PageRank

Definitions

Markov Chain

A Markov chain is a stochastic process that transitions from one state to another within a finite or countable state space. It is characterized by the property that the probability of transitioning to any particular state is dependent solely on the current state and time elapsed, not on the sequence of events that preceded it. This memoryless property is called the Markov property.

PageRank

PageRank is an algorithm used by search engines to rank web pages in their search results. Developed by Larry Page and Sergey Brin at Stanford University, PageRank assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of measuring its relative importance within the set.

Introduction

History of Search Engine Ranking

Before the advent of sophisticated algorithms like PageRank, early search engines ranked web pages based primarily on the frequency of search terms appearing on the page. This method, however, was easily manipulated by keyword stuffing, where webmasters would artificially inflate the keyword density on their pages to achieve higher rankings. Such tactics often led to poor user experience as the top search results were not always the most relevant or useful.

PageRank Algorithm

PageRank revolutionized the way search engines determined the importance of web pages by analyzing the link structure of the web. Instead of relying solely on content, PageRank considered links between pages as votes of confidence. A page linked by many other pages, especially by those with high PageRank themselves, was considered more important and thus ranked higher. The underlying assumption is that more important pages are likely to receive more links from other websites.

Mathematically, the PageRank of a page PP is calculated using the following formula:

$PR(P)=1-dN+d\sum i=1NPR(Pi)L(Pi)PR(P)=N1-d+d\sum_{i=1NL(P_i)PR(P_i)}$

where:

- $PR(P)PR(P)$ is the PageRank of page PP,
- dd is the damping factor (usually set to 0.85),
- NN is the total number of pages,
- PiP_i represents pages linking to page PP,
- $L(Pi)L(P_i)$ is the number of outbound links on page PiP_i.

Aims of the Project

The objective of this project is to implement the PageRank algorithm using two distinct methods: sampling and iteration. By doing so, we aim to compare the results and evaluate the effectiveness of each method in determining the importance of web pages within a given corpus.

Methods

Sampling Method

The sampling method estimates PageRank values by simulating a large number of random walks on the web graph. Each step in the random walk is determined by a transition model that incorporates both the probability of following a link from the current page and the probability of jumping to a random page in the corpus.

Transition Model

The transition model defines the probability distribution for the next page visit based on the current page. The probability of moving to the next page is influenced by two factors: the damping factor and the number of outbound links on the current page.

With probability dd (the damping factor), the random walker will follow a link from the current page. With probability $1-d1-d$, the walker will jump to any page chosen uniformly at random.

Iterative Method

The iterative method is a deterministic approach to calculating PageRank values by repeatedly updating the rank estimates for each page until convergence. This process starts with an initial guess and iteratively refines the PageRank values based on the structure of the web graph.

Iterative Update Equation

The iterative update equation calculates the PageRank of a page P at iteration $k+1$ using the PageRank values from the previous iteration k:

$$PR(P)^{(k+1)} = \frac{1-d}{N} + d\sum_{i=1}^{N}\frac{PR(P_i)^{(k)}}{L(P_i)}$$

where:

- $PR(P)^{(k+1)}$ is the PageRank of page P at iteration $k+1$,
- d is the damping factor (commonly set to 0.85),
- N is the total number of pages in the corpus,
- P_i represents the pages that link to page P,
- $L(P_i)$ is the number of outbound links on page P_i.

This equation reflects two components:

1. **Random Jump**: The term $\frac{1-d}{N}$ accounts for the probability of randomly jumping to any page in the corpus.
2. **Link Contribution**: The term $d\sum_{i=1}^{N}\frac{PR(P_i)^{(k)}}{L(P_i)}$ represents the contribution of PageRank from pages that link to P, scaled by the damping factor.

Equations

- **Iterative Update Equation**:

$$PR(P)^{(k+1)} = \frac{1-d}{N} + d\sum_{i=1}^{N}\frac{PR(P_i)^{(k)}}{L(P_i)}$$

These equations encapsulate the principles of the iterative method, highlighting its reliance on repeated refinements and convergence checks to produce accurate PageRank values. The method's systematic approach ensures that it effectively captures the link dynamics within a web corpus, providing a robust measure of page importance.

Conclusions

The implementation and analysis of the PageRank algorithm using both sampling and iterative methods have successfully addressed the aims of this project, which include understanding the algorithm, exploring its methods, and evaluating its practical challenges.

Project Aims Addressed

The project aimed to delve into the PageRank algorithm's intricacies, understand its functioning, and assess its practical applicability. These aims have been comprehensively addressed through:

- Detailed descriptions of the algorithm's history and workings, including the evolution of search engine ranking techniques leading to PageRank's development.
- Thorough explanations of the sampling and iterative methods used in PageRank calculation, along with the mathematical equations governing these methods.
- Hands-on implementation of the algorithm, including data parsing, transition probability calculation, and PageRank estimation using both sampling and iterative approaches.
- Identification and analysis of practical challenges encountered during implementation, such as data parsing complexities, computational efficiency issues, convergence criteria selection, and damping factor impact.

The Team Members

1-Yasser Ashraf Mohammed	**22010409**
2-Wael Ahmed	**22010290**
3-Moamen Ahmed	**22010381**
4-Abdelrahman Yousri	**22010364**
5-Abdelrahman Hesham	**22010136**