Implementation and Analysis of PageRank Algorithms

Chunbo Luo

September 24, 2023

1 Introduction

The aim of this workshop is to address several important problems you may encounter in order to understand and implement Google PageRank algorithm. This algorithm is based on "citation" that is actually represented by the links between pages.

Figure 1: Google's PageRank Algorithm

Please use Python to implement the algorithm and solve the following tasks.

2 Tasks

- **Task 1**

  Design a fully-connected directed graph where nodes are linked with each other. The nodes denote web pages and the directed links represent that the source web page contains a URL linked to the destination web page. An example is Figure 2.

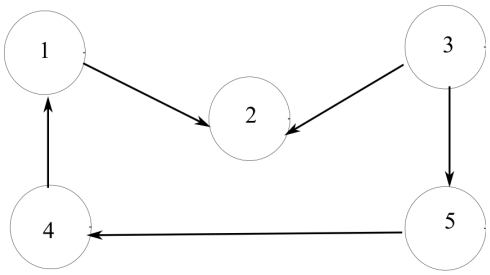

Figure 2: A fully connected directed graph.

Derive the linkage matrix (i.e., transition matrix in the lecture) from this graph, choose several sets of initial values to represent the ranks of all the pages, and calculate their page ranks respectively. *You can use the extra page in the end.*

- Can you achieve stable page ranks for all the cases?
- Most importantly, do the initial values affect the final results and why?
- Does the final result make sense in real life and why?

- **Task 2**

  Design a different directed graph that includes the following two cases:

  - Some nodes (web pages) do not link to any other nodes;
  - The graph can be divided into two or more subgraphs;

  These are also called "black hole" nodes and "isolated" groups of nodes, respectively.

  Find out the way to address these two issues and calculate the stable final page ranks for the web pages, which is independent of initial values.

Extra tasks

- **Task 3**

  For a fixed set of initial values, change the parameter d "damping factor" for addressing the issue of "isolated" group of nodes in the previous task and discuss how the parameter affects your results and why.

  For a fixed parameter d, change the set of initial values, record how many steps your program needs to run in order to obtain the final results and discuss why.

- **Task 4**

  Discuss how to get the linkage matrix in real-world situation. What will you do if you want to rank all the pages related to Exeter?

3 Appendix

Read the lecture slides and validate your code using the example from the slides.