CS240 - Network Routing

Team Hermes - 002

Nishanth Prasanna Kumar - npras017, Noah Tsai - ctsai034, Prathik Somanath - psoma005, Shubhank Joshi - sjosh052, Vivek Venkateshprasad - vv002

Project Proposal

Is the popularity of Wikipedia articles contagious ?

Abstract

Wikipedia stands as the sixth most visited website globally, attracting a diverse audience. The articles within Wikipedia form an intricate network through numerous hyperlinks interconnecting them. This project investigates the correlation between the popularity of Wikipedia articles and the popularity of their interconnected neighboring articles, as measured by the Google PageRank algorithm.

In this study, "popularity" is defined in terms of the pageviews received by an article. The primary question addressed is whether a relationship exists between the popularity of a Wikipedia article and the popularity of its adjacent hyperlink articles, utilizing a suitable algorithm. Notably, PageRank reflects the importance of a page based on the number and popularity of pages that link to it and the overall structure of the interconnected network.

The methodology involves adopting a graphical approach, where selected subsets of Wikipedia (outlined in “Data and Tools”) are treated as directed graphs. In this representation, articles are depicted as nodes, and hyperlinks serve as edges. The focus is on understanding the dynamics of popularity within the Wikipedia ecosystem through the lens of graph theory and PageRank, shedding light on the interconnected nature of information dissemination within this widely accessed knowledge platform.

Problem Statement

Problem: This research tries to address the issue of understanding the interrelationship between articles within Wikipedia, investigating facets such as association, correlation, and causality. The central problem at hand is the determination of how the popularity of one Wikipedia article may impact the popularity of another. The investigation necessitates an examination of the influence of hyperlinks, employing the PageRank algorithm as a crucial metric. The challenge lies in the validation of this influence by comparing PageRank values with the actual pageviews of specific articles, aiming to provide insights into the intricate dynamics of popularity within the Wikipedia ecosystem.

Input: Selected subsets of Wikipedia articles, where the selected subsets both represent pages that are prone to and unaffected by fluctuation in views. For example, celebrity pages are more prone to sudden influx of views due to current events, while history and math theorem pages are generally unaffected by time.

Desired Output: An understanding of the correlation between the popularity of a Wikipedia article and the popularity of its neighbors, giving insight regarding the several controlled factors involved in research to allow for comprehensive analysis of the chemical reactions between Wikipedia articles.

Previous and Relevant Work

Although there are previous research papers dedicated to exploring Wikipedia through similar lenses, there is nevertheless a lack of research with the same objective, making it challenging to design a streamlined research process on our own. One such paper [2] discusses the popularity as the number of edits made to a page and how it is affected by the granularity of the particular article. We have adopted this to find out if this categorization has the same outcome when we define popularity to be the number of pageviews.
Another paper [5] discusses the importance of the first link network and defines the first link as the article's closest neighbor. However, we set out to find the correlation of the articles with all its neighbors (All the hyperlinks contained in the article). This will give us a broader view in figuring out the flow of popularity influence in a Wikipedia network.

Data and tools

Wikipedia Categories: South American history stubs, Ancient Greece stubs, Indian Artists. We selected the first two categories because they're likely to remain consistent over time, since they're about history. We wanted to see if the correlation still holds true in the Indian Artists category, even though it's dynamic and may have more edits and added articles.

Wikipedia Python API**:** We utilize the Wikipedia Python API to collect data from the categories mentioned above. This enables us to browse through subcategories and obtain all the pages associated with them. We retrieve several details including titles, wiki-links, article IDs, and view counts within the category, and store them in JSON format.

Python NetworkX: We utilize the Python NetworkX library to build and examine the directed graph network. This allows us to analyze the properties and relationships among popular articles and their influence on neighboring articles.

Python data manipulation tools: We will utilize **NumPy** and **Pandas** for implementing PageRank due to their efficient handling of matrix operations and data manipulation, facilitating the computation and organization of data necessary for the algorithm's execution.

Google PageRanking: We utilize the Google PageRank algorithm to rank Wikipedia articles using inbound link quality(measured by the popularity of the pages linking a particular page). This method accounts for popularity and importance within our selected category, providing a comprehensive ranking system.

**We cannot use huggingface dataset because it does not contain information that we require, like hyperlinks and category.*

Contributions / Roadmap

Week 1: Collect data by scraping the selected subset of Wikipedia articles and use existing Wikipedia APIs.
Week 2: Create a network graph with the available data using Python NetworkX.
Week 3: Apply the PageRank algorithm and collect pageviews of all the pages.
Week 4: Compare, analyze and make inferences of the results obtained.
Week 5: Collate all information obtained and make it presentable.

Bibliography

[1] Aspert, Nicolas, et al. "A graph-structured dataset for Wikipedia research." Companion Proceedings of The 2019 World Wide Web Conference. 2019.
[2] Lerner J, Lomi A (2018) Knowledge categorization affects popularity and quality of Wikipedia articles. PLoS ONE 13(1): e0190674. https://doi.org/10.1371/journal.pone.0190674
[3] Spoerri, Anselm (2007) What is Popular on Wikipedia and Why? First Monday, volume 12, number 4 (April 2007), http://firstmonday.org/issues/issue12_4/spoerri2/index.html
[4] Jun Liu, Sudha Ram, Using big data and network analysis to understand Wikipedia article quality, Data & Knowledge Engineering,Volume 115,2018, Pages 80-93,ISSN 0169-023X, https://doi.org/10.1016/j.datak.2018.02.004.
[5] Mark Ibrahim, Christopher M. Danforth, Peter Sheridan Dodds, Connecting every bit of knowledge: The structure of Wikipedia's First Link Network, Journal of Computational Science, Volume 19, 2017, Pages 21-30, ISSN 1877-7503, https://doi.org/10.1016/j.jocs.2016.12.001.
[6] Consonni, Cristian, David Laniado, and Alberto Montresor. "WikiLinkGraphs: A complete, longitudinal and multi-language dataset of the Wikipedia link networks." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 13. 2019.