Open Secrets: Investigating systemic biases in Wikipedia, the free online encyclopedia

Eamon Ma, Hisbaan Noorani, Rachel Xie, Philip Harker

March 16, 2021

1 Problem description and research question

As a free online encyclopedia run by volunteer collaborators, edits to Wikipedia articles form the backbone of the website. Without its open nature, Wikipedia would not be what it is today. However, this does not guarantee greatness — for the same reasons it is a symbol of communal knowledge and free information, Wikipedia is inherently biased. Most notably, it is biased towards the countries that articles are edited most in, and the dominant cultures in those countries. "[Our] research does show that most editors to Wikipedia come from the United States and Western Europe. And, as of 2020, our survey data indicate that fewer than 1% of Wikipedia's editor base in the U.S. identify as Black or African American. Considering these data, we can say with certainty that we are missing important perspectives from the world that Wikipedia strives to serve." (Uzzell, 2021). Along with an abundance of information of interest to Western society, there is a distinct lack of information from and for marginalized groups.

The coverage of knowledge is not spread evenly across all the Wikipedia articles. It varies based on what is desired and what is available, and often, the most available information (on English Wikipedia) concerns men from "developed" countries and their interests. Only 17.82% of Wikipedia's biographies are about women. This isn't merely an issue of Wikipedia's editors specifically being unreliable, it is an issue of there not being reliable sources in the wider Internet available for editors to feel confident creating articles about topics like women in science (Erhart, 2018). In this way, we can think of Wikipedia as a representation of our society's general knowledge base.

Our research question is, **"How can we use connections / links between Wikipedia articles to determine areas where our collective knowledge is lacking?**

2 Computational plan

The data we will be using will be an offline copy of a list of all of Wikipedia's articles. There are a few reasons for this.

1. Wikipedia request that research not be done on live versions of articles as it puts unnecessary strain on their servers — we would like to respect their wishes.

2. Changes to the dataset may occur midway through computation. This would likely lead to errors in graph generation if references to articles are edited while we are trying to compute on them.

3. Computational time. Working with an offline dataset is much faster than working with an online one. If we were to use the online version, we would have to wait for it to download each article every time that a re-calculation is run, instead of a single time at the beginning and then never again.

Additionally, there will be reduced computational complexity with the offline dataset because we will need to perform fewer operations. The fourth reason is bandwidth usage. While it is bandwidth taxing to download an 18 GB database, it is even more bandwidth taxing to re-download this database every single time a regeneration of the graph is required.

Our computational plan consists of three main stages. *Stage 1, Stage 2,* and *Stage 3.*

2.1 Stage 1: Processing the data

We plan to create a graph of links between Wikipedia articles. These links are the hyperlinks visible on an article (typically in blue text) that link to other articles. External links will be ignored. We also wish to store a little bit more information about each article such as the word count, number of viewers, number of editors, number of edits, etc. This means that a simple `Graph` and `_Node` structure will not suffice. A new `Graph`, `_Node` pair with instance attributes to represent the extra data that is to be collected. The exact data we'll collect is subject to change as the viability of collecting each metric and the use that it will provide are considered.

A challenge arises when we look at where to save such graph. We will create a format that saves this graph as first; a set of nodes in the graph, second; a list of the edges in the graph, and third; a format that will save all other information for each node in the graph. This file may be something like a json or csv file.

2.2 Stage 2: Analysis

After the graph is created, nodes with characteristics determined to be found in underrepresented articles will be singled out pragmatically. After this, each article will be analyzed on both an individual and a macroscopic level. We will look for trends in the articles that we find to be underrepresented for reasons other than age of article, obscurity of topic (low demand), and other metrics. The trends discovered in this analysis will allow us to find more such articles that may not display the same severity of symptoms, and also allow us to look at the state of the internet as a whole.

2.3 Stage 3: Visualization

After this preliminary analysis is completed, a visualization system will be developed that will allow the user to find underrepresented articles, view them in relation to other articles around them, and visit the actual articles. This will allow the user to discover why these articles are underrepresented.

References

Erhart, E. (2018). Why didn't Wikipedia have an article on Donna Strickland, winner of a Nobel Prize? *Medium*. https://medium.com/freely-sharing-the-sum-of-all-knowledge/why-didnt-wikipedia-have-an-article-on-donna-strickland-winner-of-a-nobel-prize-dff8e518daaa

Uzzell, J. (2021). Who tells your story on Wikipedia. *Medium*. https://medium.com/freely-sharing-the-sum-of-all-knowledge/who-tells-your-story-on-wikipedia-6d6c29f45028