Wikispeedia, a study about user navigation behavior within information networks

Frederico Prazeres
Faculdade de Ciências da Universidade de Lisboa
fc56269@alunos.fc.ul.pt

Ricardo Sobral
Faculdade de Ciências da Universidade de Lisboa
fc56332@alunos.fc.ul.pt

José Oliveira
Faculdade de Ciências da Universidade de Lisboa
fc61009@alunos.fc.ul.pt

ABSTRACT

Understanding user navigation behavior within information networks was a topic we (the authors of this article) found interesting and worth studying. Because of that, this article investigates navigation patterns using the SNAP Wikispeedia dataset, focusing on paths players take between articles. The dataset comprises of player-generated paths where individuals navigate from one Wikipedia article to another. We aim to uncover the most traversed routes, significant nodes, and patterns in player behavior.

Utilizing network analysis techniques and graph representations, the study examines popular paths, shortest routes, and identifies central nodes within the navigation network.

The findings highlight frequent pathways, nodes acting as crucial connectors, and insights into how players navigate through a vast information network.

Through this exploration, the article contributes to the understanding of the Wikispeedia player's behavior.

INTRODUCTION

'Wikispeedia' emerges as a variant of the 'Wiki Game,' a game that has organically evolved within the Wikipedia community. Players, engaging in solitary sessions, are tasked with a unique challenge: navigating between two pre-defined Wikipedia articles or ones chosen by the player, using only the embedded hyperlinks within the articles. The goal is to traverse from the starting article to the destination ('goal') article while minimizing the number of link clicks, allowing users the flexibility of step-by-step backtracking within their navigation journey. [1]

This article aims to: find the game's dataset characteristics, unravel navigation patterns, identify frequently traversed pathways, spotlight pivotal nodes within the network and explain their importance and analyse the time and difficulty aspect of the game.

This article, will, ultimately, contribute to a better understanding of the Wikispeedia game and its player base decision making to finish the objective.

1. ANALYSIS

1.1 Dataset characteristics

1.1.1 About the dataset

We obtained the dataset from a TSV (tab-separated-file) which contains 51318 finished paths from players of the game. These finished paths contain all the nodes (articles) that the players who finished the game successfully used (from the first article to the last).

The dataset extracted from the 'Wikispeedia' game showcases interesting characteristics upon analysis of its largest strongly connected component (Largest SCC). We analyzed the largest SCC to obtain the characteristics of the dataset, therefore, loose nodes were not particularly relevant in this part of the article.

1.1.2 Network Structure

- The dataset contains 4,604 nodes, representing Wikipedia articles.
- The longest shortest path length between any two nodes (or diameter) within this component is 9, indicating a relatively short maximum distance between articles.
- It contains 119,882 edges, signifying navigational links between these articles.
- The largest strongly connected component is well-connected, allowing navigation from any node to any other node within the component.
- The dataset's largest SCC contains 3703 nodes and 58651 edges,

1.1.3 Navigation dynamics

- The average shortest path length among nodes is approximately 3.22, showcasing a relatively efficient navigation pattern.
- The distribution of node degrees within the range of 0 to 500 suggests a broad spectrum of connectivity among articles, with many articles having low degrees of connections and some having great numbers of edges.

1.1.4 Characteristics

The 'Wikispeedia' dataset, particularly its largest strongly connected component, reflects a substantial and well-connected network of Wikipedia articles. With thousands of nodes and tens of thousands of edges, this dataset embodies a dynamic environment where users traverse between a broad variety of articles.

The relatively short average path length and diameter signify efficient navigation pathways, allowing users to move swiftly between articles within a concise number of clicks.

The distribution of nodes (Fig 1.) in the graph suggests the presence of a limited number of nodes, albeit significantly connected, exhibiting notably high degrees. This phenomenon denotes the existence of hubs within the network - nodes boasting an extensive array of edges. These hubs act as pivotal connectors, serving as a bridge between the initial and final nodes. In subsequent sections of the article, particular attention will be devoted to identifying and analyzing these nodes. This pursuit aims to deepen our comprehension of player behavior and the underlying reasoning behind their navigational choices.

