**GOOGLE PAGERANK ALGORITHM**

1. **Introduction**

PageRank is one the of oldest and most important algorithms of Google. Google Search uses it to rank web pages in its search engine. PageRank is created by Larry Page and Sergey Brin, founders of Google. PageRank is a way of measuring the importance of each website pages on the internet. Thanks to the superiority of PageRank and others algorithms, Google has beated competitors such as Yahoo or Bing to become the word’s best search engine.

According to Google: “PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.” (Source: Wikipedia – PageRank).

This algorithm is very complicated to the one who just study data structure for a little time, but it’s a good application for Graph data structure implement in real life. So in this section, we will just clarify some basics about PageRank algorithms which directly relate to Graph data structure.

1. **Algorithm implementation**

PageRank value (sympol: PR(E)) is calculated by webgraph – Graph with the vertex is the website and the edge is the link to the website (these websites include some government website .gov and some authority else). Each link to a website will be counted as a increasing of value PageRank of that website. So, a website which has many links from other high PR(E) value website will have a high PR(E) value too, if a website just has some small link’s PR(E) value link to it, its PR(E) value will not sure to be high).

PageRank also represents for probability that a user randomly clicks on the link on the internet websites. In PageRank, a website which PR(E) = 0.5 means: 50% chance of a user clicking on a random link to go to that website (of course in real life this percent hard to be happen).

Example of website’s PR(E):

A picture containing diagram

Description automatically generated

In this example, I use value percent for easy implementation. Website C has 11% that user will randomly click to the link which go to it, website B has 39%. Because website B just has one link but that link is from a important website (website A – 45%) so PR(E) of website B is very high. Website C has many links but its links are not very important with low PR(E).

1. **Algorithm formula**

Now we will prove the formula for the PageRank algorithm.

Assume that on the internet has four website: A, B, C, D. So at the beginning, no links are implement so the percent for each website is 0.25 (25%).

A picture containing text, pool ball, table

Description automatically generated Example 1:

When website B, C, D have a link to website A, each website will tranfers its PR(E) to website A. So,

A picture containing text, pool ball, vector graphics

Description automatically generatedExample 2:

Now website B has another link to website C, website D has another links to website C and B. Website B must divide its PR(E) by 2 to tranfers its value to two website C and A. Website D must divide its PR(E) by 3 to tranfers its value to website A, B and C. So PR(A) now:

In general, we have the formula for PageRank value to any website U:

v is the website that has a link to website U. B(U) is the set of many v. L(v) is the number of links that goes from website v.

1. **Algorithm in coding**

According to the formula above, we can turn it to code. To calculate the PageRank value of a website: In a directed graph, detect the root vertex that stand for the website. Then we must to find out each vertex that adjacent to root vertex and number of edges begin of that vertex. We can use BFS algorithm to go to all adjacent vertices to vertex root, but remember to stop when the root vertex has no more unvisited predecessor.

Pseudo-code:



**IMAGE IMPLEMENTATION IN COMPUTER**

In computer, we can use 2D array data structure to store the value that creates the image: pixel.