# Mathematics on Wikipedia: Analyzing the Most Important Mathematical Topics Using PageRank and Visitation Data

Kylie Falkey, Derek Papierski, Colin Prim

**Abstract**

The PageRank algorithm does not take user usage into account in its calculations. This means that pages with incredibly low importance to users can be ranked incredibly high by the algorithm. In an attempt to alleviate this issue, we took a dataset on Wikipedia pages about mathematical topics and performed various transformations to the data to accurately account for user visitation rates. While we were ultimately unsuccessful, we discovered that weighting a graph by visitation of incoming nodes may increase adherence to user-defined importance.

**Introduction**

PageRank is an algorithm that analyzes links between a set of web pages to determine the relative importance of each one in comparison to the others. A page's importance is determined by the importance of pages that link to it. However, it is possible that the importance as defined by the PageRank algorithm does not coincide with the importance of pages to users. A page may be declared important by PageRank which is utterly useless to a user (e.g., a records page for a set of important pages).

To see if this theory proves true, we first compared PageRank importance to visitation rates. We then changed the parameters of the PageRank algorithm in an attempt to better match PageRank importance to visitation. Lastly, we attempted to weight the data by visitation rates to examine the feasibility of a new approach.

**Related Work**

Several researchers' analysis has revolved around the *Wikipedia Math Essentials* dataset. This dataset details cross-references on Wikipedia pages, and as such presents a useful, closed graph structure that is easy and reliable to analyze [4].

Mahmood [5] evaluated this dataset from the perspective of an educator during the COVID-19 pandemic, hoping to determine a reading order for students who are learning online. Research by Mahmood found that this dataset follows the small world phenomenon, meaning that almost every node can be linked to any other node with only a few (an average of 2.79) intermediary steps. Additionally, it was found that the betweenness[1] of a node and its in-degree[2] are positively correlated, though this correlation did not hold with out-degrees.

Thalhammer and Rettinger [4] investigated whether certain links within Wikipedia pages are more important. Using an expanded dataset and various PageRank algorithms, the researchers found no strong relationship between link exclusion and the quality of the PageRank results. They determined that link-structure rankings (such as PageRank) and page-view-based rankings are not strictly reliable by themselves, but instead suggest a combination of the two systems be developed.

---

[1] Betweenness: For between a node is from others. Measured by the number of shortest paths that pass through a node.

[2] In-degree: The number of links to a node.

**Methodology**

A dataset called *Wikipedia Math Essentials* was chosen from UC Irvine's Machine Learning Repository [2]. The dataset contains edges of a directed graph such that an entry of [a, b] represents an edge from Node A to Node B. The dataset also contains weights for each edge, denoting the number of times a link to Page B appears on Page A. Lastly, the dataset includes the daily visits of each page. The dataset was imported and transformed for use according to a Kaggle notebook titled *loading_wikipedia_math_essentials* [1].

To begin our analysis, we performed exploratory data analysis (EDA) on the dataset. The dataset contained 1068 nodes (each corresponding to a Wikipedia Page on a certain mathematical topic), and 27,709 edges. Each node has an average of 25.36 outgoing edges, meaning that each page in the set has an average of 25.36 links to other pages in the set (note that this does not include links to pages that are not in this dataset). Several pages had significantly more outgoing edges (most notably 'Mathematics,' which has 654 outgoing edges) or ingoing edges ('Logarithms,' has 143 ingoing edges). Betweenness and centrality were also measured, and will be discussed later. The 10 pages with the highest number of total visitations were: '*Normal distribution*,' '*Standard deviation*,' '*Fibonacci number*,' '*Pi*,' '*Roman Numerals*,' '*Prime number*,' '*Poisson distribution*,' '*Mathematics*,' '*Golden ratio*,' '*Taylor series*.'