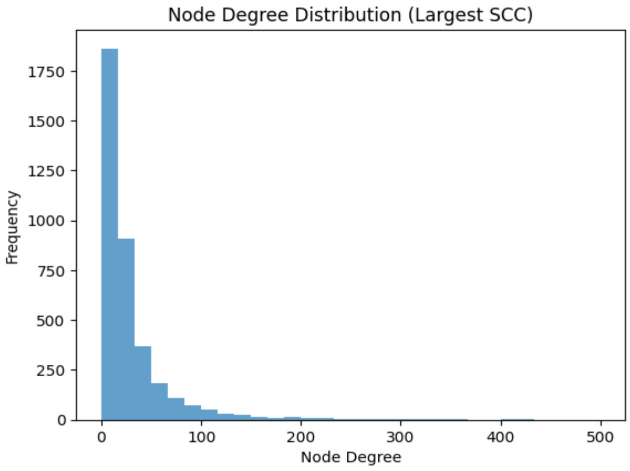


Figure 1 - Node distribution of the dataset (SCC)

1.2 Hubs and their significance

1.2.1 Significance

The discovery and analysis of hubs within the 'Wikispeedia' network have substantial importance in comprehending the dynamics of article traversal and player navigation behavior. Hubs, characterized by nodes with notably high degrees and betweenness centrality, serve as pivotal connectors within the network. Besides understanding the behavior of the player, finding these hubs may also reveal critical articles and topics that act as central points within the Wikipedia information network.

These highly connected nodes play a dual role in the Wikispeedia context. They serve as efficient bridges between articles, enabling players to navigate swiftly between disparate topics and may also signify articles of great relevance or broad appeal. Studying these hubs unveils essential patterns in player behavior and logic within the Wikispeedia game context.

In this paper we'll try to explain why certain articles attain hub status, whether due to their inherent importance, topical significance, or preference of the user.

1.2.2 Hubs and meaning

To uncover hubs within the 'Wikispeedia' dataset, a Python script leveraging libraries such as NetworkX was used. It revealed nodes that act as crucial connectors.

In terms of metrics, we focused on node degree (number of edges a node has) and betweenness centrality (which indicates if a node acts as a bridge to other nodes in the network).

We observed that most nodes identified with the highest betweenness centrality metrics align closely with those boasting the highest degrees within the dataset as shown in figure 2 and 3.

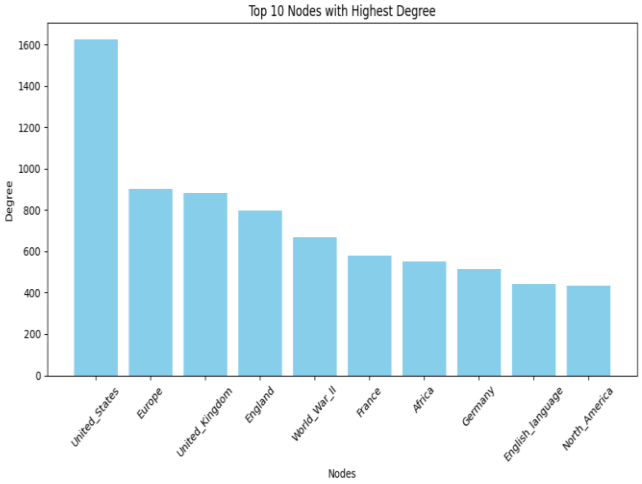


Figure 2 - Top 10 Nodes with highest degree in the dataset

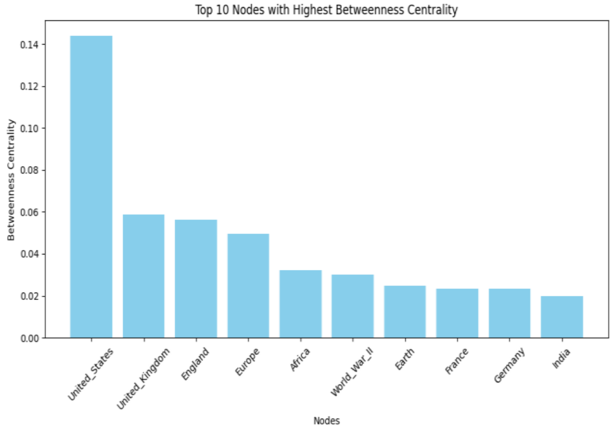


Figure 3 - Nodes with highest Betweenness Centrality

From the figures we can affirm that nodes such as 'United States,' 'United Kingdom,' 'England,' and others not only exhibit substantial connectivity, represented by their high degrees, but also emerge as pivotal bridges, as highlighted by their significant betweenness centrality values. This intersection between high degree and betweenness centrality reflects their importance as hubs within the network.

