The Hyper-linked Imperative: Wikipedia's Journey to Philosophy

LT6 - Castillo, Jasmin, Petilo, Sangalang

October 29, 2024

Abstract

This study investigates a notable phenomenon in Wikipedia's link structure wherein following the first valid hyperlink in article texts predominantly leads to the "Philosophy" page. The criteria for a valid hyperlink is the following: (1) not within a parenthesis, (2) not a link to either a meta page, a page outside Wikipedia, or a broken link, and (3) not an in–page citation. To demonstrate this phenomenon, we developed a web crawler to systematically analyze this pattern.The crawler follows the first link of a random article until it reaches the Philosophy page, encounters a terminal page with no valid links, or detects a loop.

Our methodology employs Selenium with Python to traverse article connections while adhering to the criteria for hyperlink validity. For the English Wikipedia, results indicate that approximately 67% of 1,000 random pages lead to Philosophy; 23% encounter a cyclic reference; and 10% lead to invalid links. The median degree of separation between a random article and the Philosophy page is 18. This pattern suggests an inherent hierarchical structure in Wikipedia's knowledge organization, where specific concepts gradually lead to more general philosophical principles. However, the Cebuano Wikipedia has a different network structure. In comparison, approximately 69% of articles terminate after encountering a loop, while 31% end up on invalid links.

1 Introduction

Human knowledge is an interconnected, complex web of concepts, ideas, and relationships. A naive but convincing demonstration of this is the behavior of Wikipedia articles wherein any random page will most likely be connected to the Philosophy page, much like the branches of a tree converging toward its central trunk. Using a custom web crawler built with Selenium and Python, we would like to illustrate this interconnectedness through the traversal of random Wikipedia articles and analyzing its path towards the Philosophy page. We limit the network to follow only the first valid hyperlink of each page. Our definition of a valid hyperlink excludes those within parentheses, meta pages, external links, broken links, and in-page citations. We also conduct the same study for the Cebuano Wikipedia and observe the differences in network structure.

Wikipedia is a collaborative digital encyclopedia where volunteer editors, called Wikipedians, create and maintain free content using the MediaWiki platform. Its main purpose is to benefit readers by presenting information on all branches of knowledge. With its launch dating back to January 15, 2001, the English Wikipedia has accumulated almost

seven million total articles with over 800 admins maintaining these pages. Alternatively, the Cebuano Wikipedia was launched in June 2005 and has accumulated a little over six million articles initially created through automatic programs. Here, we define articles as referring to a page containing an encyclopedia entry which is different from a page (encompasses all the material on Wikipedia, including encyclopedia topics, talk pages, documentation, and special pages such as Recent Changes).

In this case, we are drawing an analogy between the encyclopedic nature of Wikipedia and the network of human knowledge. To prove the interconnectedness of human knowledge, we build a custom Wikipedia crawler that traverses articles until it reaches the Philosophy page. We will then analyze the network structure and form conclusions based on degree of separation, termination cases, etc.

2 Methodology

To analyze Wikipedia's network structure, we employed a custom web crawler using Selenium with Python to find the path patterns from random articles to the Philosophy page. The study focused on two distinct Wikipedia versions: the English Wikipedia and the Cebuano Wikipedia.

2.1 Data Collection

We developed a custom web crawler using Selenium with Python to automatically traverse Wikipedia's article network. The crawler was designed to:

1. Start from a randomly selected Wikipedia article
2. Identify and follow the first valid hyperlink in the main text
3. Continue this process until reaching one of three terminal conditions:
    - Arriving at the Philosophy page
    - Encountering a page with no valid links
    - Detecting a cyclic reference (loop)

2.2 Link Validity Criteria

To ensure consistency and meaningful results, we followed a strict criteria for determining a link's validity. A hyperlink was considered valid only if it met the following conditions:

1. Located outside parentheses and not italicized
2. Not pointing to a meta page
3. Not an external link
4. Not a broken link
5. Not an in-page citation

2.3 Sampling and Analysis

We analyzed a sample size of 1,000 random articles and each article's path was tracked and recorded. We then measured the degree of separation between the starting article and the terminal page. Finally, we calculated the following metrics:

1. Percentage of paths leading to the Philosophy page, cyclic references, and invalid or terminal pages
2. Median path length for successful routes to the Philosophy page
3. Most common last three links leading to the Philosophy page

2.4 Data Processing and Visualization

To ensure data quality and reliability, we implemented error handling to manage network issues and page loading failures. We also verified link validity before following each hyperlink. Additionally, we maintained a history of visited pages to detect cycles. Finally, we logged all crawling sessions for verification and analysis purposes.

