Republic of the Philippines
North Eastern Mindanao State University
Formerly Surigao del Sur State University
LIANGA CAMPUS
Lianga, Surigao del Sur 8307
www.sdssu.edu.ph

PageRank Algorithm as a Search Engine Instrument for Disinformation Detection

Dexter L. Alolod

Arabella A. Morgado

College of Information Technology Education

March 2023

Table of Contents

Introduction

Fake news is not a new issue, but it poses a greater challenge now. According to Benjamin (2018), the flow of information has risen dramatically, with information now spreading globally within seconds through the Internet. As disinformation has spread quickly, it has a big effects on how people think and act as well as on society as a whole. Most people use search engines as their main source of information. Because of this, it is important to come up with a way to find and promote reliable information sources while reducing the amount of disinformation.

Google's PageRank algorithm is a well-known search engine ranking tool. The algorithm analyzes the significance of web sites based on the quantity and quality of links going to them, with the goal being to promote trustworthy information sources while reducing disinformation. Yet, the efficiency of the PageRank algorithm in detecting and reducing disinformation remains undetermined. As search engines have become the primary tool for accessing information, this study will look at PageRank's ability to spot false information.

Objectives

The main objectives of this research paper are as follows:

- To understand the concept of disinformation and its impact on society.
- To explore the role of PageRank algorithm in identifying and ranking reliable and non-credible sources.
- To analyze the effectiveness of PageRank algorithm in combating disinformation in search engines.

Problem Statement

The spread of disinformation poses a huge threat to democratic countries because it may be used to impact public opinion and political debate. Existing methods for detecting fake news and disinformation are frequently ineffective, despite the fact that there are a number of them. The PageRank algorithm has the potential to give a more reliable way for identifying disinformation, as it takes the network structure and relationships between information sources into account.

Literature Review

The rapid spread of false information throughout the internet is becoming a major cause for concern for both academics and government policymakers. As the number of information that is available online has been growing at an exponential rate, it has become increasingly challenging to differentiate between genuine information and disinformation. As a result of this, there is an increasing demand for resources that can assist in recognizing false information and preventing this further propagation. The PageRank algorithm, which was created by Google, is one example of such a tool. This method was designed to assess the relative significance of individual web sites based on the quality and number of connections to those pages. This study examines the application of the PageRank algorithm as a tool for detecting disinformation that may be implemented into search engines.

The PageRank algorithm has been demonstrated to be an effective tool for identifying and combating the spread of disinformation on the Internet. Wang et al. (2021) investigated the interlinking structure of websites that promote anti-vaccination misinformation and discovered that these sites frequently link to one another, producing a web of disinformation. By observing how websites link to one another, they determined that the PageRank algorithm could be used to identify and prevent the spread of false information. This strategy can be especially helpful for identifying and targeting th e most important sources of disinformation, as these sites typically have the highest PageRank rankings. It may be capable of significantly decrease the spread of false information online by focusing on these high-ranking websites.

Similarly, Lazer (2019) examined the connections of sites that disseminated fake news during the 2016 US presidential election. He discovered that these sites were frequently connected, creating a web of propaganda. Lazer stated that the PageRank algorithm may be used to identify fake news and prevent its spread by analyzing the interlinking of websites. Hence, using the PageRank algorithm to detect and target these interconnected fake news sites could be an effective strategy for preventing the online spread of disinformation.

Overall, the PageRank algorithm offers an excellent method for detecting and combating the spread of disinformation on the internet. By studying the linking patterns of websites, the algorithm can help with determining the sources of disinformation and preventing its spread. However further study is required to evaluate the algorithm's usefulness in detecting disinformation, the reviewed studies indicate that it is a viable tool for combating this increasing problem. Future search engines and other online platforms will likely depend more heavily on the PageRank algorithm to counteract the spread of disinformation and promote the accuracy of web content.

Methodology

This methodology describes how to examine data sets in order to locate accurate data. To strengthen the reliability of measures, we want to identify non-credible information to prevent the spread of fake news among users. We also provide efficient methods for computing PageRank, which uses a mathematical formula to assign a score to each web page. Therefore, we provide a detailed description of the steps followed by a flowchart of an entire algorithm.