Their prominence extends beyond mere informational depth. These articles encapsulate diverse narratives, meaning that these specific Wikipedia pages contain a wide variety of sections that allow the player to enter a line of articles that will ultimately lead to the end goal. For example, if a player is stuck in an American singer's page and he's trying to reach the Great Wall of China the logical approach is to enter the United States (country) article, because it will be easy to find the China (country) article and then the Great Wall. This player line of thought is observable in the figures shown above as only one article out of the top 10 does not belong in the country or region category.

In short, the cultural diversity, historical heritage, and geopolitical significance present in an article plays an important role in defining hubs in the Wikispeedia game.

Leaving the Wikispeedia game context, we can also find many of these articles in the all-time most viewed articles of Wikipedia [2], meaning that these articles not only serve as bridges in the game but have a deep cultural importance.

1.3 Path analysis

1.3.2 Path Importance

In order to comprehend the player's logic and behavior, path analysis is absolutely necessary. In this part of the article, we delved into the path component of this dataset.

1.3.3 Most traversed paths

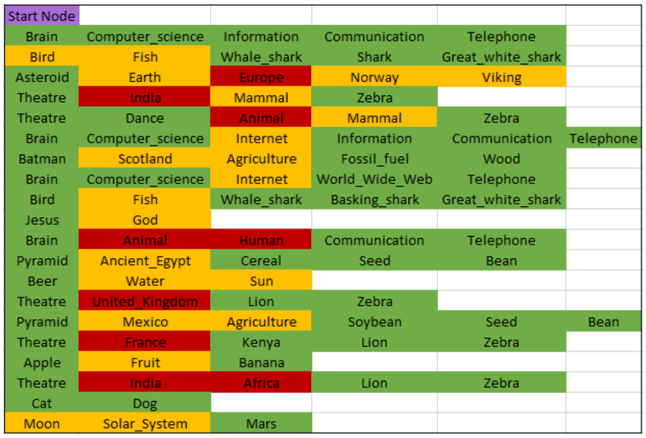
Visualizing the most traversed paths in this dataset is the first (and probably most) important step to understand the Wikispeedia player's logic. It helps in discerning patterns in player behavior, showcasing sequential patterns that players tend to follow.

Using a python script, we found the 20 most traversed paths in the dataset. The first path appeared 144 times and the last 19 times.

These paths were then placed in a table (Fig 4), in which, nodes with very high degrees (possible hubs) were colored with red and nodes with medium degrees colored with yellow. The nodes with relatively low degrees were displayed with the green color.

It's important to note that these paths appear more times in the dataset because the start and end nodes are given to the players quite often, meaning that the assignment of the articles in the Wikispeedia game may not be 100% random.

This also means that many players took the same decisions with these start and end nodes. With these paths appearing so often we can take some conclusions of how the player (in general) plays the game.



Start Node					
Brain	Computer_science	Information	Communication	Telephone	
Bird	Fish	Whale_shark	Shark	Great_white_shark	
Asteroid	Earth	Europe	Norway	Viking	
Theatre	India	Mammal	Zebra		
Theatre	Dance	Animal	Mammal	Zebra	
Brain	Computer_science	Internet	Information	Communication	Telephone
Batman	Scotland	Agriculture	Fossil_fuel	Wood	
Brain	Computer_science	Internet	World_Wide_Web	Telephone	
Bird	Fish	Whale_shark	Basking_shark	Great_white_shark	
Jesus	God				
Brain	Animal	Human	Communication	Telephone	
Pyramid	Ancient_Egypt	Cereal	Seed	Bean	
Beer	Water	Sun			
Theatre	United_Kingdom	Lion	Zebra		
Pyramid	Mexico	Agriculture	Soybean	Seed	Bean
Theatre	France	Kenya	Lion	Zebra	
Apple	Fruit	Banana			
Theatre	India	Africa	Lion	Zebra	
Cat	Dog				
Moon	Solar_System	Mars			

Figure 4 - Top 20 traversed paths

Excluding the [Jesus-God] and [Cat-Dog] paths, we can observe certain behaviors in these paths.