2.5 Concurrent Execution

To optimize the data collection process and reduce overall execution time, we implemented concurrent execution using Python's concurrent.futures module. The concurrent crawling system was designed with the following parameters:

- Utilized `ThreadPoolExecutor` with `max_workers=15`
- Processed articles in batches with `batch_size=30`

3 Results

The analysis of the link paths from random articles to the Philosophy page in both the English and Cebuano Wikipedia revealed significant patterns in the network structure as well as interesting differences between the two.

Table 1 summarizes the results of the analysis on the network structure.

Table 1: Detailed Comparison of Wikipedia Link Analysis

	English Wikipedia	**Cebuano Wikipedia**
Sample Size	1000	1000
Success Rate	67.2%	0%
Average Path Length	18	0
Invalid Links	22.8%	30.9%
Loop Occurrence	10.0%	69.1%

3.1 English Wikipedia

As seen in Figure 1, the majority of the articles in the English Wikipedia reach the Philosophy page at 67.2%. The next most common reason for termination is reaching invalid links; loops are the least common reason for termination.

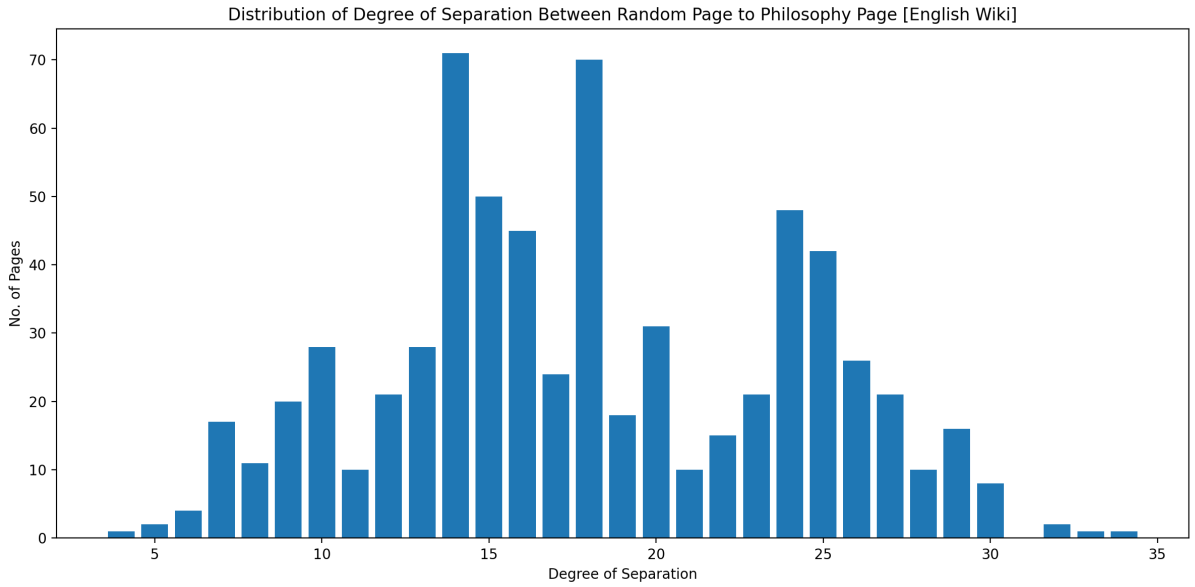

Figure 1: Count by termination reason for English Wikipedia

The median degree of separation in the English Wikipedia is 18, and less than 1% of the random articles have a degree of separation of 6. Taking into account the size of the English Wikipedia, which is 6,902,357 articles, we find that an estimated total of Wikipedia articles with degree 6 is only 27,609.

