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 Morgado

College of Information Technology Education

March 2023

Table of Contents

Introduction

Disinformation has become a major issue in modern times, causing confusion, uncertainty, and even harm to individuals and society as a whole. As search engines have become the primary tool for accessing information, it is critical to have reliable sources at the top of the search engine results. PageRank algorithm, developed by Google, has played an important role in identifying and ranking reliable sources. This research paper aims to explore the role of PageRank algorithm in identifying and ranking reliable sources to combat disinformation in search engines.

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 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

Research has also been done that looks at how false information, often known as disinformation, has been rapidly spreading via numerous web site pages, making it crucial to identify and stop its spread. The PageRank algorithm, created by Google to determine the relative significance of individual web sites, might be useful here. The purpose of this literature study is to examine the possibility of using Google's PageRank algorithm to detect and counter the spreading of disinformation. The PageRank algorithm is a method that calculates a web page's relative relevance depending on the quantity and quality of links referring to it. Search engines place greater value on pages that include links from other, higher-quality sources. It is generally agreed that the PageRank algorithm is an

efficient means of ranking online sites by their relative importance, and as such, it has been adopted by many search engines.

Disinformation refers to deliberately false information that is disseminated online. It may have far-reaching effects, such as swaying public opinion or even encouraging violence, and comes in a variety of forms, including propaganda and false news. With the use of the PageRank algorithm, Wang, X., Zhang, M., Fan, W., & Zhao, K. (2021) analyzed the interlinking structure of websites that promotes anti-vaccine disinformation. In their research, they discovered that these sites often linked to one another, forming a web of disinformation. Nonetheless, reputable sites that presented factual data concerning vaccinations often linked to one another and to official government and medical resources. In a recent study, Lazer, D. (2019) looked at how sites that distributed fake news during the 2016 US presidential election were linked to one another. These sites were discovered to often connect to one another, creating a web of propaganda. As opposed to this, reputable news websites often referenced other reputable news websites as well as official government sources.

The PageRank algorithm may prove to be a powerful instrument in the fight against disinformation. The algorithm may help in the fight against the spread of misleading information by examining the linking patterns of websites to determine their origins. The PageRank algorithm has the potential to be used in the future to the detection of disinformation on news web sites, which have become a significant source of fake news. PageRank, in the end, it offers a viable strategy for fighting disinformation and protecting the reliability of data found online.

Methods

We used a web graph consisting of web pages containing information about a particular subject, such as vaccinations or politics. The algorithm is then applied to the web graph in order to rank web sites according to their PageRank score. The scores are evaluated to see if they properly identify credible sources and detect disinformation.

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.

The dataset will be cleaned and pre-processed to eliminate duplicates and irrelevant data. Each web page's PageRank score will be determined by applying the PageRank algorithm to the web graph. The results will be examined to evaluate the efficiency of the PageRank algorithm in detecting and prioritizing credible sources and combatting disinformation in search engines.

Data Training

We trained the PageRank algorithm using a dataset consisting of web pages that had previously been labeled as reliable or containing misinformation. 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 detected disinformation.

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. Consequently, 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 who nodes are responsible for their disinformation in the network. The general acceptability of the result was measured using PageRank algorithm in this web graph.

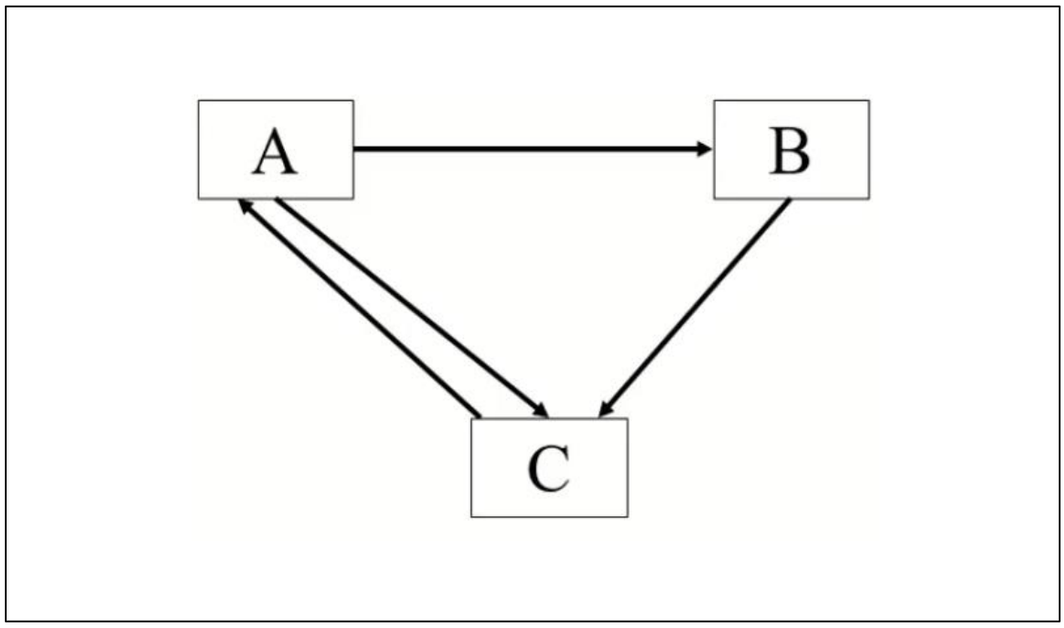

Figure 1: A web page graph: A sample graph to show the construction of a web page graph

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 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 high PageRank scores may indicate the presence of disinformation, and further investigation may be required.

