CSC111 Project Proposal: Mapping Wikipedia Article Structure

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

Tuesday, March 16, 2021

Problem Description and Research Question

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 betray 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.) These categories also have sub categories. This leads us to our project goal:

1. **Graphically represent connections between Wikipedia articles by hyperlinks, using categories, subcategories, and article size.**
2. **Locate "special 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.

Computational Plan

Articles on the English Wikipedia can be represented as a *directed graph*, where each node is a Wikipedia article, each edge leaving the node is a hyperlink in that article to another article, and each edge entering the node is a hyperlink to that article from another article. Theoretically, we can represent all 6,500,000+ articles as a part of a single graph. However, the graph will not be connected, because there are a very small number of Wikipedia articles that have no hyperlinks and no hyperlinks linking to them. Although, the vast majority (99.9%+) of Wikipedia articles have hyperlinks, so we can safely ignore the very small sections of Wikipedia that are not connected to the "main graph". Each node instance will contain at least 4 instance attributes:

1. name of the article (stored as a string. Every Wikipedia article has a unique name).
2. list of hyperlinks in the article (stored as a set of article names).
3. The category that the article belongs to (according to the categories Wikipedia itself uses).

4. The size of the article in kilobytes (stored as a float).

There are several ways we can implement the category instance attribute (a string for instance).

Using Less Data

Assume that it takes anywhere from 1 to 5 kilobytes to represent a single node. If we want to represent all of Wikipedia, it will take about 6GB to 30GB of storage. Additionally, doing computations on that much data will take a lot of time. Thus, graphs will be restricted to the size of a category or a maximum path size extending from a chosen node. In this case, we will omit hyperlinks to articles that are not in our dataset. It would otherwise be computationally infeasible in our setting to construct a graph of a large portion of Wikipedia. If we think of Wikipedia as a massive category, we can think of any subcategory of Wikipedia as a nearly accurate depiction of Wikipedia's connectivity, purely at a decreased scale that one can cognitively extrapolate if desired.

Measuring Hyperlink Properties

Consider a single Wikipedia article (node). If we assume that the hyperlinks of that node are completely random, we would expect the percentage of all articles in our dataset that eventually link to our article to be very close to the percentage of hyperlinks in the article which eventually link back to that very same article. However, the hyperlinks are probably not random: they are chosen for a variety of reasons. Thus, by comparing these two percentages, we can get a good look at how the hyperlinks are chosen. We can plot these percentages against each other in a scatter plot (using plotly), where every point in the plot is an article, and we can color the points to indicate which category they belong to. Similar measures of connectivity in the graph can give us additional insights.

Using packages to visualize data

Our first goal is to visualize the entire graph. This can be done with networkx or plotly. The main difference between the use of plotly here and its use in previous assignments is that a very large amount of data will be graphed using data points and graphing formats that we have not used before.

API usage

To achieve the desired graph, we will first collect relevant data to compile the graph. The data will be obtained using the MediaWiki Action API, which is the proprietary API for Wikipedia. The API features an accessible collection of properties for any given Wikipedia page. The properties pertaining to our graph are the name of the article, relevant hyperlinks, category, and size. This information can be retrieved under varying nomenclature from the "Properties" query for a page. To define the scope of a graph, we will additionally need more broad data, such as the collection of subcategories and pages belonging to a category. This can be performed using the *Categoryinfo* and *Categorymembers* queries. To find all pages belonging to a given category, we may use the *Allpages* query. Keep in mind, it is likely we will use additional queries as we continue to develop the project. To summarise, the MediaWiki Action API is highly versatile and is equipped to retrieve any Wikipedia data or properties we will need.

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/