Exploration on HITS and PageRank Improvements: Using GenAI to aid stability and experience

By Zhao Qixian U2321752L 17 April 2025

1. Introduction

PageRank, famously employed by Google, assesses pages based on the global structure of the web, while HITS focuses on identifying authoritative and hub pages based on local network structures (Page et al., 1999; Kleinberg, 1999). This report aims to compare PageRank with the original HITS algorithm and 2 of its stable variants proposed by Michael Jordan, Andrew Ng, and colleagues (Ng, Zheng, & Jordan, 2001). The objectives of the report include a detailed literature review of 4 algorithms, performing numerical experiments on both small and larger graphs to validate their convergence properties, and exploring potential enhancements by integrating large language models such as GPT to propose a Generative Search Engine.

2. Literature Review

PageRank Algorithm

PageRank, introduced by Page et al. (1999), evaluates the significance of a webpage based on the probability of a random surfer landing on the page through random navigation. Mathematically, PageRank is calculated using an iterative process where the score of each page is derived from the scores of pages linking to it, adjusted by their number of outbound links. The process incorporates a damping factor to model the behavior of a random surfer who might randomly jump to any page. The primary advantage of PageRank is its robustness derived from analyzing the global network structure, effectively reducing susceptibility to spam and artificially inflated link structures. However, its drawbacks include being computationally intensive for large networks and potentially favoring older, more established webpages, making it less dynamic.

HITS Algorithm

The HITS algorithm, proposed by Kleinberg (1999), identifies two distinct types of pages: hubs, which link extensively to authoritative pages, and authorities, which are frequently linked by hubs. Authority and hub scores are iteratively calculated based on mutual reinforcement between hubs and authorities. HITS excels at clearly distinguishing between hubs and authorities, making it particularly effective for topical searches and query-driven results. Nonetheless, it suffers from certain drawbacks, including susceptibility to topic drift, where irrelevant but popular pages might dominate results, and sensitivity to noisy or spammy links that may destabilize the computed scores.

Stable Variants of HITS by Jordan and Ng

To address stability and convergence issues inherent in the original HITS algorithm, Ng, Zheng, and Jordan (2001) introduced variants aimed at enhancing algorithm robustness. These variants modify the calculation of authority and hub scores through normalization techniques and adjusted weighting schemes, thereby significantly reducing sensitivity to noise and mitigating the topic drift problem. Despite these improvements, these stable variants introduce additional complexity in computation and necessitate careful parameter tuning to ensure optimal performance.

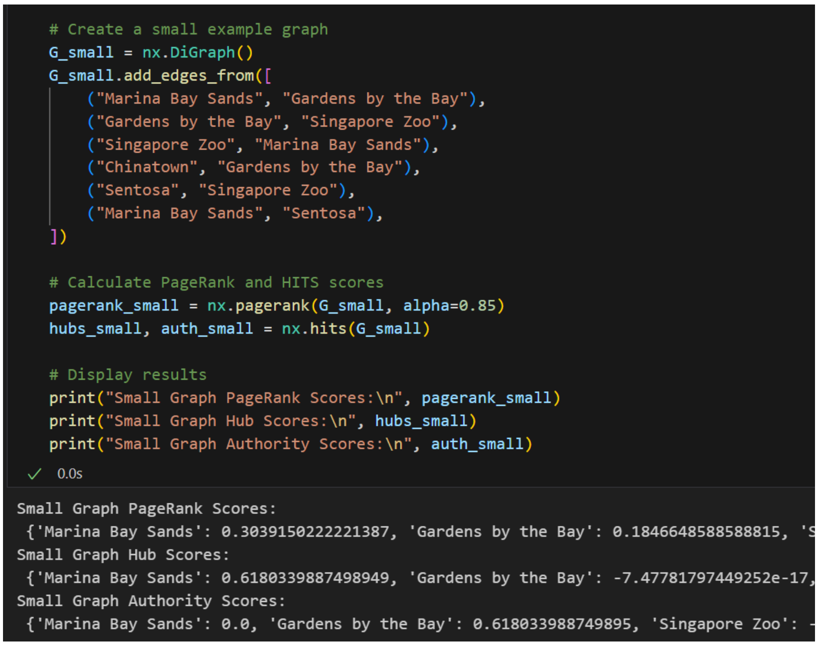
Comparative Analysis of PageRank and HITS