Step 1: Construct a Web Graph for Disinformation Detection

To detect disinformation, we construct a web graph, where nodes represent web pages and edges represent links between the pages. The web graph should include pages that are likely to contain disinformation. This graph would clearly depict what nodes are responsible for their disinformation in the network. The general acceptability of the result was measured using the PageRank algorithm in this web graph.

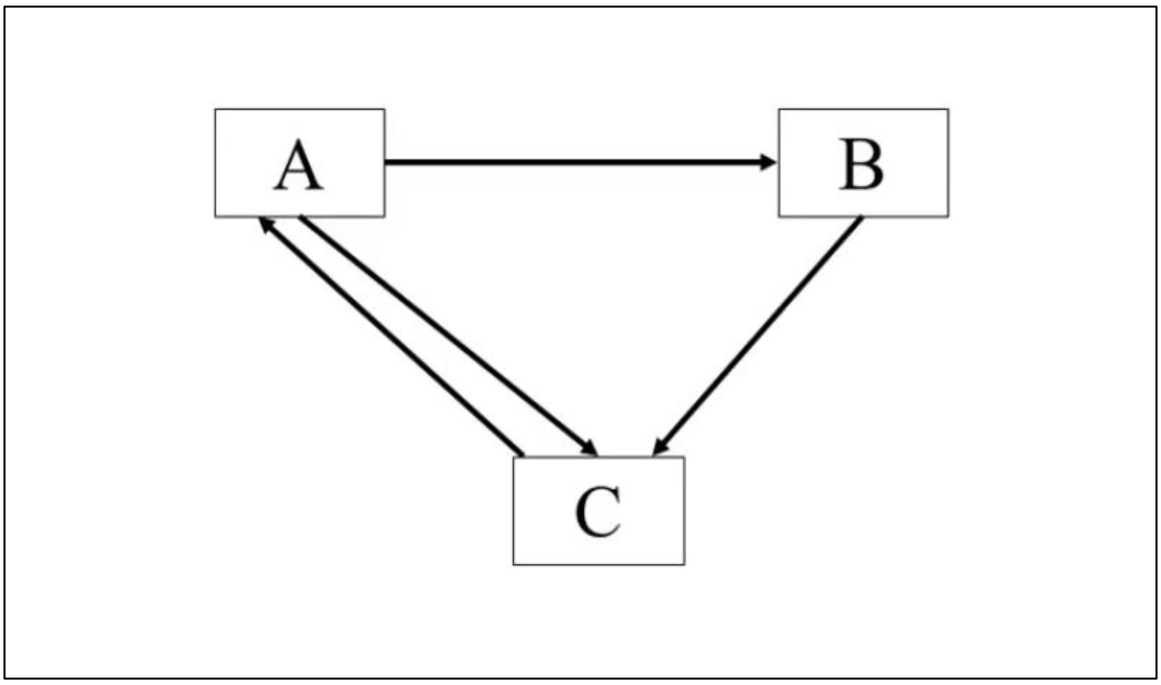

Figure 1: A web page graph: A sample graph to show the construction of a web page graph

Step 2: Calculate the Initial Web Page Scores

The next step is to figure out each web page's initial PageRank score. One common way to do this is to give each web page an initial score of 1/N, where N is the number of web pages. A random number can also be used to decide what the initial score is.

Step 3: Implement the PageRank Algorithm

The PageRank algorithm is an iterative process used to compute each web page's PageRank score. The algorithm begins by assigning an initial PageRank score to each website. Next, for each cycle, the algorithm recalculates each web page's PageRank score based on the scores of its inbound and outbound links.

Step 4: Calculate the new Web Page scores

To calculate the new PageRank score, the algorithm uses the PageRank formula that includes the damping factor, the set of web pages that link to a particular page, the number of outbound and inbound links from each of the web pages in the set, and the PageRank of each web page in the set.

PR(A) = (1-d) + d (PR(Ti)/C(Ti) + ... + PR(Tn)/C(Tn))

- PR(A) represents the PageRank of web page A.
- d is the damping factor, which is typically set to 0.85.
- Ti...Tn represents the set of web pages that link to web page A.
- C(Ti)...C(Tn) represents the number of outbound links from each of the web pages in the set Ti...Tn.
- PR(Ti)...PR(Tn) represents the PageRank of each web page in the set of Ti...Tn.

Step 5: Update the web page scores

After generating the new PageRank scores for each website, the algorithm updates the scores by replacing the new scores for the old ones. The procedure is repeated until the PageRank scores converge, which occurs when the PageRank scores no longer change considerably between iterations.

