WEB MINING: PAGE RANK
Oprea Olivia Maria-Magdalena IA1B

1. Importance and practical applications of the algorithm

PageRank is an algorithm used by Google Search to rank websites in their search engine results. It appeared as a result of Larry Page and Sergey Brin's work while they were Ph.D. students at Stanford University.

The importance of PageRank in web mining is connected to the fact that it can effectively rank and organise large amounts of information on the internet. PageRank makes it possible for search engines like Google to quickly and accurately return the most relevant results for a user's query.

The way to measure the credibility of a website is to use the number and quality of links pointing to it. Thus, PageRank helps to ensure that the websites that appear at the top of search results are the most reputable and trustworthy sources of information. This reduces the chance of directing users to unreliable or poor quality websites, while also making it easier for them to find the information they need.

In addition to being used in search engines, PageRank has also been used in other fields. Some noteworthy mentions are bioinformatics, social networks, and recommender systems. The algorithm is also used for evaluating the importance of a webpage, and it is used to evaluate the quality of backlinks.

PageRank also has a direct impact on SEO (Search Engine Optimization) strategies and website's visibility. The better the PageRank score is, the higher a website will appear in the search results, thus increasing the website's visibility and the number of visitors to the website.

Some practical applications of the PageRank algorithm in web mining include:

Search engines: As mentioned before, PageRank is used by Google and other search engines to rank websites in their search results. This helps users find the most relevant and credible information for their queries.

Link analysis: PageRank can be used to evaluate the importance or quality of links between websites. This can be useful for identifying key influencers or authoritative sources in a particular field.

Recommender systems: PageRank can be used to recommend items or resources to users based on their browsing history or other behaviours. This can be applied in e-commerce, social media, and other platforms.

Social network analysis: PageRank can be used to identify key nodes or influential users in a social network. This can be useful for understanding how information spreads and identifying opinion leaders.

Information retrieval: PageRank can be used to retrieve and organise relevant information from large datasets such as news articles, scientific papers, and other documents.

SEO: As mentioned before, PageRank can help website owners to improve their website visibility and ranking in search engine results, by providing useful information on how to optimise their website's backlinks and other SEO strategies.

For short, PageRank is an algorithm that assigns a "rank" to each webpage on the internet, based on the number and quality of links pointing to it, it is widely used in web search and information retrieval, it is used to evaluate the importance of a webpage, and it is used to evaluate the quality of backlinks.

2. The algorithm general presentation

PageRank is an algorithm used to rank websites and evaluate their importance in a network of web pages.

The algorithm assigns a "rank" to each website based on the number and quality of links pointing to it. The basic idea is that a link from one website to another is a "vote" of credibility for the second website. The more votes a website has, the more credible it is deemed to be.

The PageRank algorithm uses a mathematical formula to evaluate the importance of each page in the network. The formula takes into account the number of inbound links to a page, as well as the importance of the pages linking to it. The algorithm then assigns a score or rank to each page based on this information.

The page rank algorithm uses a combination of a "damping factor" and the "random surfer model" to calculate the importance of a web page. The damping factor is a value between 0 and 1 that is used to simulate the probability that a user will continue clicking on links from one page to another.

The PageRank algorithm is based on the idea of a random surfer: the algorithm assumes that a user randomly clicks on links on the internet, and that the probability of a user visiting a page is directly proportional to the number and quality of links pointing to it.

PageRank is an iterative algorithm and it starts with a random set of initial ranks, the algorithm repeatedly updates the ranks of the pages until it reaches a stable state where the ranks no longer change significantly.

3. Known results and issues

PageRank has been widely used for searching different topics on the web and for retrieving information. It has been successful in ranking web pages based on their importance and relevance. Some of the known results of the PageRank algorithm include:

Improved search results: By using the number and quality of links pointing to a website as a measure of its credibility, PageRank helps to ensure that the websites that appear at the top of search results are the most reputable and trustworthy sources of information.

Increased website traffic: Websites that appear at the top of search results are more likely to receive more traffic. As a result, PageRank can be used to increase the visibility and popularity of a website.

