CSC111 Project Report: Mapping Wikipedia Article Structure

Benjamin Fitzgerald, Samuel Reeder, Vedant Swamy, Raymond Zeng-Xu

Sunday, April 2, 2023

Introduction

Wikipedia is an online encyclopedia that contains more than six million English articles. It is actively maintained and expanded by tens of thousands of volunteers called "editors", who work on articles according to general guidelines. The encyclopedia is a highly popular source of knowledge: Wikipedia.org is the 6th most visited website in the world. Due to its size and importance, the organization of information on Wikipedia is of great academic interest: Wikipedia is (in some sense) a reflection of collective human knowledge. Therefore, insights into the structure of Wikipedia articles can reveal valuable information about how knowledge is organized in society, or how different categories of knowledge relate to each other. A well-known example of this is the phenomenon termed "getting to philosophy". Nearly all Wikipedia articles contain at least one *hyperlink*: a word in a Wikipedia article that links to another article, (possibily itself). If the first non-italicized, non-parenthesized hyperlink in a Wikipedia article is chosen, and then the first hyperlink in that article, and so forth, it has a pretty good chance of getting to the Wikipedia article on Philosophy. Actually, more than *90 percent* of Wikipedia articles will eventually lead to the philosophy article if this algorithm is performed. This phenomena occurs because it is standard practice on Wikipedia to start an article with a general description of a topic. This general description usually includes a broader category that the topic belongs to (for example, the topic "chemistry" belongs to the broader category "science"). Intuitively, philosophy is a very broad category, since it is the foundation of every area of study. So in this way, Wikipedia "reflects" how we internally categorize information, i.e. chemistry $\rightarrow$ is a subset of physics $\rightarrow$ is a subset of math $\rightarrow$ is a subset of philosophy. This raises an interesting question: what other articles have a large portion of Wikipedia link to them eventually? What other relationships can we find by exploring hyperlinks in an iterative way? On the English Wikipedia, nearly all articles are placed in specific categories: (History and events, geography and places, human activities, etc.). Each article can belong to multiple categories. Thus, each article on Wikipedia is associated with a set of categories that it belongs to. As a result, we can also ask: what is the relationship between the size of an article, its connections, and its categories? For example, is there an correlation between the size of an article (it terms of its word count) and how many categories it belongs to? This leads us to the goal of this project:

1. **Graphically represent connections between Wikipedia articles by hyperlinks, using categories and article size.**
2. **Locate "general" articles that many articles eventually hyperlink to.**
3. **Draw conclusions about how knowledge on Wikipedia is organized according to article hyperlinks.**

Note: for the third goal, the way data is organized by hyperlinks is likely not the same as the *article categories* mentioned above.

Datasets

We obtained all our data from the MediaWiki API. It provides the hyperlinks going out from an article, the categories articles belong to, and the size of the articles (in terms of word count).

Computational Overview

The *DirectedGraph* Class

We used directed graphs to represent connectivity between Wikipedia articles and other Wikipedia articles that are directly connected by reference through hyperlinks. We preserve most of the common properties of directed graphs

that are present in lectures but supplement them with additional properties and use lists as opposed to sets in some places. Each node on the graph is represented as a connected article.

API

Sending a request using *requests* module's *.get()*, we take an article and make a call to the MediaWiki API to get all desired properties of the article.

Graph Visualization

Once we gather the properties of the parent article, we recursively perform the same above steps for all articles referenced as a hyperlink in the parent, until the maximum recursion depth is reached. We establish edges between the above parent article and each daughter. They are not all added simultaneously. Additionally, we surf the graph to identify whether newly added articles connect to any previous articles and establish those edges. If category filters are specified, it performs the same steps but omits articles that do not adhere to the category rules and restrict recursion through their respective hyperlinks. Importantly, networkx's *spring_layout()* uses the Fruchterman-Reingold algorithm, that minimizes the amount of overlapping edges to produce neater-looking graphs. Plotly's *graph_objects* module provides necessary functions for creating a scatter plot, allowing us to manipulate point size, color, and placement to represent node properties.

Statistics.py

The Statistics.py file contains three functions that take a DirectedGraph instance as input: The first one calculates the vertex with the most number of edges. The second one calculates the vertex with the maximum ratio of output to input edges. The third one calculates the vertex with the maximum ratio of input to output edges. If there is a tie, then all the tied vertices are returned.