When comparing PageRank and HITS, several factors emerge distinctly. PageRank typically exhibits higher computational complexity due to its requirement for global computation across the entire web graph. Conversely, the HITS algorithm generally involves lower computational complexity, though its original form suffers from stability and convergence issues. The stable variants of HITS proposed by Jordan and Ng alleviate these issues but moderately increase computational demands. Regarding convergence, PageRank guarantees convergence under standard conditions, thanks to its damping factor, while the original HITS algorithm has notable convergence problems, significantly improved by its stable variants. In terms of stability and resistance to manipulation, PageRank is generally more robust against spam but less adaptable to dynamic content. The original HITS algorithm is highly vulnerable to spam and manipulation, whereas its stable variants provide substantial improvements in resilience and adaptability.

3. Numerical Experiments

Small Graph Illustration

Initial understanding was established through a directed graph of Singapore attractions, including hardcoded nodes like Marina Bay Sands, Gardens by the Bay, and Sentosa. PageRank computed consistent node importance based on global connectivity. In contrast, HITS identified authoritative nodes and hubs distinctly, emphasizing local network influence.



```
# Create a small example graph
G_small = nx.DiGraph()
G_small.add_edges_from([
    ("Marina Bay Sands", "Gardens by the Bay"),
    ("Gardens by the Bay", "Singapore Zoo"),
    ("Singapore Zoo", "Marina Bay Sands"),
    ("Chinatown", "Gardens by the Bay"),
    ("Sentosa", "Singapore Zoo"),
    ("Marina Bay Sands", "Sentosa"),
])

# Calculate PageRank and HITS scores
pagerank_small = nx.pagerank(G_small, alpha=0.85)
hubs_small, auth_small = nx.hits(G_small)

# Display results
print("Small Graph PageRank Scores:\n", pagerank_small)
print("Small Graph Hub Scores:\n", hubs_small)
print("Small Graph Authority Scores:\n", auth_small)
```

✓ 0.0s

```
Small Graph PageRank Scores:
 {'Marina Bay Sands': 0.3039150222221387, 'Gardens by the Bay': 0.1846648588588815, 'S
Small Graph Hub Scores:
 {'Marina Bay Sands': 0.6180339887498949, 'Gardens by the Bay': -7.47781797449252e-17,
Small Graph Authority Scores:
 {'Marina Bay Sands': 0.0, 'Gardens by the Bay': 0.618033988749895, 'Singapore Zoo': -
```

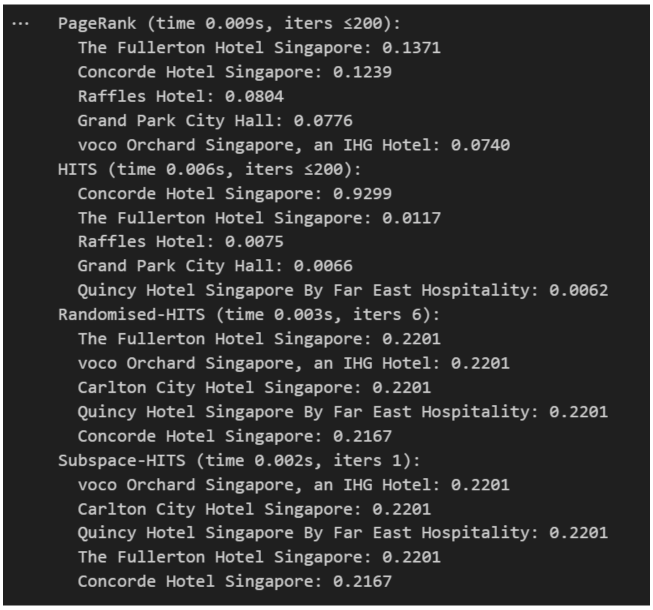
Real World Larger Dataset: Singapore TripAdvisor Reviews

