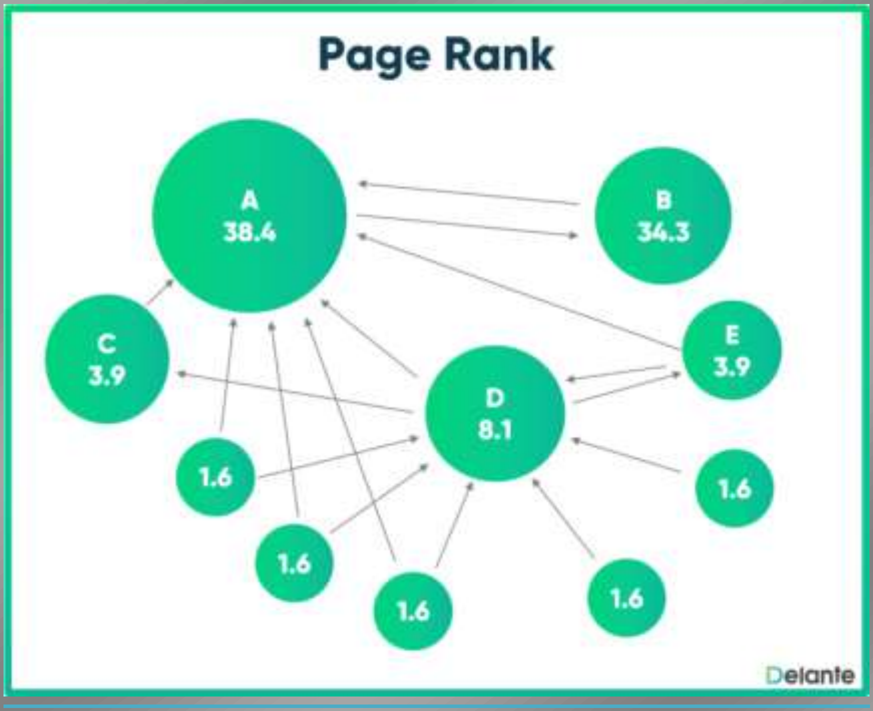
Report on the PageRank Algorithm



Definitions:

"Markov chain":

A stochastic model that describes a sequence of events where the probability of each event depends only on the state reached in the previous event. Named after the Russian mathematician Andrey Markov.

PageRank:

An algorithm used by Google Search to rank web pages in their search engine results. It estimates the importance of web pages by counting the number and quality of links to a page.

Introduction:

Prior to advanced search engines, web pages were ranked using simple metrics such as keyword frequency and metadata, which often led to the manipulation of search results and poor user experience. To address these issues, Google introduced the PageRank algorithm in the late 1990s. PageRank transformed search engine ranking by evaluating not just the content of web pages but also the quality and quantity of links pointing to them.

The PageRank algorithm, created by Larry Page and Sergey Brin during their time at Stanford University, is based on the principle of "voting" or "recommendations" among web pages. A page receives a higher PageRank if it has more incoming links, indicating its importance.

This project aims to implement and understand the PageRank algorithm, explore both its sampling and iterative methods, and evaluate its effectiveness in ranking web pages.

Methods:

-1) Sampling Method:

-This method simulates a random surfer navigating web pages by randomly following links.

- At each step, the surfer either follows a link with a damping factor probability or jumps to a random page in the corpus with a probability of (1 - damping factor).

- This process is repeated for a large number of samples to estimate the PageRank of each page.

-2) Iterative Method:

- In this method, PageRank values of web pages are updated iteratively until they converge.

- Initially, each page is assigned a uniform PageRank value (1/N, where N is the total number of pages).

- The PageRank values are then recalculated based on the previous values and the link structure of the web pages using the PageRank formula.

Equations:

The PageRank formula is:

$$PR(p) = \frac{1-d}{N} + d \sum_{i \in M(p)} \frac{PR(i)}{L(i)}$$

- \(PR(p) \): PageRank of page \(p \)

- \(d \): Damping factor (typically set to 0.85)

- \(N \): Total number of pages

- \(M(p) \): Set of pages that link to page \(p \)

- \(PR(i) \): PageRank of page \(i \)

- \(L(i) \): Number of outbound links from page \(i \)

Implementation and Practical Challenges:

Implementing the PageRank algorithm involved several challenges such as parsing HTML files to extract links, building transition models, and ensuring convergence in the iterative method.

- Parsing HTML: We used regular expressions to extract links from HTML files, which required careful handling to avoid extracting irrelevant information.

- Building Transition Models: Accurate transition models were constructed, taking into account pages with no outgoing links and calculating probabilities for random jumps.

- Convergence: Ensuring convergence in the iterative method involved careful monitoring of PageRank values and updating them until the changes were below a certain threshold.

Results:

Sampling Method:

- Through numerous random walks, the estimated PageRank values for the pages in our test corpus stabilized, accurately reflecting the link structure.

Iterative Method:

- Starting with uniform PageRank values, the iterative method successfully updated these values through multiple iterations until they converged. The final PageRank values were consistent with those obtained from the sampling method, confirming the correctness of our implementation.

Conclusions:

In summary, we successfully implemented the PageRank algorithm using both sampling and iterative methods. This project provided a deeper understanding of how search engines rank web pages and the significance of link analysis in determining page relevance. By achieving the project's objectives, we demonstrated the effectiveness of the PageRank algorithm in ranking web pages based on their importance and popularity.

The project highlighted the robustness of the PageRank algorithm, its mathematical basis in Markov chains, and its practical implications for improving search engine results. Future work could focus on optimizations such as efficiently handling larger datasets and incorporating additional ranking factors to further enhance the accuracy and utility of the PageRank algorithm.

Our team member :

1- Mohamed Gamal Ghareeb	22010215
2- Kamel Mostafa Kamel	22010377
3- Mohamed Ahmed Mohamed	22010211
4- Moataz Mohamed Abdul Hamid	22011663