Instructions

Simply call the *main* function in *main.py*. You must pass the article title, depth, and what color on the graph should indicate (either category size or article size). The function will return a graph, as well as relevant statistical data pertaining to it. The graph will inclde the passed article, as well as the hyperlinks that leave it, and the hyperlinks that leave those articles, and so forth up to the given depth. A depth of 1 returns the article itself. Be sure that the given article is a valid English Wikipedia article title.

Changes to Project Plan

We graphed the number of categories each article belongs to instead of *which* categories each article belongs to. We also only graphed a very small subset of all English Wikipedia articles (nothing close to the entire thing).

Discussion of Results

1. Graphically Represent Connections

Overall the graphical representation of connections is very successful. Even though only a few thousand articles can be graphed at a time, each graph nicely contains all the information we sought: the connections between articles, how many connections they have, their size (word count) and how many categories they occupy. We Did not represent *which* category they occupy. We can make detailed graphs of any Wikipedia article and thousands of its neighbors.

2. Locate "General" Articles

The graphs and statistical data give us an idea of which articles are general. For example, below is the output of main("History", 2, True, set()).

"History" has a large number of connections compared to its neighbors. So it can be reasonably assumed that "History" is a general topic, since its neighbors have fewer connections. It is plausible that, if history was a general topic, some of its neighbors would have a very large number of connections, because the neighbors of history would

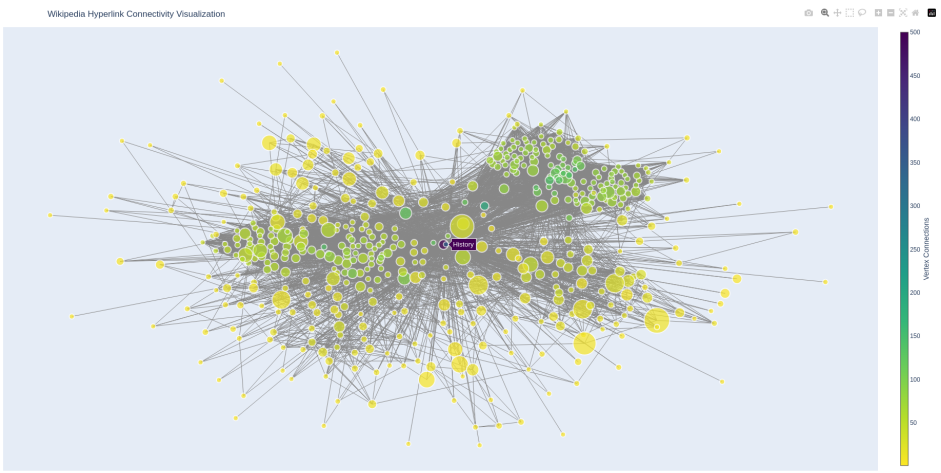

Figure 1: Output of *main("History", 2, True, set())*

be connected to each other. However, this is note the case. It is still unclear which articles behave similarly to the "getting-to-philosophy" phonomenon. Thus, while we have a measurement of generality from the graph, we still do not know what pecent of the entire Wikipedia eventually hyperlinks to a specific article.

3. Drawing Conclusions About how Data is Organized

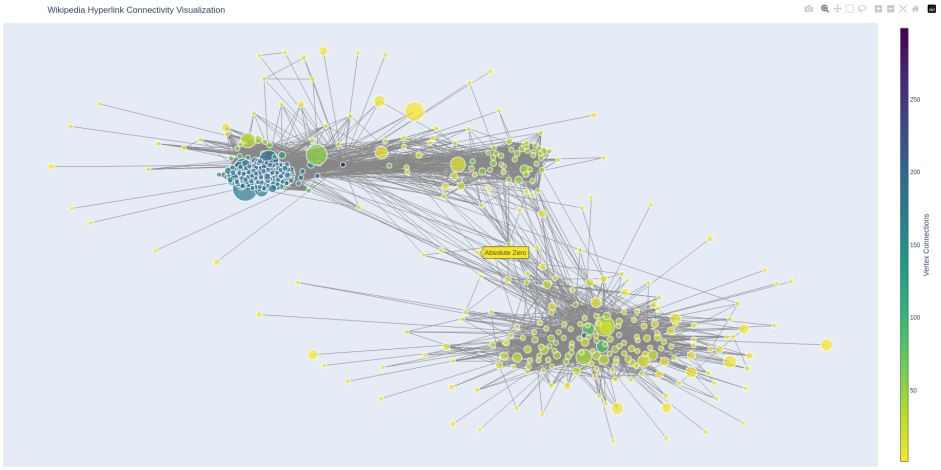

Figure 2: Graph starting at "Absolute Zero" with color representing connections.