Step 6: Analyze the Results for Disinformation Detection

Once the PageRank scores have converged, the results are analyzed to identify the relative significance of each web page on the network. Pages with higher PageRank scores are considered to provide more credible information than those with lower scores. Pages with unusually low PageRank scores may indicate the presence of disinformation, and further investigation may be required.

FLOWCHART OF THE ALGORITHM

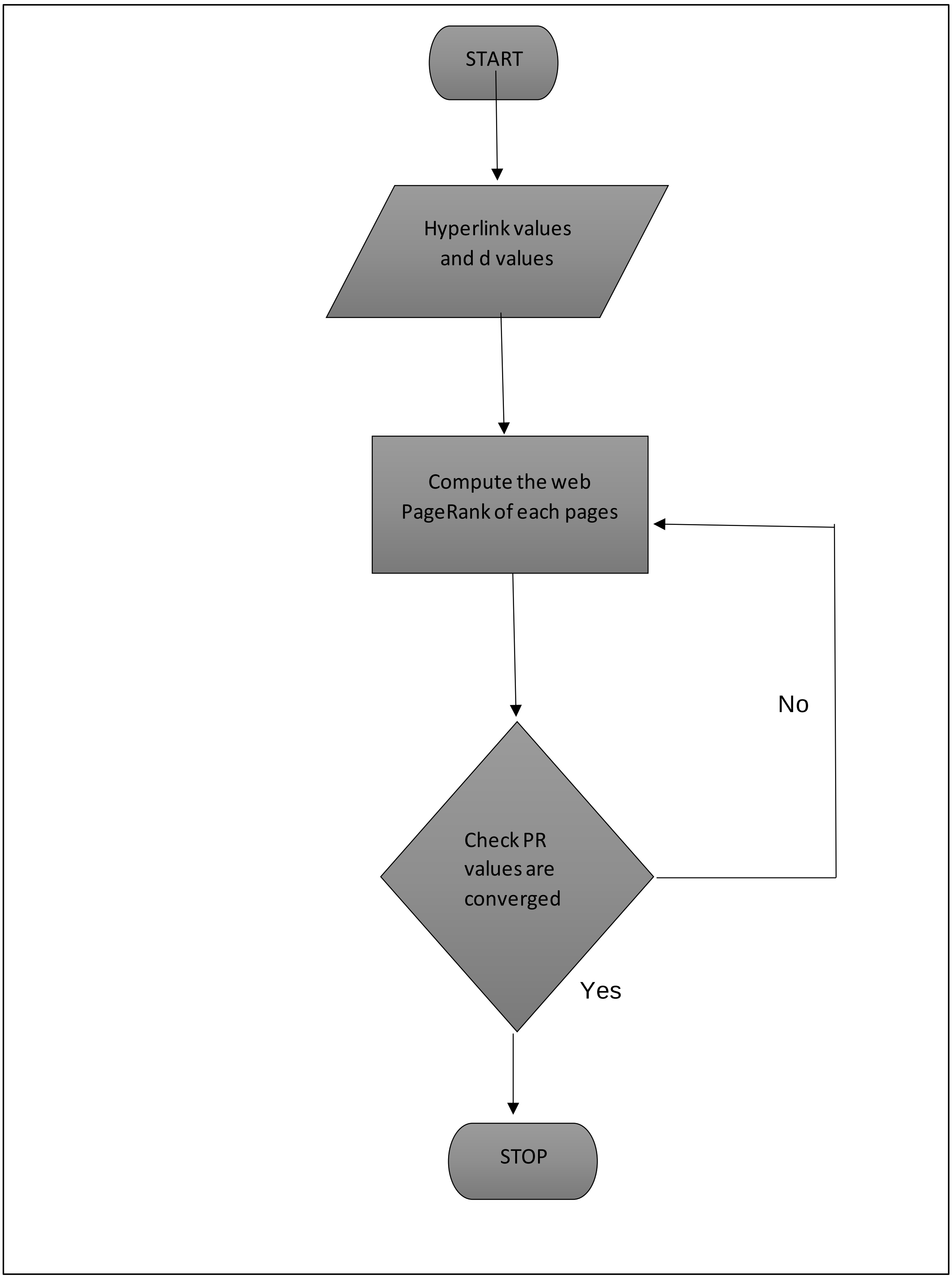
START
Hyperlink values and d values
Compute the web PageRank of each pages
No
Check PR values are converged
Yes
STOP

Figure 2: Algorithm for determination of disinformation in a page, the flowchart gives the algorithm to detect disinformation using a web graph to estimate the general acceptability of tweets and the credibility of sources.

