disturbances compared to HITS.

While SALSA has been foundational in link analysis, recent developments have seen its principles applied in other fields. For instance, the SALSA optimizer introduces an automatic step size selection method for machine learning algorithms, enhancing optimization processes without the need for manual tuning[7].

2.1.4 Other Algorithms

Building upon these foundational algorithms, several variations have been developed to address specific challenges

1. **Weighted Page Rank:** This variant assigns different weights to link based on their importance, considering factors like the relevance of the linking page and the position of the link within the page. Such weighting schemes aim to provide a more nuanced ranking for pages[8].

2. **Trust Rank:** Designed to combat web spam, Trust Rank propagates trust scores from a set of manually identified trustworthy seed pages through the web graph. Pages receiving high trust scores are deemed reliable, improving search result quality by demoting potential spam pages[9].

2.2 Comparative analysis

The following analysis highlights the comparative strengths and weaknesses of Page Rank, HITS, and SALSA algorithms based on critical evaluation criteria, including handling spam, computational efficiency, relevance to user queries, and scalability for large-scale datasets.

2.2.1 Handling Spam

One of the significant challenges for ranking algorithms is mitigating the impact of web spam (malicious attempts to manipulate rankings).

- **Page Rank:** Due to its reliance on the link structure, Page Rank is moderately resistant to spam. However, link farming (artificially creating interconnected pages) can inflate Page Rank scores. Extensions like Trust Rank enhance spam resilience by propagating trust scores from a manually curated set of trustworthy seed pages[9].

- **HITS:** The HITS algorithm is more vulnerable to spam because it operates on a local subset of the web graph. Malicious actors can easily exploit its reliance on hubs and authorities by creating false hubs linked to their target pages[10].

- **SALSA:** SALSA, with its stochastic approach, inherits some robustness from its random walk methodology. However, its effectiveness in handling spam is generally inferior to Page Rank due to its dependence on the bipartite graph, which can be manipulated.

2.2.2 Computational Efficiency

The computational cost of an algorithm is crucial for real-time information retrieval in search engines.

- **Page Rank:** Page Rank is computationally efficient on a global scale due to its ability to pre-compute rankings for the entire web graph. Its iterative eigenvector computation converges quickly in practice, although handling dangling nodes (pages with no outbound links) adds complexity[?] .

- **HITS:** HITS is less computationally efficient than Page Rank because it processes the web graph locally for each query[11]. The iterative updates for authority and hub scores must be recomputed for every query, making resource-intensive[10].

- **SALSA:** SALSA balances computational demands better than HITS by combining global and local graph analysis. However, its reliance on separate random walks for hubs and authorities increases the complexity compared to Page Rank[12].

2.2.3 Relevance to the User Query

An algorithm's ability to deliver results relevant to user queries determines its practical utility.

- **Page Rank:** As a global ranking algorithm, Page Rank does not inherently consider query relevance. While it ranks pages based on importance, additional mechanisms like query-dependent weights are required to align results with user queries[1].

- **HITS:** HITS excels in query relevance since it operates on a local graph constructed from pages matching the query. This focus ensures that both hubs and authorities are closely aligned with the query topic[10].

- **SALSA:** By combining features of HITS and random walks, SALSA achieves a balance between global importance and query relevance. However, it may under-perform compared to HITS in highly specific query contexts due to its hybrid approach[12].

2.2.4 Scalability for Large-Scale Datasets

Scalability is a critical factor for algorithms operating on the vast and ever-growing web.

- **Page Rank:** Page Rank is highly scalable and suitable for large-scale datasets. Its pre-computable nature and efficient implementation in distributed systems like Map-Reduce make it a preferred choice for search engines[13].

- **HITS:** The scalability of HITS is limited by its query-specific computations. Constructing and analyzing local sub-graphs for every query is infeasible for large-scale datasets[10].

- **SALSA:** SALSA's scalability lies between that of Page Rank and HITS. While it does not require global computation for every query, its hybrid approach introduces additional complexity, making it less scalable than Page Rank[12].

3 Conclusion

In conclusion the study of link-based ranking algorithms demonstrates the important role that they play in information retrieval systems. Page Rank, as the cornerstone of Google's search engine, sets a benchmark for evaluating the importance of web pages, based on their connectivity. It managed to influence various modern algorithms that build upon its principles to address some challenges like combating web spam and incorporating personalized search strategies.

While Page Rank, HITS, and SALSA each offer unique strengths, they also highlight the need for adaptability in a dynamic digital environment. For instance, Trust Rank and other advanced models illustrate efforts to mitigate spam's impact, while hybrid methods like SALSA bridge global and local ranking considerations. Despite their utility, these algorithms face limitations, particularly when it comes to handling personalized and context-specific search queries, as well as ensuring efficiency at scale in an ever-expanding web ecosystem.

The role of these algorithms in shaping digital information landscape is indisputable. They have enabled the organization of massive web data, improved user access to relevant information, and inspired continuous innovation in search technologies. Future research could explore integrating artificial intelligence more deeply into these models, enhancing their ability to adapt user needs and addressing emerging challenges in information retrieval.

Bibliography

[1] Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank Citation Ranking : Bringing Order to the Web. The Web Conference. 1999; `https://api.semanticscholar.org/CorpusID:1508503`.

[2] Langville, A.; Meyer, C. *Google's PageRank and Beyond: The Science of Search Engine Rankings*; Princeton University Press, 2011.

[3] Brew, M. The Evolution of the PangeRank Algorithm. **2023**, `https://marketbrew.ai/the-evolution-of-the-pagerank-algorithm`.

[4] Morrison, J. L.; Breitling, R.; Higham, D. J.; Gilbert, D. R. GeneRank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* **2005**, *6*, 233 – 233.

[5] Shah, K.; Singh, S.; Iyer, P. S. Evaluating News Authenticity Using The Hits Algorithm: An Analytical Approach To Identifying Reliable Sources. *Educational Administration: Theory and Practice* **2024**, *30*, 2623–2629.

[6] Khan, M.; Mello, G.; Habib, L.; Engelstad, P.; Yazidi, A. HITS based Propagation Paradigm for Graph Neural Networks. *ACM Transactions on Knowledge Discovery from Data* **2023**, *18*.

[7] Kenneweg, P.; Kenneweg, T.; Fumagalli, F.; Hammer, B. No Learning Rates Needed: Introducing SALSA - Stable Armijo Line Search Algorithm. *arXiv preprint arXiv:2407.20650* **2023**, `https://arxiv.org/abs/2407.20650`.

[8] Xing, W.; Ghorbani, A. Weighted PageRank algorithm. Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. 2004; pp 305–314, `https://ieeexplore.ieee.org/document/1344743`.

[9] Gyöngyi, Z.; Garcia-Molina, H.; Pedersen, J. Combating web spam with trustrank. Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30. 2004; p 576–587, `https://dl.acm.org/doi/10.5555/1316689.1316740`.

[10] Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM* **1999**, *46*, 604–632, `https://doi.org/10.1145/324133.324140`.

[11] Richard Osei Adu, W. A., Klinsman Kwaku Boateng HITS vs. PageRank: A Comparative analysis of Web Search Algorithms. *International Journal of Computer Applications* **2024**, *186*, 32–36.

[12] Lempel, R.; Moran, S. SALSA: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.* **2001**, *19*, 131–160, `https://doi.org/10.1145/382979.383041`.

[13] Dean, J.; Ghemawat, S. MapReduce: simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113, `https://doi.org/10.1145/1327452.1327492`.