The first one is the search for a great article that will lead the player to a more desirable topic and ultimately the end article. This can be seen especially in the [Theatre-United_Kingdom-Lion-Zebra] path in which the player finds a hub (the United Kingdom node) to switch to a more appropriate topic. We see this pattern in almost every path where the start and end nodes are very distant (topic-wise).

In the other cases where these articles were not so far off from each other (topic-wise), the players did not feel the need to switch to a greater article, since they acknowledged that they already were in the right way to find the end article.

In short, the average player of the Wikispeedia game has two different approaches according to the situation he is currently:

- When in a situation where the topic of the current article is far off from the objective, he actively searches for a great article (a hub in the dataset context) to direct him.
- If he finds himself in an article with relatively the same topic as the objective, he will proceed only to find the end article.

1.3.4 Backtracking

An essential part of the game of Wikispeedia, backtracking is the ability a player must go back to the previous Wikipedia article he's in. This often happens when the player feels he made a bad decision when picking the article and wants to undo said decision.

So far, we have yet to take the backtrack node (represented as '<') into consideration, as it was deemed irrelevant for the most part of the study. However, we felt that not investigating this particular aspect of the dataset would make this path analysis very incomplete.

In this section of the article, we will uncover how frequent this game mechanic is and how important it is to the player. We will try to understand how this mechanic is used and why it may be used.

1.3.4.1 *Backtracking in the dataset*

To uncover the importance of backtracking in the dataset, it was necessary to know its frequency. We discovered that backtracking was present in 8995 paths. This represents 17.5% of all the paths in the dataset. A great quantity of these backtracks were not isolated as some paths contained a great number of backtracks (these could go up to 45!).

From the 17.5% figure we can also conclude that most players were confident in their decision making without ever needing to go back to the previous article.

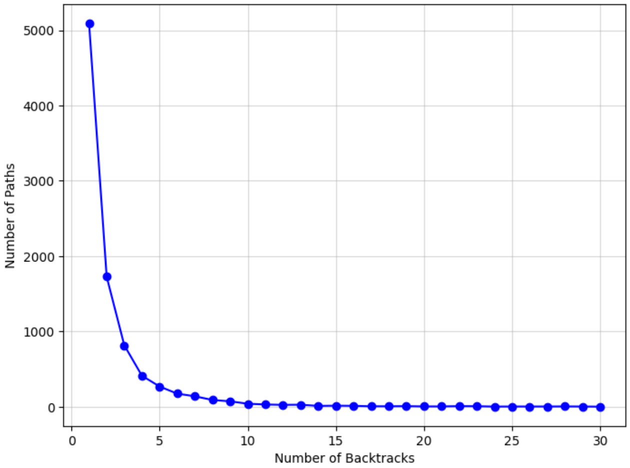


Figure 5 - Backtrack distribution in the dataset

The distribution (observable in Fig. 5) reveals that very few of these paths had more than 5 backtracks. This indicates that most of the players who needed to backtrack did not need to backtrack again many times.

1.3.4.2 *About backtracking*

We have analysed the backtracking distribution, but it is also important to understand how this mechanic may influence the game strategy.

The first step we took in this direction was to investigate the average degree of the nodes before and after backtracking occurred. We proceeded this way to understand if the average player backtracked to reach a more connected node.

We also investigated if the player hesitated in the game. In other words, we found how many times the players who backtracked went exactly to the same article again.

This path is an example of this occurrence: [London-Europe-Backtrack(<)-Europe].

After this investigation, these were our findings:

- The average degree of the nodes before and after backtracking were 119 and 166, respectively.
- The number of same nodes present before and after a backtrack was 1332.

What could this mean in terms of player behavior? We concluded the following:

- Players backtracked to find slightly more relevant articles.
- Occasionally, there was hesitation among the players as 2.5% of the paths in the dataset contained a backtrack between two equal nodes. In these paths in particular, the player may have opted to go back to the same article he backtracked because there weren't any good options to proceed in his search.