The dataset utilized TripAdvisor reviews of Singapore attractions, creating a directed graph where nodes represented attractions, and edges captured user transition frequencies between attractions. We have 27 nodes 606 edges in this dataset.

Algorithm Implementations and Performance Analysis

The runtime analysis indicated PageRank as relatively slower but stable, while Randomised-HITS and Subspace-HITS achieved faster and stable convergence. Spearman rank correlation analyses revealed strong correlations between PageRank and Randomised-HITS, indicating similar rankings, whereas Subspace-HITS and original HITS provided distinct yet stable results.

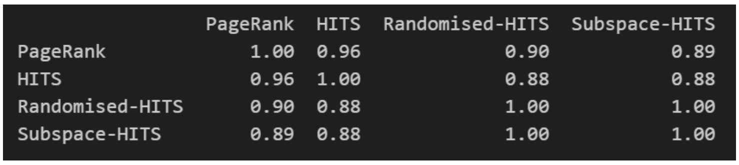
Specifically, the experiments show a clear speed-accuracy trade-off: Subspace-HITS is fastest and converges in essentially one spectral step, closely shadowed by Randomised-HITS, whose teleportation guarantees quick, stable authority rankings. Standard HITS takes marginally longer and yields a highly skewed score distribution that exaggerates the top authority, while PageRank—though still efficient—remains the slowest of the four and produces more evenly spread importance scores. Overall, adding teleportation or restricting the iteration to the dominant eigen subspace both stabilise HITS and bring its results closer together, whereas PageRank retains a broader, more conservative ranking profile.