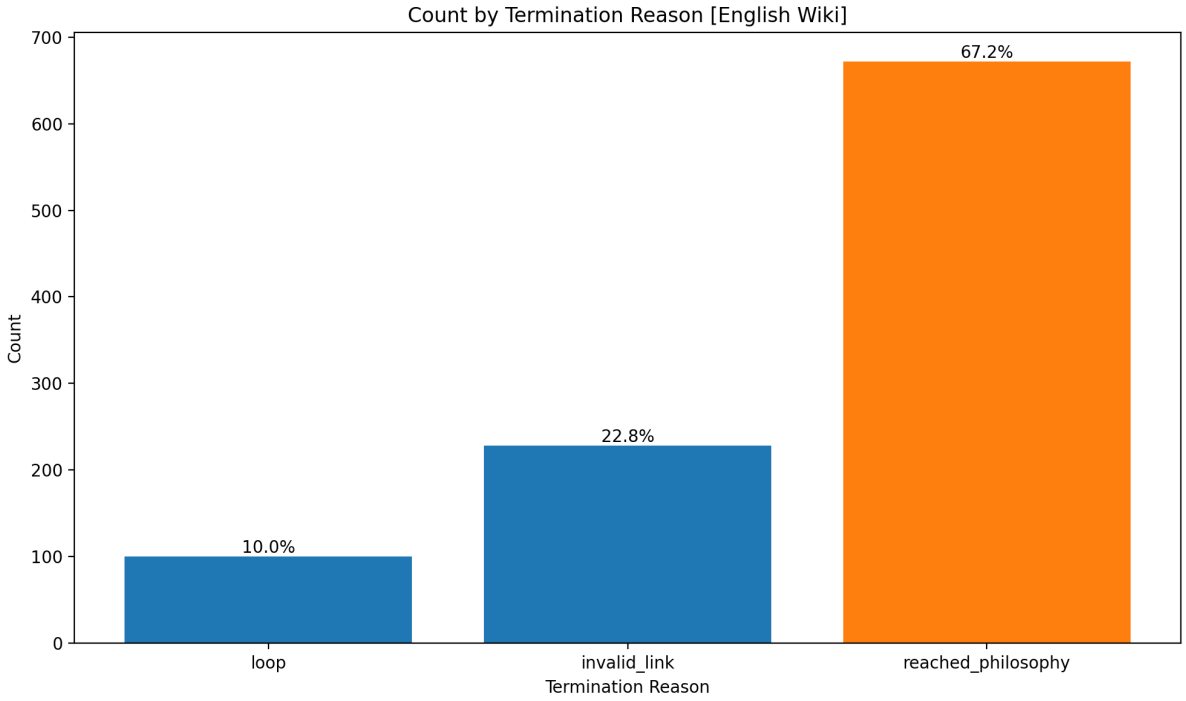

Figure 2: Distribution of Degree of Separation Between Random Page to Philosophy Page in the English Wikipedia

The analysis of English Wikipedia's link structure reveals a clear hierarchical pattern. A significant majority (67.2%) of articles eventually lead to the Philosophy page, demonstrating a natural progression from specific topics to philosophical concepts. The typical path length is 18 steps (median degree of separation), indicating that knowledge tends to become increasingly abstract over multiple transitions. Notably, very short paths are rare – less than 1% of articles reach Philosophy in just 6 steps, which translates to approximately 27,609 articles out of the total 6.9 million. When articles don't reach Philosophy, this is most commonly due to invalid links rather than cyclic references, suggesting a generally well-structured link hierarchy in the English Wikipedia.

3.2 Cebuano Wikipedia

In comparison, the Cebuano Wikipedia never reaches the Philosophy page as illustrated in Figure 3. However, the most common termination reason in this case is that most pages encounter a cyclic reference.