Better organisation of information: PageRank makes it possible to organise large amounts of information on the internet in a logical and meaningful way. This makes it easier for users to find the information they need.

Identifying key influencers: PageRank can be used to evaluate the importance or quality of links between websites, which can be useful for identifying key influencers or authoritative sources in a particular field.

Recommender systems: PageRank can be used to recommend items or resources to users based on their browsing history or other behaviours.

SEO: PageRank can help website owners to improve their website visibility and ranking in search engine results, by providing useful information on how to optimise their website's backlinks and other SEO strategies.

However, there are also some known issues and limitations associated with the algorithm:

Link manipulation: Because the algorithm assigns importance based on the number and quality of links pointing to a website, some people have attempted to manipulate search results by artificially creating large numbers of links to their own websites. This is called "link farming" or "link buying" and it is considered as a black hat SEO technique.

Spam: Since the PageRank algorithm assigns importance based on the number of links pointing to a website, spammers have attempted to artificially inflate the importance of their sites by creating large numbers of low-quality links.

Lack of freshness: PageRank does not take into account the freshness of the content. Because of this, websites with outdated information can still obtain a high rank, which may not provide the best result to the user.

Lack of context: PageRank does not take into account the context of the query or the specific topic that the user is searching for. This can lead to irrelevant results appearing at the top of the search results.

Personalization: Since PageRank does not take into account the individual preferences of each user, a lack of personalization in the search results can easily be seen.

4. Datasets used (names and samples)

I used my own datasets as I made my own simple version of the page rank algorithm.

It uses square matrices NxN.

For example:

Example 1:
0,1,1,0,0,1,0
1,0,0,1,1,0,0
1,0,0,1,0,1,0
0,1,1,0,0,0,1
0,1,0,0,0,1,1
1,0,1,1,1,0,0
0,0,0,1,1,0,0
Example 2:
0,1,1,0,0
1,0,0,1,1
1,0,0,1,0
0,1,1,0,0
0,1,0,0,0
Example 3:
0,1,1,0,0,1,0,1,1
1,0,0,1,1,0,0,0,1
1,0,0,1,0,1,0,0,0
0,1,1,0,0,0,1,1,1
0,1,0,0,0,1,1,0,1
1,0,1,1,1,0,0,0,0
0,0,0,1,1,0,0,1,1
1,0,0,1,0,0,1,0,0
1,1,0,1,1,0,1,0,0

5. Results (including samples) and evaluation

The PageRank algorithm that I provided calculates the importance of each page in a network of web pages based on the number of inbound links and the importance of the pages linking to it.

The algorithm starts with initialising the PageRank scores for all pages in the network to a starting value, such as 1, and then it uses a mathematical formula to iteratively update the scores of each page until they converge.

The final output of the algorithm is a score vector "v" representing the pageRank scores for all pages in the network, where v[i] is the pageRank score of page i.

This score can be used as a measure of the importance of each page in the network, with higher scores indicating more important pages.

It's important to keep in mind that the output of the algorithm is an approximation and the value of the PageRank score does not have a clear meaning. Moreover, the output is affected by the starting value, damping factor and the number of iterations, which are all adjustable parameters and have to be chosen based on the specific use case.

The code that I provided earlier is a simplified example of how the PageRank algorithm can be implemented in Python using the NumPy library and it is not meant to be used as a production-ready implementation.

In terms of evaluating this specific code, there are a few things to consider:

The code assumes that the input matrix G is a dense matrix, which may not be the case for large datasets. If the input matrix is sparse, it would be more memory-efficient to use a sparse matrix representation.

The code uses a while loop to iterate until the scores converge, but this method may not be efficient for large datasets. A more efficient approach would be to use an iterative method that guarantees convergence, such as the power iteration method.

The code does not provide any error handling or input validation, which would be necessary in a production-ready implementation.

The code is not optimised for distributed computing and it would not be able to handle large datasets with very large numbers of web pages.

6. References

https://en.wikipedia.org/wiki/PageRank

https://www.geeksforgeeks.org/page-rank-algorithm-implementation/

https://www.semrush.com/blog/pagerank/

https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af

https://neo4j.com/docs/graph-data-science/current/algorithms/page-rank/

http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html