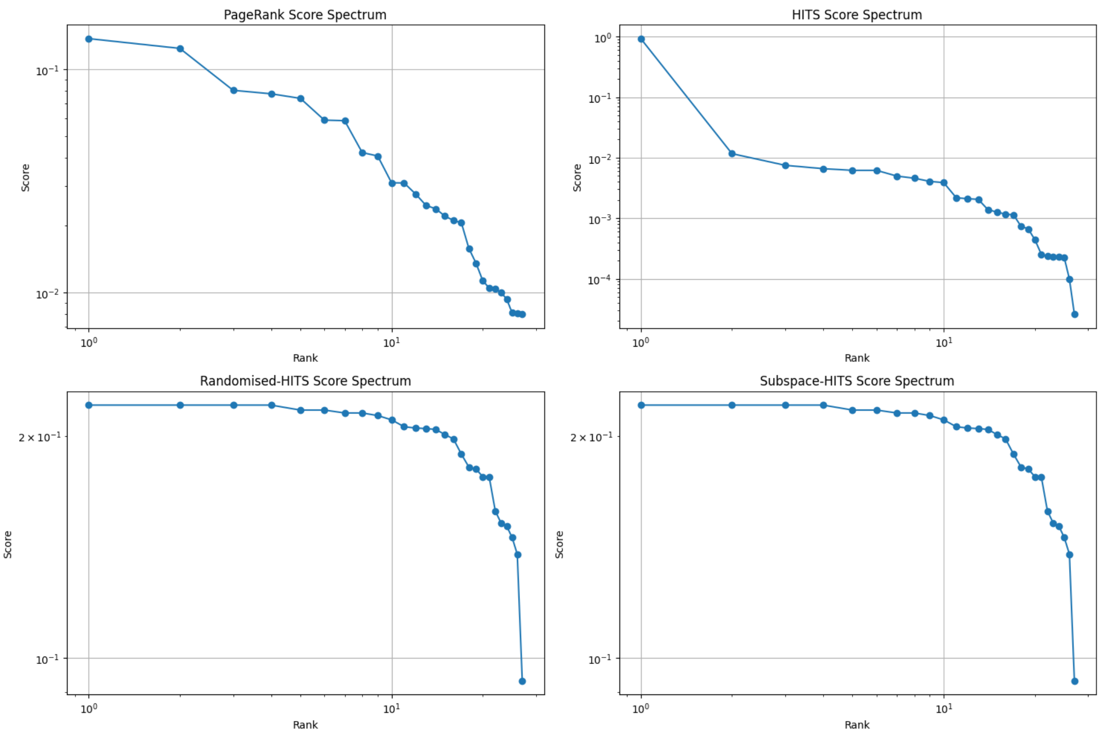
```
…   PageRank (time 0.009s, iters ≤200):
      The Fullerton Hotel Singapore: 0.1371
      Concorde Hotel Singapore: 0.1239
      Raffles Hotel: 0.0804
      Grand Park City Hall: 0.0776
      voco Orchard Singapore, an IHG Hotel: 0.0740
    HITS (time 0.006s, iters ≤200):
      Concorde Hotel Singapore: 0.9299
      The Fullerton Hotel Singapore: 0.0117
      Raffles Hotel: 0.0075
      Grand Park City Hall: 0.0066
      Quincy Hotel Singapore By Far East Hospitality: 0.0062
    Randomised-HITS (time 0.003s, iters 6):
      The Fullerton Hotel Singapore: 0.2201
      voco Orchard Singapore, an IHG Hotel: 0.2201
      Carlton City Hotel Singapore: 0.2201
      Quincy Hotel Singapore By Far East Hospitality: 0.2201
      Concorde Hotel Singapore: 0.2167
    Subspace-HITS (time 0.002s, iters 1):
      voco Orchard Singapore, an IHG Hotel: 0.2201
      Carlton City Hotel Singapore: 0.2201
      Quincy Hotel Singapore By Far East Hospitality: 0.2201
      The Fullerton Hotel Singapore: 0.2201
      Concorde Hotel Singapore: 0.2167
```

Score Distribution and Rankings

The log-log spectra confirm a classic heavy-tailed pattern: importance plummets after the leading nodes, then decays roughly linearly on a log scale. Standard HITS shows the sharpest drop, funnelling almost all authority into a single star, while Randomised- and Subspace-HITS flatten the top of the curve, producing a small plateau of co-equal leaders before the tail sets in—evidence that damping or subspace projection tempers HITS's extremity. PageRank sits between these extremes, retaining a more graduated ranking. The Spearman matrix reinforces the picture: PageRank and HITS remain closely aligned (ρ ≈ 0.96), each diverging modestly from the two stabilised variants (ρ ≈ 0.90), whereas Randomised- and Subspace-HITS are virtually indistinguishable (ρ ≈ 1.00). Practically, this means PageRank or vanilla HITS suits scenarios needing fine-grained differentiation, whereas the stable HITS variants are better when the goal is to surface a tight, definitive shortlist of top attractions.