By looking at various graphs, we can see how article size relates to category. An interesting conclusion we can draw is that very large articles tend to be a part of a large number of categories. This makes sense, because topics that are a part of many categories simultaneously tend to be very important (for example, the article on Albert Einstein is the categories for important poeple, physics, science, etc). We can also see that articles that have a large number of categories may not necessary have a large number of nodes (for example, the "Absolute Zero" graph). The graphs of a large random sample of articles will need to be examined before any definite statistical conclusion can be drawn. We did not determine the relationship between different categories.

Discussion

We were able to gather some interesting insights into the structure of Wikipedia articles, but we were unable to determine which articles most Wikipedia articles eventually link to when repeatedly traversing hyperlinks. There is

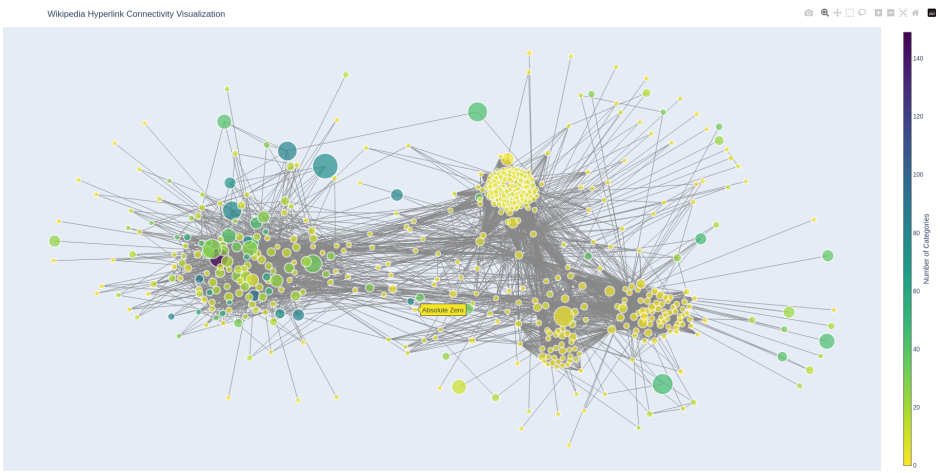

Figure 3: Graph starting at "Absolute Zero" with color representing categories.

likely a strong pattern between catagory size, node connections, and article size. To further our research we should call *main()* on a large random sample of articles in order to work out these patterns using regression. Another possible direction of research is to use more computational resources so that the graphs include more articles and possibly even the entire English Wikipedia itself. Additionally, it should be possible to determine the relationship between different categories by filtering them using the "categories" parameter of *main()*.

References

1. "API:Main Page - MediaWiki." *MediaWiki.* Retrieved March 8, 2023 from https://www.mediawiki.org/wiki/API:Main_page.
2. "API:Properties - MediaWiki." *MediaWiki.* https://www.mediawiki.org/wiki/API:Properties.
3. "API:Categoryinfo - MediaWiki." *MediaWiki.* Retrieved March 8, 2023 from https://www.mediawiki.org/wiki/API:Categoryinfo.
4. "API:Categorymembers - MediaWiki." *MediaWiki.* Retrieved March 8, 2023 from https://www.mediawiki.org/wiki/API:Categorymembers.
5. "API:Allpages - MediaWiki." *MediaWiki.* Retrieved March 8, 2023 from https://www.mediawiki.org/wiki/API:Allpages.
6. "Getting to Philosophy." *Wikipedia.* Retrieved March 8, 2023 from https://en.wikipedia.org/wiki/Wikipedia:Getting_to_Philosophy
7. "Top Websites." *Similarweb.* Retrieved March 8, 2023 from https://www.similarweb.com/top-websites/ https://en.wikipedia.org/wiki/Wikipedia:Getting_to_Philosophy
8. "Force-directed graph drawing." *Wikipedia.* Retrieved April 2, 2023 from https://https://en.wikipedia.org/wiki/Force-directed_graph_drawing