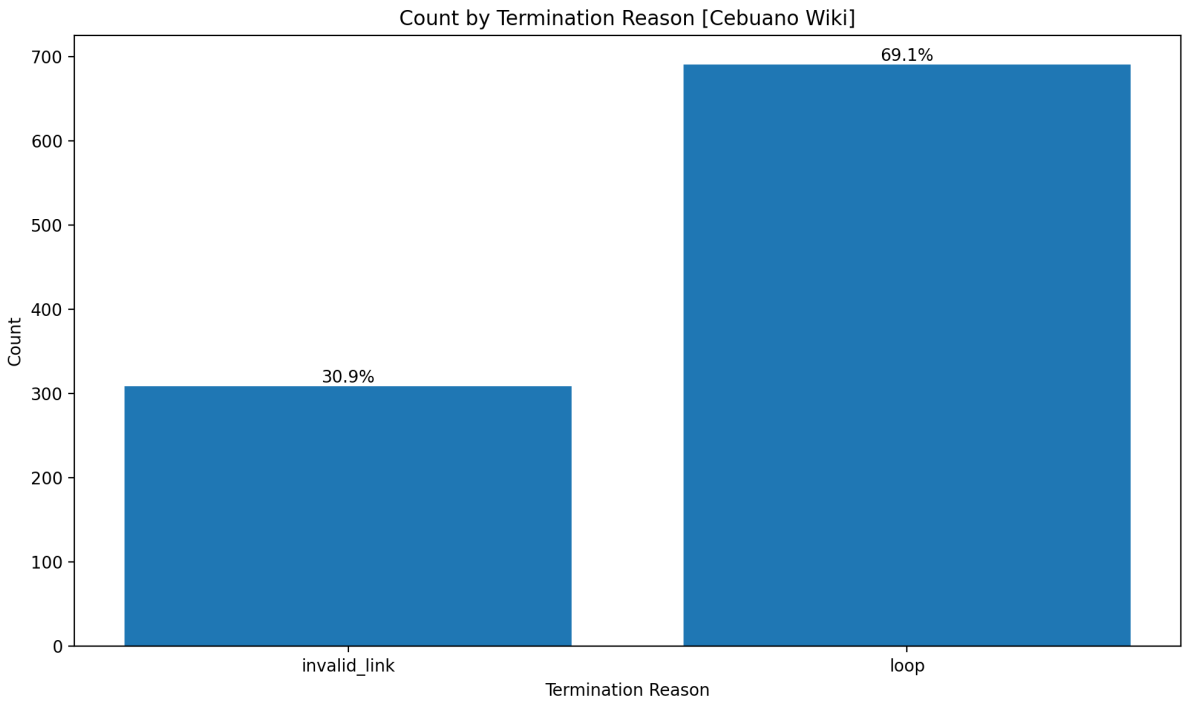

Figure 3: Count by termination reason for Cebuano Wikipedia

Upon inspecting the terminal links in the Cebuano Wikipedia, we found that 78.5% of the random articles end up terminating on the Pransiya page as shown in Figure 4.

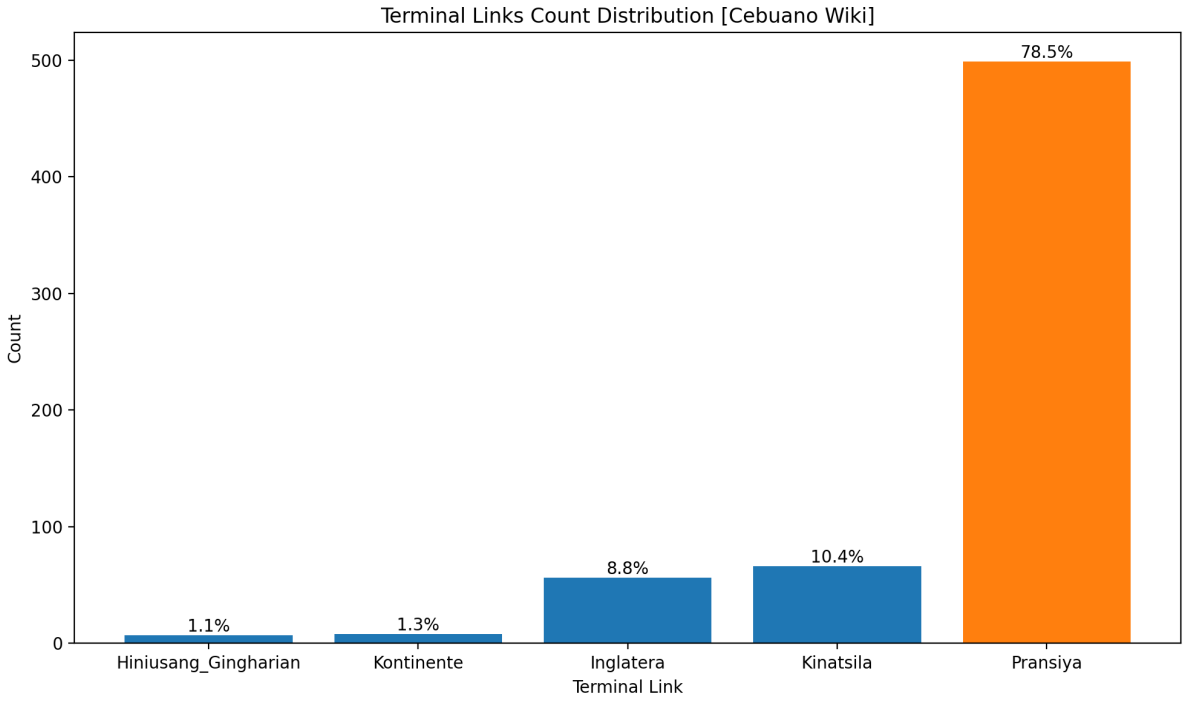

Figure 4: Count distribution of terminal links in the Cebuano Wikipedia

Finally, in Figure 5, we visualize the number of links it takes before the traversal is terminated. The median number of links before it reaches termination is 10.

