PageRank algorithm

Yogesh Agarwala
EE19B130

Introduction

PageRank is an algorithm Google Search uses to rank webpages in their search engine results. Even though PageRank is used for ranking pages, it is named PageRank after its discoverer and developer, Larry Page, who was one of the co-founders of Google. The motivation behind this algorithm was that – before google, when we performed a standard search on the internet using keywords or textual patterns, we would get back millions of hits, most of which were really low quality and with a few valuable pages buried in the millions.

Page came up with the idea of using the web structure of the internet to identify important documents. We can think of the whole internet as a graph where a user is on a page(node) with a link(directed edge) to another page, and users are kind of randomly travelling around the world wide web. A page is hypothesized to be “more important” if viewed a significant fraction of the time by these random browsers.

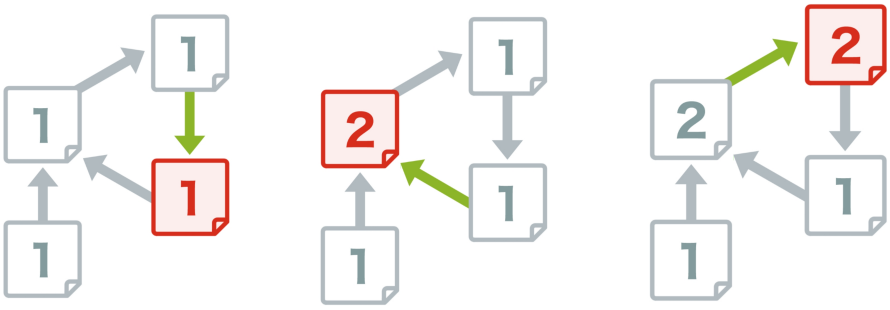
Description

PageRank works by counting the number of links and the quality of each link directed towards a page to estimate the website's importance. The underlying assumption is that important websites are more likely to get referred from other websites. In the below images, the squares are webpages, and the arrows represent the links between them.

In the PageRank algorithm, the more links a webpage has pointing to it, the more important that page is determined to be. We'll give the pages that have no links pointing to them a score of 1. So the score of a page with links pointing to it is the combined score of the pages pointing to it. A link from a highly linked webpage has a high value, and it's said to be the most important page.

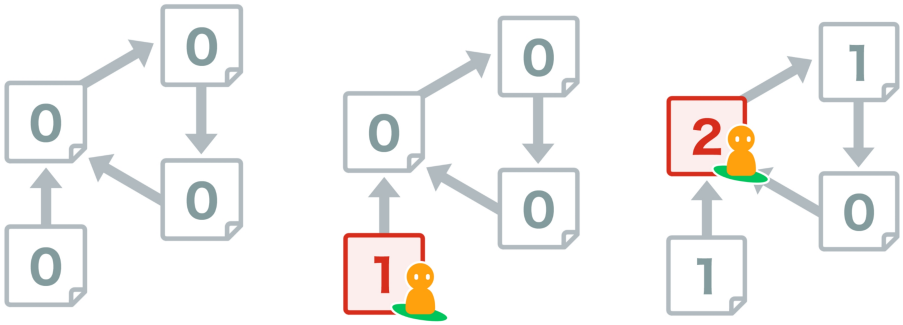
But what if the links form a loop?

If we calculate the score of each page in order, as we can see, it loops infinitely, and the scores of the pages in the loop increase without end. This loop problem is resolved using a calculation method called the “random surface model”.



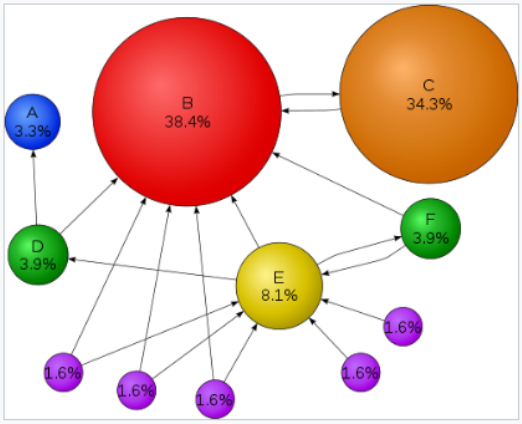
We assume there is a surfer who starts from a random webpage and keeps clicking on links until he eventually gets bored. Then he starts on some other random page and does the same. The probability that a surfer visits a given page is taken as the score for PageRank.

Let α be the probability that the person will teleport to some other page. Then 1-α is the probability that one link among those on the current page will be chosen. We'll consider what happens when the links form a loop. The number on each page represents the number of times the web surfers have visited that page.



Let's speed up the passage of time. We can run the simulation until the total number of page visits reaches 1000 and express the results as percentages. These values express the probability that someone is viewing the page at a given point in time. Using this method, we can calculate a score even if the links create a loop.

Applications



A simple illustration of the Pagerank algorithm. The percentage shows the perceived importance, and the arrows represents hyperlink.

1. **Rank web pages:** The most popular use of PageRank is, of course, to determine the relevance of web pages. Though it is not the only algorithm used by search engines to order search results but it was the first one to be used and is still a relevant one.
2. **Predicting traffic in urban cities:** Taking streets as nodes and intersections as edges, we can represent the city as a well-connected graph. Applying PageRank on this graph has been observed to predict the traffic flow on individual roads as well as on connected roadmaps quite accurately.
3. **PageRank in Biology - GeneRank, ProteinRank, IsoRank:**. Most of these applications use PageRank to reveal localized information about the graph based on some form of external data. E.g. In GeneRank, the idea is to use a graph of known relationships between genes to find genes that are highly related to those which are promoted/repressed in the experiment.
4. **Neuroscience:** We can generate an undirected network using fMRI scans where we treat voxels as nodes and the time-correlation between the voxels by edges. Applying a version of PageRank developed for undirected graphs, neuroscientists could identify areas of the brain that vary together as the subject ages.
5. **PageRank in Mathematical Systems:** Graphs and networks arise in mathematics to abstract the properties of systems of equations and relationships between simple sets.

Complexity

Let N - number of nodes (web pages)
E - number of edges (links)
Then the time complexity of PageRank is approximately O(N+E)

Because even in a complete graph, PageRank has to touch at least every node, which is O(N) and two times each edge which is O(E). So the Big-O complexity of PageRank will be O(N+E). There are several assumptions while calculating this time complexity, like we are running PageRank on a given graph and not re-calculating all the values each time a node is added because then the complexity can be as bad as O(N*E)