1.4 Temporal and difficulty analysis

In this section, we studied the time and difficulty of the game. Analysing these two components is important and it can lead to interesting conclusions.

It may also be important to address that the difficulty component is solely based on player input. The player after each round could evaluate the difficulty of the round. This difficulty ranged from 1 (easy) to 5 (brutal). Because the difficulty evaluation was optional, many paths in the dataset did not contain this component. As a result of this, only 55.54% of paths were evaluated for their difficulty.

In this section of the article, we analysed both aspects individually and, in the end, we performed a joint analysis.

For better readability: every circle in the scatter plots below represents a round of the game Wikispeedia.

1.4.1 *Difficulty Analysis*

The first step to take to gain a better understanding of the game's difficulty is to observe how players classified the paths they took.

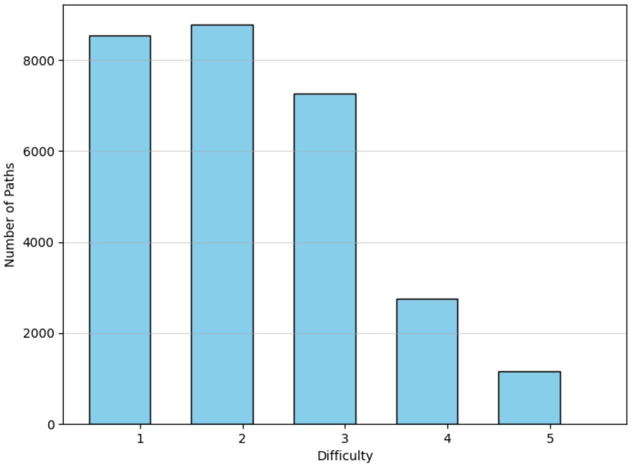


Figure 6 - Difficulty distribution in the dataset

For that, we created a histogram (Fig. 6). From the plot, we can observe that most players did not find the assigned path very hard to achieve. The great majority didn't to evaluate the round higher than 3 and only 7% of the paths in the dataset contained a rating of 4 or 5. The average difficulty found was 2.27.

We also felt that it was necessary to compare the path length and the difficulty to see if there was any correlation between both. By looking at Fig. 7 we can observe just that.

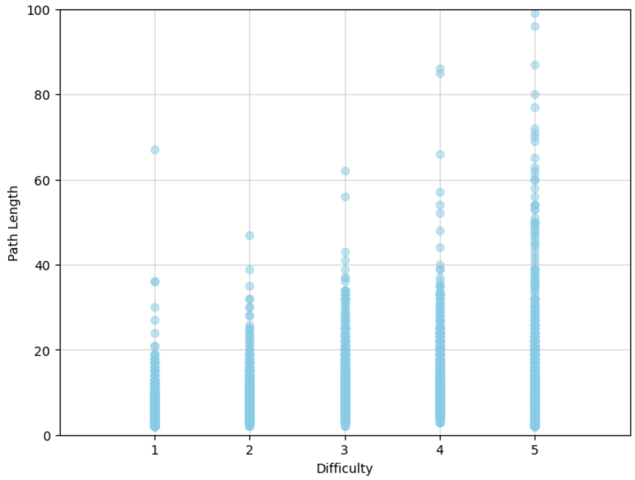


Figure 7 - Comparison between path length and difficulty

There is an apparent correlation between difficulty and path length. The greater the path length, the more probable it is to be classified with a higher difficulty. This is not an unexpected conclusion; it is logical that a player who has gone through many articles to reach the objective would classify this path as having a high difficulty.

1.4.2 Temporal Analysis

The time was one of the columns in the dataset. It was registered in seconds. The average time throughout the dataset was 2 minutes and 38 seconds.

As we did earlier with difficulty, we compared the path length and time taken to complete the objective.

