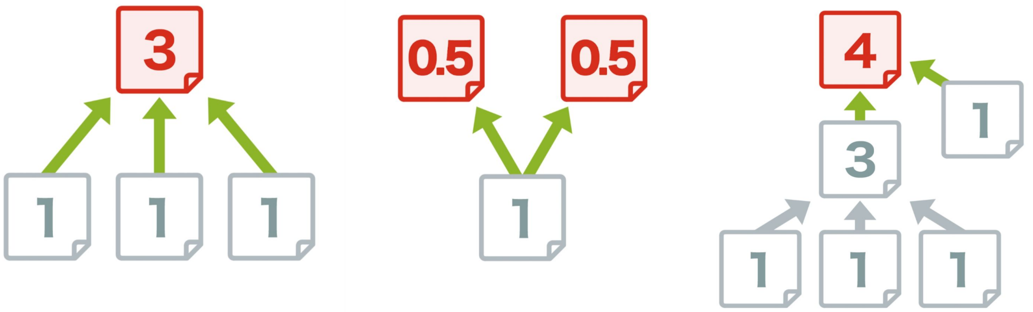
Problem 8(i): Pagerank

Weighted link method: The basic thinking behind PageRank

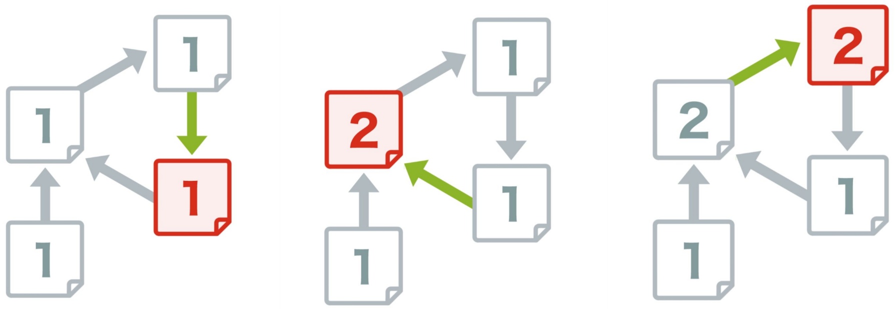
1. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.
2. Let's say in the below images, the squares are webpages and the arrows represent the links between them.
3. In the PageRank algorithm, the more links a webpage has pointing to it, the more important that page is determined to be.



4. 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.
5. A link from a highly linked webpage has a high value, and it's said to be the most important page.

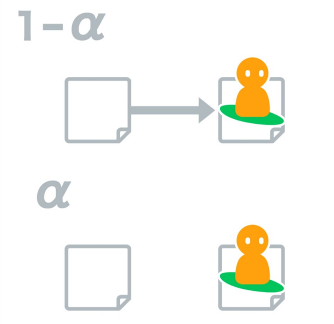
Random surface model: solves the problem when links forms a loop

1. 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 increases without end.

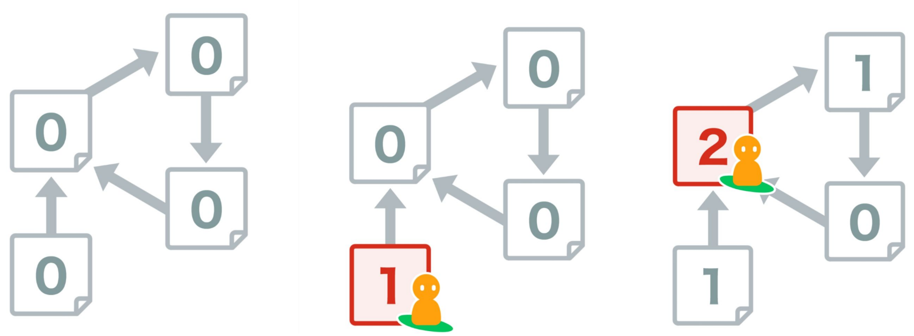


2. This loop problem is resolved using a calculation method called the random surface model.

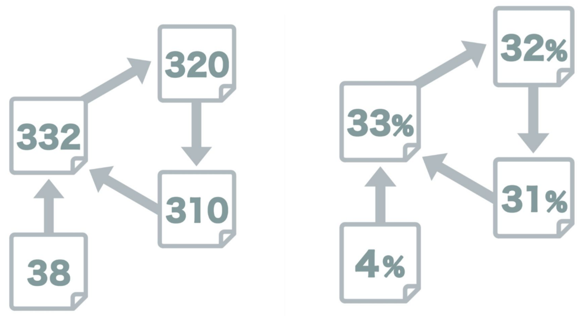
3. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the d damping factor is the probability that at each page the "random surfer" will get bored and request another random page.
4. If we define the actions of web surfers, we get something like the following.
5. Probability 1-a is the probability that one link among those on the current page will be chosen.
6. Probability a is the probability that the person will teleport to some other page.



7. As before we'll consider what happens when the links form a loop
8. The number on each page represents the number of times the web surfers has visited that page.



9. 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 **(In reality, more practical calculations methods are used rather than simulations, discussed in the next topic).**



10. We can say that these values express the probability that someone is viewing the page at a given point in time.
11. Using these values for webscores is the random surface model’s method.
12. As we can see by using this method we can calculate a score even if the links create a loop

PageRank Calculation:

The original PageRank algorithm was described by Lawrence Page and Sergey Brin in several publications. It is given by

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

where

PR(A) is the PageRank of page A,
PR(Ti) is the PageRank of pages Ti which link to page A,
C(Ti) is the number of outbound links on page Ti and
d is a damping factor which can be set between 0 and 1.

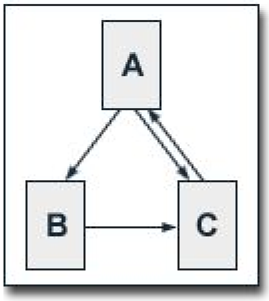
So, first of all, we see that PageRank does not rank web sites as a whole, but is determined for each page individually. Further, the PageRank of page A is recursively defined by the PageRanks of those pages which link to page A.

The PageRank of pages Ti which link to page A does not influence the PageRank of page A uniformly. Within the PageRank algorithm, the PageRank of a page T is always weighted by the number of outbound links C(T) on page T. This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T.

The weighted PageRank of pages Ti is then added up. The outcome of this is that an additional inbound link for page A will always increase page A's PageRank.

Finally, the sum of the weighted PageRanks of all pages Ti is multiplied with a damping factor d which can be set between 0 and 1. Thereby, the extent of PageRank benefit for a page by another page linking to it is reduced.

Calculation Example:



We regard a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. According to Page and Brin, the damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5. The exact value of the damping factor d admittedly has effects on PageRank, but it does not influence the fundamental principles of PageRank. So, we get the following equations for the PageRank calculation:

PR(A) = 0.5 + 0.5 PR(C)
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

These equations can easily be solved. We get the following PageRank values for the single pages:

PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

It is obvious that the sum of all pages' PageRanks is 3 and thus equals the total number of web pages. As shown above, this is not a specific result for our simple example. In practice, the web consists of billions of documents and it is not possible to find a solution by inspection. Because of the size of the actual web, the Google search engine uses an approximative, iterative computation of PageRank values.

Conclusion:

Though in reality google search results are not arrived at by PageRank alone. However, that doesn’t change the fact that the PageRank algorithm was revolutionary in its two concepts:

- Those of calculating webpages values from link structures
- And they can carry out calculations even when links form loops.