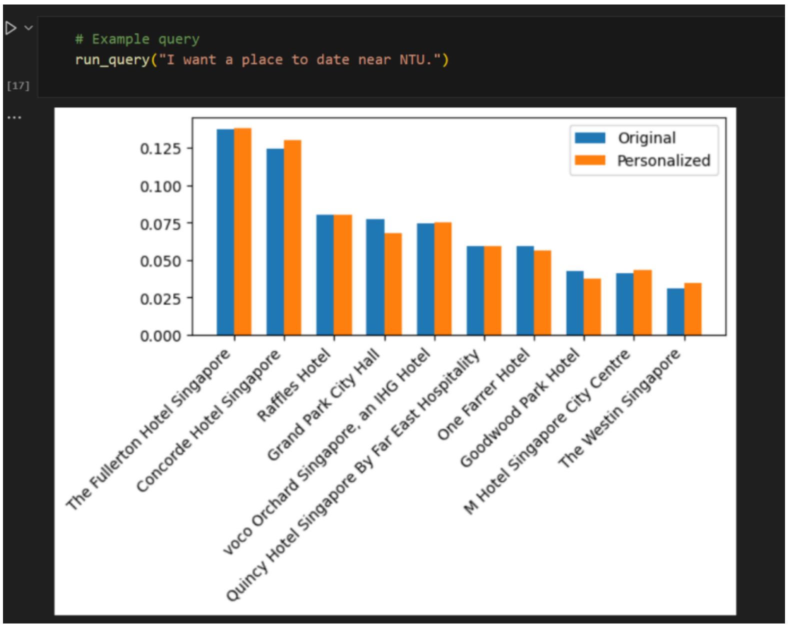


```
                 PageRank  HITS  Randomised-HITS  Subspace-HITS
PageRank             1.00  0.96             0.90           0.89
HITS                 0.96  1.00             0.88           0.88
Randomised-HITS      0.90  0.88             1.00           1.00
Subspace-HITS        0.89  0.88             1.00           1.00
```

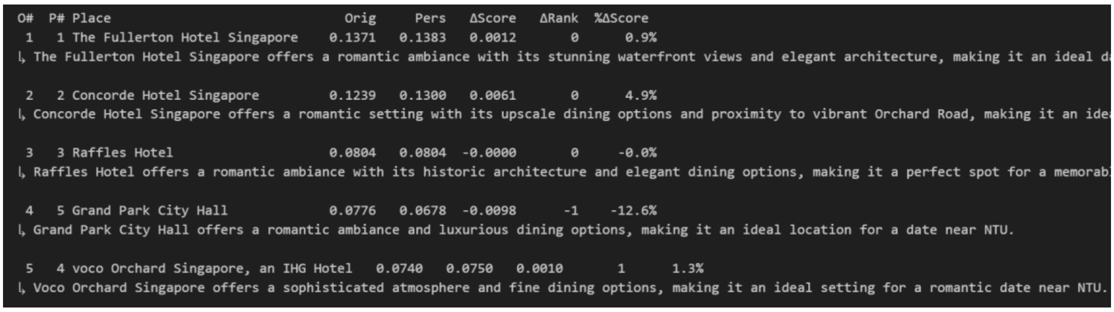


Personalised PageRank with GenAI

To inject real-world meaning and user intent into an otherwise purely topological measure, I augmented PageRank with GPT-powered text embeddings. Each attraction name is first embedded once and cached; at query time, the user's is embedded and cosine-matched against every node. The similarity scores are then re-scaled into a stochastic personalization vector and fed directly into the PageRank iteration (personalization=p). This single line turns the classic random-surfer model into a Personalized PageRank that blends structural importance with semantic relevance.



```
# Example query
run_query("I want a place to date near NTU.")
```



```
O#  P# Place                          Orig     Pers    ΔScore   ΔRank  %ΔScore
 1   1 The Fullerton Hotel Singapore    0.1371   0.1383   0.0012       0      0.9%
↳ The Fullerton Hotel Singapore offers a romantic ambiance with its stunning waterfront views and elegant architecture, making it an ideal d

 2   2 Concorde Hotel Singapore         0.1239   0.1300   0.0061       0      4.9%
↳ Concorde Hotel Singapore offers a romantic setting with its upscale dining options and proximity to vibrant Orchard Road, making it an ide

 3   3 Raffles Hotel                    0.0804   0.0804  -0.0000       0     -0.0%
↳ Raffles Hotel offers a romantic ambiance with its historic architecture and elegant dining options, making it a perfect spot for a memorab

 4   5 Grand Park City Hall             0.0776   0.0678  -0.0098      -1    -12.6%
↳ Grand Park City Hall offers a romantic ambiance and luxurious dining options, making it an ideal location for a date near NTU.

 5   4 voco Orchard Singapore, an IHG Hotel   0.0740   0.0750   0.0010       1      1.3%
↳ Voco Orchard Singapore offers a sophisticated atmosphere and fine dining options, making it an ideal setting for a romantic date near NTU.
```