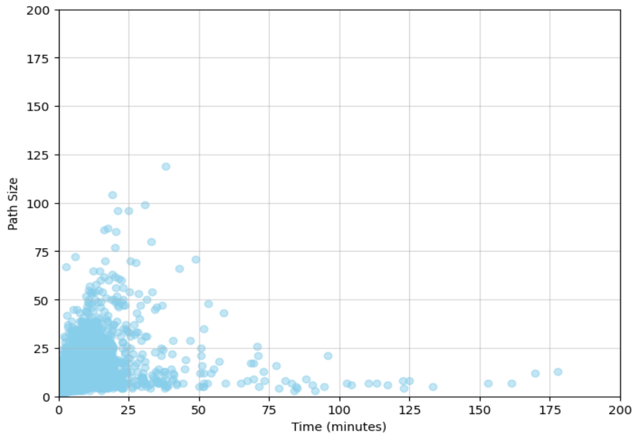


Figure 8 - Comparison between path length and time taken to complete the objective

Before looking at Fig. 8 we certainly expected a more linear correlation between the path length and time taken to complete the task. That wasn't the case, the time taken didn't rise linearly with the path length.

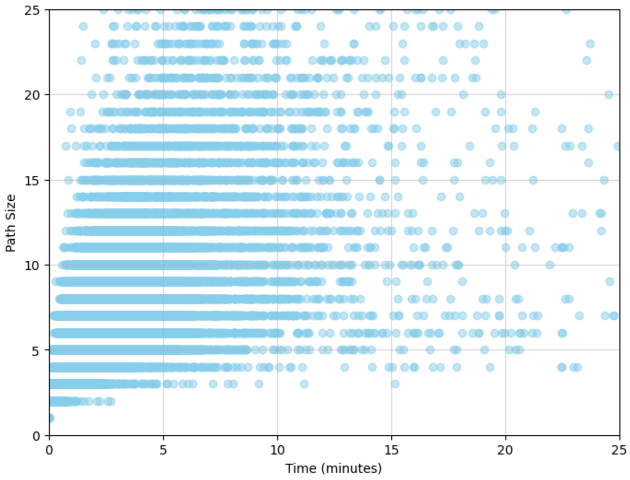


Figure 9 - Figure 8 Enhanced

Zooming in on the most critical part of the graph, we still can't observe any correlation, this meaning that a greater path does not imply a greater time to perform it.

1.4.2 Time and difficulty

In this section of the article, we investigated if the time and difficulty of the paths had any relation with each other.

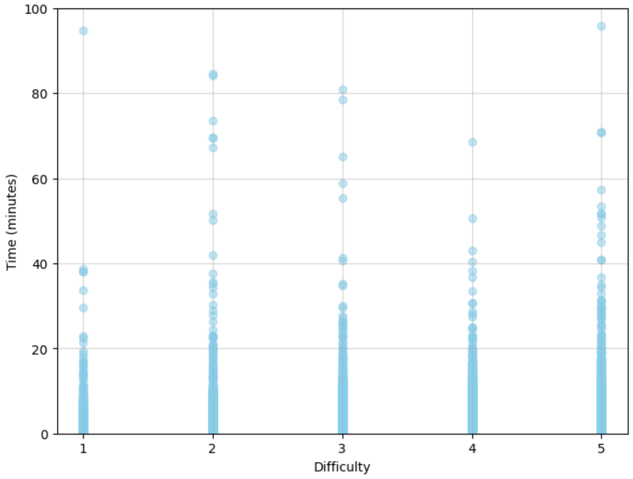


Figure 10 - Time and difficulty analysed

And indeed, they have. It is quite subtle, but we can observe a certain growth in time when the difficulty also rises. From this we can conclude that a harder path should take longer to complete.

CONCLUSION

In this article we provided insights into the Wikipediae game and its player behaviour. We explored the dataset's structure, highlighting a well-connected network with a relatively short average path length, identified hubs crucial for navigation (and their meaning) and revealed nodes with high degrees and betweenness centrality, often representing pivotal connectors and central points in the network.

A path analysis was performed to identify frequently traversed paths, revealing player tendencies to switch articles to more central nodes when topics are distant. We also analysed the backtracking component of the dataset, uncovering its frequency and impact on player strategies.

We studied the player-rated difficulty, and time taken to complete paths, showing correlations between path length and difficulty while highlighting a subtle relationship between that same difficulty and time.

With this article we hope to have led the reader to a better understanding of the Wikispeedia game and player.

REFERENCES

[1] Robert West, Joelle Pineau, and Doina Precup. 2009. Wikispeedia: an online game for inferring semantic distances between concepts. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1598–1603.

[2] Most popular Wikipedia pages https://en.wikipedia.org/wiki/Wikipedia:Popular_pages