FLOWCHART:

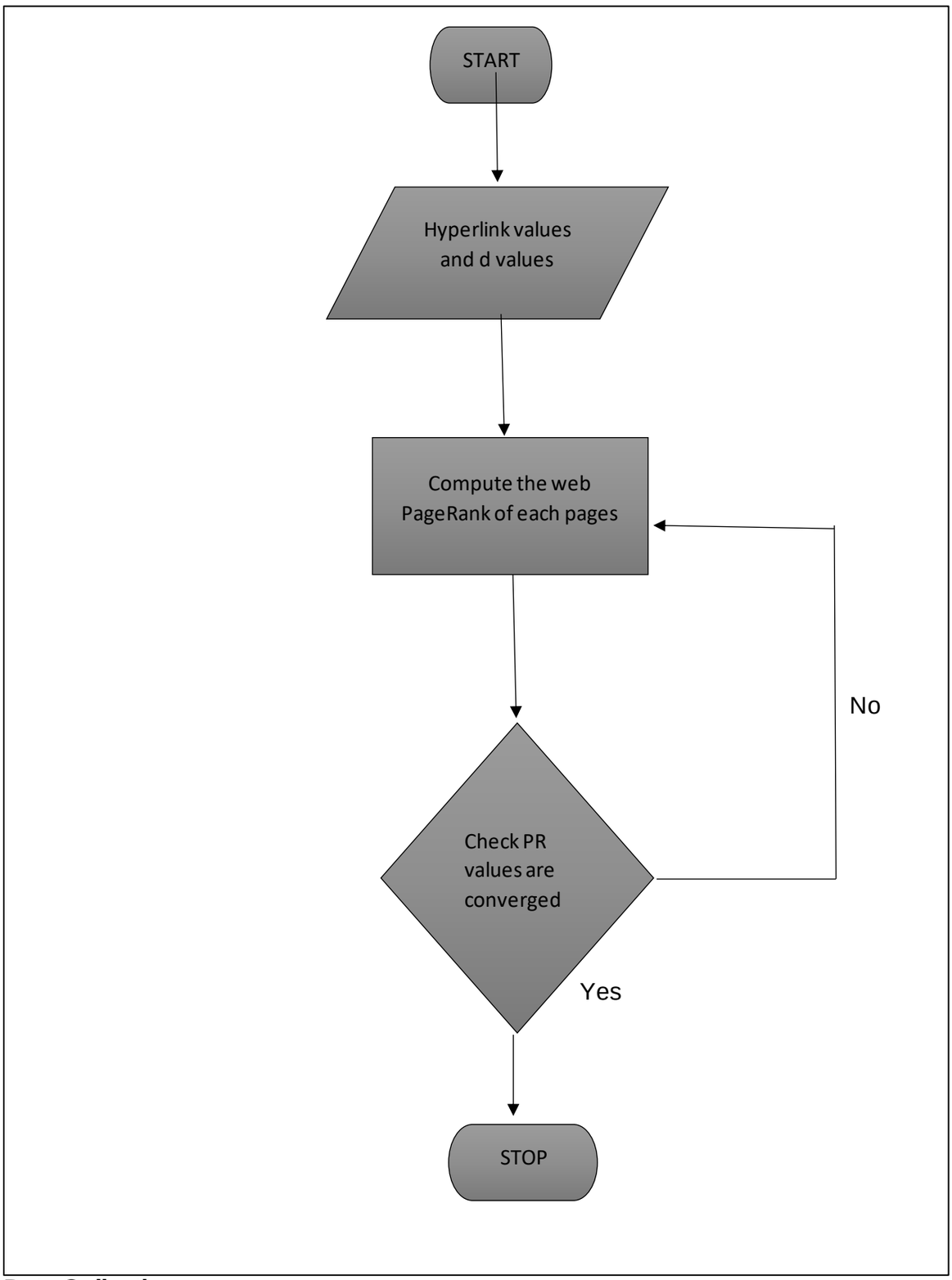

Figure 2: Algorithm for determination of disinformation in page. The flowchart giving the algorithm to detect disinformation using a web graph to estimate the general acceptability of tweets and credibility of sources.

Data Collection

In order to conduct this research, we collected a list of websites known to disseminate false information. The dataset was compiled from credible sources, such as news articles and reports from recognized NGOs that monitor disinformation campaigns.

The websites content, domain authority, page rank, backlinks, and other relevant metrics were gathered using various techniques, such as web scraping. This information was

then used to evaluate the performance of the PageRank algorithm in recognizing and ranking credible sources.

Data Preprocessing

Before conducting any analysis, the collected data had to be preprocessed to ensure its quality and compatibility 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 computing complexity of the analysis and to ensure that the attention was placed on the most important factor.

Conclusion

In conclusion, the dissemination of disinformation has become a significant problem, causing confusion, anxiety, and even damage to people and society. As search engines have become the major method for gaining access to information, it is essential that only credible sites appear at the top of search engine results.

Google's PageRank algorithm is important for finding and ranking credible sources. This research study investigates the function of PageRank algorithm in detecting and prioritizing reliable source to counteract disinformation in search engines. Examining the connecting patterns of websites to discover their origins, the algorithm may prove to be a powerful asset in the battle against disinformation, according to the literature study. 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:

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