Following this, we ran NetworkX's PageRank algorithm [3] on the dataset with edges default edge weights. The algorithm found the most important pages to be as follows: '*Rotations in 4-dimensional Euclidean space*,' '*24-cell*,' '*Symmetric rotation*', '*Branch point*,' '*Queue (abstract data type)*,' '*Lambert W function*,' '*Duality (optimization)*', '*Elliptic partial differential equation*,' '*10*,' '*Bucket Sort*.' We then tried running the PageRank algorithm with a multitude of parameter changes. The most notable change was increasing the error tolerance (part of the measure of convergence) to 0.01[3]. This changed the set of important pages to be: '*Tuple*,' '*Fibonacci number*,' '*Eigenvalues and eigenvectors*,' '*0*,' '*Logarithm*,' '*Monty Hall problem*,' '*Bipartite graph*,' '*Laplace transform*,' '*Cartesian product*,' '*1*.'

Lastly, we attempted to change the weighting system to account for visitation. Weights were applied to edges based on (a) the total visitation of the incoming node, (b) the total visitation of the outgoing node, (c) the average of the total visitations of the incoming and outgoing nodes, (d) the normalized[4] visitation of the incoming node, (e) the normalized visitation of the outgoing node. System (a) produced the closest to accurate results, being: *'Dodecahedron*,' '*Queue (abstract data type)*,' '*Hamiltonian path*,' '*10*,' '*Bucket sort*,' *'Decagon*,' '*Heap (data structure)*,' '*Fibonacci number*,' '*Chi-square distribution*,' '*Exponential distribution*.' The results of all experiments can be found on the attached spreadsheet.

[3] This was done using the experimental weighting system (a) described in the next paragraph.

[4] Normalized: Total visitation for a given page minus the average total visitation for all pages.

**Experimental Discussion**

We found the result of the PageRank algorithm unexpected. Several pages, most notably '*Rotations in 4-dimensional Euclidean space*' and '*24-cell*[5],' seemed unimportant from a human perspective, but were deemed as the two most important by the algorithm. It is believed that this is due to the fact that these pages had the highest eigenvector centrality of all pages in the graph. These pages were frequently the two most important pages, however, weighting the graph by visitation of incoming nodes (system (a)), seemed to negate this issue.

We were unable to produce an algorithm that was more than 10% similar to the set of highest visited pages. We thought at first that this could be because of unexpectedly high connectivity, but the graph is actually less than 5% connected, so we do not suspect this was the issue. It could instead be that the structure of the graph causes issues. The PageRank algorithm does not account for page length. Longer pages are likely to have more outgoing links, thereby reducing their overall importance in the calculations [4]. We believe that this is the root cause of the inaccuracy that we are facing.

**Contribution**

All three group members participated actively in the project. Our contributions were as follows:

- Kylie
    - Comparing parameter combinations
    - Visual Presentation
- Derek
    - Exploratory Data Analysis (EDA)
    - Initial comparisons of PageRank importance and visitation
- Colin
    - Re-weighting graph by visitation
    - Written report

**Conclusion**

We attempted to better align the PageRank algorithm with inferred user-defined importance (determined by the total number of visits). By weighting the graph by visitation, we obtained some promising results that began to align the algorithm with our expected results, however, the results were not accurate enough to be satisfactory. In the future, we plan to alter the algorithm to accommodate for the lengths of pages as well. We hope to improve our algorithm as much as possible to hasten the advancement of the data science community.

**References**

[1] A. FranciscoPalm, "loading_wikipedia_math_essentials," 20 December 2021. [Online]. Available: https://www.kaggle.com/code/mapologo/loading-wikipedia-math-essentials/notebook. [Accessed 1 March 2024].

[2] UC Irvine Machine Learning Repository, "Wikipedia Math Essentials," 19 April 2021. [Online]. Available: https://doi.org/10.24432/C5V32F. [Accessed 1 March 2024].

[3] NetworkX Contributors, "Reference," NetworkX, 4 April 2024. [Online]. Available: https://networkx.org/documentation/stab

[5] A 24-cell is a 4-dimensional object.

le/reference/index.html. [Accessed 21 April 2024].
[4] A. Thalhammer and A. Rettinger, "PageRank on Wikipedia: Towards General Importance Scores for Entities," 20 October 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-47602-5_41. [Accessed 22 April 2024].
[5] S. S. Mahmood, "Studying the Wikipedia Math Essentials Pages using Graph Theory Metrics," International Journal of Advances in Soft Computing and its Applications, vol. 14, no. 1, 2022.