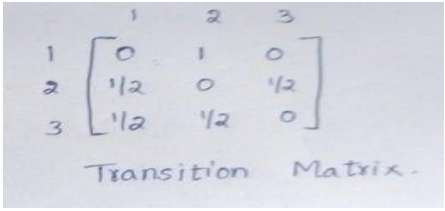
[A] Definitions:

Markov Chain: PageRank is akin to a random surfer navigating the internet, resembling a Markov Chain. It predicts system behavior transitioning from one state to another based solely on the current state, using transition probabilities between states.

How it works:A Markov chain comprises a set of states and a transition probability matrix, where each entry represents the likelihood of transitioning from one state to another.

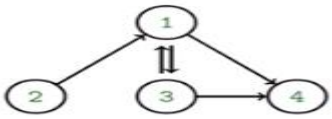
This matrix must adhere to specific properties:

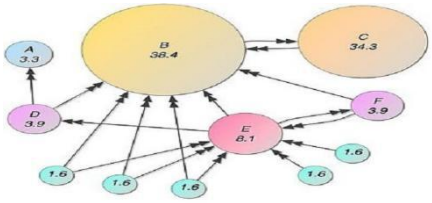
1. Entries range from 0 to 1
2. The matrix is square and non-negative
3. Each row sums to 1
4. Forming a stochastic matrix.



These properties ensure the predictive nature of the Markov chain, facilitating the analysis of state transitions based solely on the current state without considering previous states.

Google PageRank: is a link-based algorithm for ranking web pages, focusing on the interconnectedness of pages within the web graph rather than their content. The web graph represents the World Wide Web as a directed graph where pages are nodes and hyperlinks are edges.





PageRank evaluates a page's importance by considering both incoming and outgoing links, with higher reputation pages contributing more weight to the ranking.

[B] Introduction:

[1]Description of the history of how search engines used to rank important pages:

The vast expanse of the internet presents a challenge in locating relevant information. Search engines play a crucial role in assisting users to navigate through the abundance of

online data. These search engines employ algorithms to assess the relevance of web pages to a user's search query, a process that has evolved significantly over time.

In the early stages of the internet, search engines primarily relied on keywords and metadata to rank web pages. However, as the internet expanded, these simplistic methods were proved to be not enough or inadequate. The necessity for a more sophisticated approach, one that considers the interconnectedness of web pages, became apparent

The advent of PageRank by Larry Page and Sergey Brin in the late 1990s revolutionized search engine optimization. PageRank shifted the focus from mere keyword matching to evaluating the links between web pages. Pages that garnered numerous links from reputable sources were rewarded with higher rankings in search results.

[2] Description of PageRank Algorithm:

The PageRank algorithm, pioneered by Larry Page and Sergey Brin at Google, revolutionized the way web pages are ranked in search engine results. Unlike traditional methods, as discussed earlier, that primarily relied on keyword frequency and metadata, PageRank introduced a novel approach based on the *Analysis of web page interconnections*.

At its core, PageRank operates on the principle of importance by association. It views the web as a network of interconnected pages, with hyperlinks as pathways between them. The fundamental premise is that a page is considered more important if it's linked to by other important pages.

[3] Aims of the Project:

The primary objective or aim of this project is to go through the complexities of the PageRank algorithm and explore avenues for enhancement.

Through deconstructing the algorithm and experimenting with modifications, our aim is to gain a detailed and comprehensive understanding of its functionality and identify opportunities for refinement.

Ultimately, our aspiration is to contribute to the improvement of search engines, facilitating the efficient and accurate retrieval of information for users.

[C] Methods:

In the code, there are several functions that work together to calculate the PageRank scores. The main function serves as the entry point of the program. It prompts the user to input a directory path containing HTML pages. The code then utilizes the crawl function to extract links from these pages and create a corpus, which represents the interconnected web pages.

Sample:

The sample page rank function calculates the PageRank scores using a random sampling approach. It starts by randomly selecting a page from the corpus and performs a specified number of iterations. In each iteration, it updates the sample

count dictionary to keep track of how many times each page has been visited. The transition model function is used to calculate the transition probabilities for moving from one page to another during the sampling process.

Iterative:

The iterate page rank function, on the other hand, uses an iterative approach to calculate the PageRank scores. It initializes the ranks dictionary with initial values and then iteratively updates the scores based on the influence of linked pages. This iteration continues until the difference between the new and old scores falls below a specified epsilon value.

At the end of the main function, the PageRank results are printed for both the sampling and iteration methods. The results provide insights into the importance of each page within the corpus, with pages having higher scores considered more important.

Implementation and practical challenges:

- Ensuring convergence in the iterative PageRank algorithm
- Ensuring the equation are written and computed write

Conclusion:

The journey through the PageRank algorithm unveils its transformative impact on web search, shifting the paradigm from basic keyword matching to sophisticated link analysis. By leveraging the principles of Markov chains, PageRank assigns importance to web pages through their interconnectedness, fundamentally enhancing search accuracy. This project not only demystifies the mechanics behind PageRank but also explores innovative modifications to optimize its performance. As we delve into the intricacies of both sampling and iterative approaches, the potential for refining search algorithms becomes evident, promising a future where users can navigate the vast expanse of online information with unprecedented precision and efficiency.

Citations (each number indicates its order):

1. Teja, R. (2021) Google page rank and Markov chains, Medium. Available at: Here (Accessed: 13 May 2024).

2. What is Google Pagerank? (no date) Page One Power. Available at: Here (Accessed: 13 May 2024).

3. Page rank algorithm and Implementation (2022) GeeksforGeeks. Available at: Here (Accessed: 13 May 2024).