Data Collection

To do this research, we put together a list of websites that are known to spread false information. The data set was put together from reliable sources, like news articles and reports from well-known NGOs that monitor disinformation campaigns.

Various methods, such as web scraping, were used to get the website's content, domain authority, page rank, backlinks, and other important system of measurement. Then, this information was used to critic how well the PageRank algorithm could find and rank trustworthy sources.

Data Preprocessing

Before conducting any analysis, the collected data had to be preprocessed to make sure it was good and would work with the analytical tools. The preprocessing steps included:

Data Cleaning

The researchers filtered the data of any redundant or irrelevant information. Incomplete data that might affect the analysis were also eliminated.

Data Reduction

We filtered the dataset to only include factors relevant to the study. This was done to reduce the computational complexity of the analysis and to ensure that the attention was placed on the most important factor.

Data sets

We carried out experimental evaluation on data sets obtained from a popular search engine 'Google'. We collected from various sources, including news websites and online forums. The web graph will be constructed by using a Google web crawler or Googlebot that will identify the links between different web pages and create a directed graph where the nodes represent web pages, and the edges represent links between the pages.

The dataset will include the following variables:

- Node ID: Unique identifier for each web page.
- PageRank score: Score assigned to each web page based on the PageRank algorithm.
- Inbound links: is a link that is received by a web page from another web page.
- Outbound links: is a link that is created by a web page to point to another web page.
- Source category: Category of the web page (reliable or disinformation).
- Website URL: The web address of the website

Data Training

We trained the PageRank algorithm using a dataset consisting of web pages that had previously been labeled as reliable or containing disinformation. The algorithm was then applied to the respective web graphs of these pages in order to compute their PageRank scores. These ratings were then compared to the pre-existing categories in order to evaluate the algorithm's efficiency in recognizing reliable source and detecting disinformation. The algorithm was optimized through a process of iteration and retraining to attain a high degree of accuracy in identifying credible sources and detecting disinformation.

Data Testing

To evaluate the efficiency of the PageRank algorithm in identifying disinformation, we used a group of vaccine-related web pages. Each web page was personally tagged as either reliable or containing disinformation. Then, we applied the PageRank algorithm to the web graph and compared the PageRank scores of credible and false web pages. The results were analyzed to determine if the algorithm accurately identified reliable sources and detect disinformation.

Conclusion

In conclusion, spreading false information has become a major problem that causes misunderstanding, anxiety, and even harm to people and society. Since search engines have become the main way to find information, it is important that only trustworthy sites show up at the top of search results.

Google's PageRank algorithm is important for finding and ranking credible sources. This research study investigates the function of the PageRank algorithm in detecting and prioritizing reliable sources to counteract disinformation in search engines. According to the literature study, by examining the connecting patterns of websites to discover their origins, the algorithm may prove to be a powerful asset in the battle against disinformation. The methodology section explains in detail how to utilize the PageRank algorithm to identify reliable information and prevent the spread of fake news among users. In conclusion, the PageRank algorithm provides a realistic technique for combating misinformation and safeguarding the credibility of online content.

References:

Benjamin, Terri-Anne, Shashi, Muhammad, & Juhi. (2018). *Rajaratnam School of International Studies*.

Kou, Y., & Wu, Y. (2019). A Modified PageRank Algorithm for Fake News Detection on Twitter. *Journal of Intelligent Information Systems, 53(3), 487-499.* https://doi.org/10.1007/s10844-018-0521-6

NewsVerifier: *A New Approach to Fake News Detection by Leveraging Linguistic Features and Hierarchical Clustering. arX*

Kumar, K. S., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-Centric Computing and Information Sciences*, *4*(1). https://doi.org/10.1186/s13673-014-0014-x

Wang, X., Zhang, M., Fan, W., & Zhao, K. (2021). Understanding the spread of COVID-19 misinformation on social media: The effects of topics and a political leader's nudge. *Journal of the Association for Information Science and Technology*, *73*(5), 726–737. https://doi.org/10.1002/asi.24576

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706