Towards a Generative Search Engine

A generative engine could therefore blend these signals hierarchically: start with PageRank to filter noise, inject Subspace-HITS for millisecond-level personalization, and overlay Randomised-HITS to refresh topical clusters as conversations evolve. GPT-style models would then craft narratives that explicitly reference each signal (“Highly rated overall, and frequently linked by other date-night blogs”), turning opaque scores into persuasive, context-rich explanations. Beyond text, the hub/authority duality could guide multimodal prompts—surface hub images for inspiration and authority details for planning—while teleportation parameters become controllable dials that users (or agents) adjust in human language. In short, the differing strengths of PageRank, HITS, and their stable variants map naturally onto the responsiveness, transparency, and diversity goals of a truly generative search experience.

4. Conclusion

The comparative analysis of PageRank, HITS, and its stable variants demonstrates each method's unique strengths and limitations. PageRank excels in global robustness, whereas HITS and its stable variants offer nuanced, topic-specific insights with improved convergence characteristics. Incorporating LLMs opens new avenues, allowing personalized, semantically-rich recommendations. The proposed Generative Search Engine concept exemplifies the future of web and recommendation systems, blending traditional algorithmic robustness with advanced language understanding, significantly enriching user interactions.

5. References

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5), 604-632. https://doi.org/10.1145/324133.324140

Ng, A. Y., Zheng, A. X., & Jordan, M. I. (2001). Stable algorithms for link analysis. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 258-266. https://doi.org/10.1145/383952.384025

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/422/

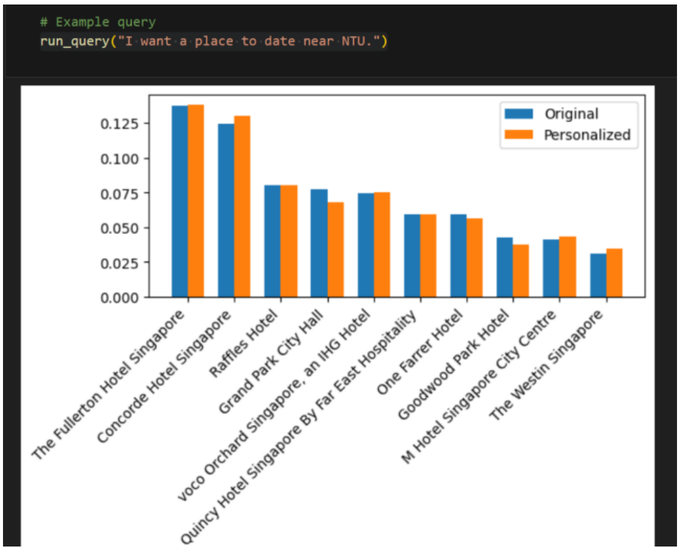
6. Code Implementation:

The full code can be found at my GitHub repo:

Full code can be found here:

https://github.com/ZhaoQixian/Cloud_Computing/blob/main/Assignment2/Pagerank_vs_HITS.ipynb

Before running the notebook, add your OpenAI API key and replace the example run_query("I want a place to date near NTU.") with any query you like. The attraction embeddings are already cached in node_embeddings.pkl, so only the query itself is embedded. run_query() then calls the API once for that embedding and once for each generated place description (up to 10). Using text-embedding-ada-002 plus gpt-4o-mini, the total cost for a typical run should be below one US cent.



```
# Example query
run_query("I want a place to date near NTU.")
```