AMRITA VISHWA VIDYAPEETHAM

MATHEMATICS FOR COMPUTING & ELEMENTS OF COMPUTING

Page rank algorithm and Google Search Engine

Group - 13

Team Members

Adhwaidh K	-	CB.SC.U4AIE24003
Adith S	-	CB.SC.U4AIE24004
Chaitanya Varma	-	CB.SC.U4AIE24017

Introduction

- PageRank is a crucial algorithm for ranking web pages based on their importance using linear algebra.
- It has applications in search engines, social network analysis, academic citations.
- This study explores and compares algebraic (Gauss elimination, matrix multiplication) and iterative (power iteration) methods for PageRank computation.

- Challenges like handling dangling nodes, computational complexity, and slow convergence are addressed.
- The objective is to provide efficient PageRank computations, visualize results, and analyze performance across different methods.

Objectives

- Develop a comprehensive understanding and implementation of PageRank using Gauss elimination, matrix multiplication, and power iteration methods.
- Visualize PageRank distributions for web graphs and explore its applications using NetworkX.
- Implement custom crawling and PageRank computations for real-world data.
- Compare computational methods and analyze performance trade-offs.

Literature Review

S. No.	Title	Advantages	Limitations
1.	The Anatomy of a Large-Scale Hypertextual Web Search Engine	- Introduced a scalable method for ranking web pages. - High accuracy for determining page importance.	- Computationally expensive for large datasets
2.	Google's PageRank and Beyond	- Provided mathematical rigor for PageRank and iterative methods - Focus on matrix algebra optimizations	- Heavy computational burden for extremely large matrices

Literature Review

S. No.	Title	Advantages	Limitations
3.	Network Analysis in Python	- Easy implementation of PageRank with built-in functions - Visualization features for better interpretation	- Limited customization for complex variations of PageRank - Performance issues with very large graphs

Research Gaps

- Limited comparison of matrix algebra and power iteration methods for real-world web data.
- Lack of visualization and user-friendly insights from PageRank outputs.
- Minimal exploration of custom web-crawled datasets in PageRank studies.

Problem Statement

- **Web Growth:**

  Efficient page ranking is needed for large, interconnected networks.

- **Algorithm Efficiency:**

  Current ranking methods face scalability issues.

- **Computational Complexity:**

  Matrix methods are computationally expensive.

- **Convergence Issues:**

  Iterative methods are slow for dense networks.

- **Visualization:**

  Limited tools for interpreting ranking results.

- **Comparative Study:**

  Need for evaluating different page ranking approaches.

Methodology

- **Data Collection:**

  Crawl web pages using Python's requests and BeautifulSoup.

- **Matrix-based Methods:**

  Implement Gauss elimination and matrix inversion for PageRank computation.

- **Iterative Approach:**

  Power iteration with convergence checks.

- **Network Analysis:**

  Visualize PageRanks using matplotlib and NetworkX.

- **Performance Analysis:**

  Compare the accuracy and efficiency of each method.

Timeline

- **Week 1:** Literature review and problem formulation
- **Week 2-3:** Implement algebraic methods and power iteration
- **Week 4:** Integrate web crawling and custom graph creation
- **Week 5:** Visualization and performance comparison
- **Week 6:** Documentation and report preparation

Thank You!