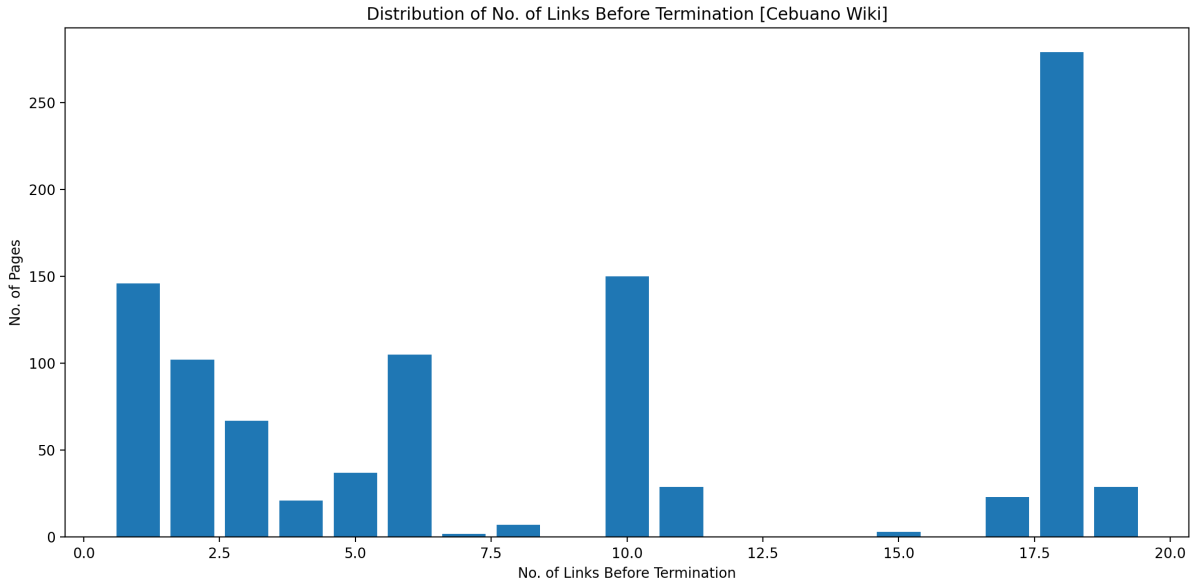

Figure 5: Distribution of number of links before termination in the Cebuano Wikipedia

The Cebuano Wikipedia demonstrates a markedly different link structure from its English counterpart. Rather than leading to Philosophy, articles predominantly become trapped in loops, with 78.5% of all paths ultimately terminating at the Pransiya page. This suggests a less hierarchical and more cyclical organization of knowledge in the Cebuano Wikipedia.

4 Conclusion

Our analysis reveals fundamental differences in the knowledge organization between the English and Cebuano versions of Wikipedia, highlighting how language-specific variations can significantly impact the structure of digital encyclopedias. The English Wikipedia demonstrates a clear hierarchical organization, with 67.2% of articles eventually leading to the Philosophy page through an average of 18 steps, suggesting a natural progression from specific topics to more abstract philosophical concepts. This pattern indicates an implicit organizational structure where knowledge gradually converges towards fundamental philosophical principles.

However, the Cebuano Wikipedia presents a markedly different structure. Instead of converging on the Philosophy page, 78.5% of articles become trapped in cyclic references, predominantly terminating at the Pransiya page. This striking contrast suggests that smaller Wikipedia versions may develop different organizational patterns, possibly influenced by their size, community dynamics, and cultural context. The concentration of links leading to a single country page, rather than an abstract concept, might reflect differences in content creation patterns and editorial practices between the two Wikipedia versions.

The contrasting structures between English and Cebuano Wikipedia demonstrate that knowledge organization in digital encyclopedias varies significantly across languages, suggesting that cultural and linguistic factors strongly influence how information is interconnected in collaborative platforms.