Outline

- PageRank – Introduction
- Page Ranking Algorithm
- Different scenario of Page Ranking Algorithm
- Weighted Page Rank Algorithm
- Matching
- Edge Cover
- Min-max Theorems
- Min Cover & Max Matching

Motivation

When searching for information on the WWW, user perform a query to a search engine. The engine return, as the query's result, a list of Web sites which usually is a huge set. So the ranking of these web

sites is very important. Because much information is contained in the link-structure of the WWW, information such as which pages are

linked to others can be used to augment search algorithms.

PageRank - Introduction

PageRank is a link analysis algorithm which assigns a numerical weighting to each Web page, with the purpose of "measuring" relative importance.

- Based on the hyperlinks map
- and web keyword searches

Why Page Ranking Algorithm?

Page ranking algorithms are used by the search engines to present the

- search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest.

History of Page Ranking Algorithm

- PageRank was developed by Larry Page (hence the name *Page*-Rank) and Sergey Brin.
- It is first as part of a research project about a new kind of search engine. That project started in 1995 and led to a functional prototype in 1998.
- PageRank algorithm is used by the famous search engine that is Google. This algorithm is the most commonly used algorithm for ranking the various pages.

Page Ranking Algorithm and Data Mining

Page Ranking algorithm is belongs to the category of "**Web Structure Mining**".

Web Structure Mining (WSM) generates the structural summary about the Web site and Webpage. It tries to discover the link structure of the hyperlinks in inter documents level. So web structure mining categorizes the web pages on the basis of the hyperlink and finds the similarity and relationship of information between different Web sites. This type of mining can be performed at intra-page or at inter-page (hyperlink level).

Algorithm

The Page Rank algorithm is given by

1) Calculate page ranks of all pages by following formula:

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

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, but it is usually set to 0.85

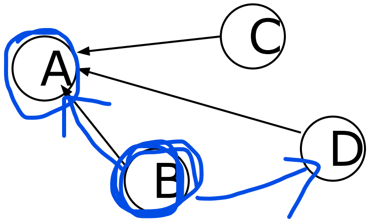
2) Repeat step 1 until values of two consecutive iterations match.

Simplified PageRank algorithm

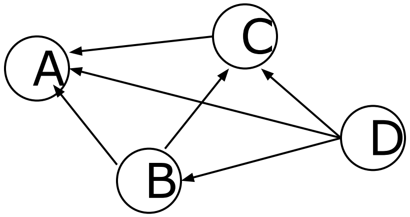
A

- □ Assume four web pages: **A**, **B**,**C** and **D**. Let each page would begin with an estimated PageRank of 0.25.

$PR(A) = (1-d) + d\left(\frac{PR(B)}{C(B)} + \frac{PR(D)}{C(D)} + \frac{PR(C)}{C(C)}\right)$



$$PR(A) = PR(B) + PR(C) + PR(D).$$



$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$

- □ L(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$

PageRank algorithm including Damping Factor

We assume page A has pages $T_1 ... T_n$ which point to it i.e., are links. The variable d is a damping factor, which value can be set between 0 and 1. We usually set the value of d to 0.85. $PR(T_1)$ is the incoming link to page A and $C(T_1)$ is the outgoing link from page T_1 (such as $PR(T_1)$). The PageRank of a page A is given by the following (1):

$$PR(A) = (1-d) + d(PR(T_1) /C(T_1) +...+PR(Tn /C(Tn)) \qquad (1)$$

- The damping factor is used to stop the other pages having too much influence; this total vote is damped down by multiplying it by 0.85.

Intuitive Justification

- 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.
  - The d damping factor is the probability at each page the "random surfer" will get bored and request another random page.
- A page can have a high PageRank
  - If there are many pages that point to it
  - Or if there are some pages that point to it, and have a high PageRank.

Random Surfing

The page ranks form a probability distribution over web pages, so the sum of all web pages' page ranks will be one and the d damping factor is the probability at each page the random surfer will get bored and request another random page. Another simplified version of PageRank is given by:

$$PR(u) = C \sum_{v \in B(u)} \frac{PR(v)}{N_v} \qquad (1)$$

Where, u represents a web page,

B(u) is the set of pages that point to u,

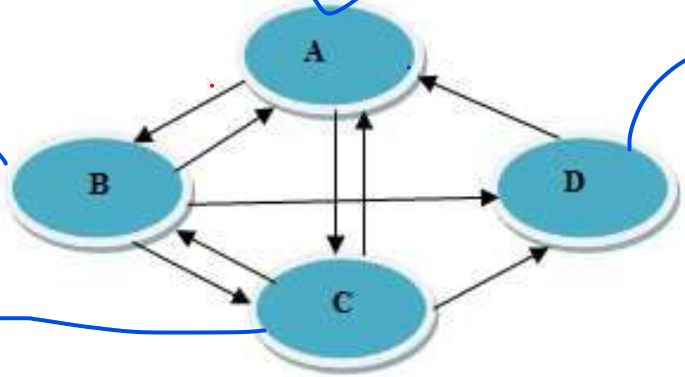
PR(u) and PR(v) are rank achieves of page u and v respectively,

N_v indicates the number of outgoing links of page v, c is a factor applied for normalization.

Hyperlink Structure

Let us take an example of hyperlink structure of four pages *A*, *B*, *C* and *D* as shown in Figure. The PageRank for pages *A*, *B*, *C* and *D* can be calculated by using (1).

$\frac{PR(B)}{C(B)} = \frac{1}{3}$



$\frac{1}{3}$

ভ্যালু

$\frac{1}{1}$

লিংক

Figure 2: Hyperlink structure of four pages

Calculation of First Iteration

Let us assume the initial PageRank as 1 and do the calculation. The value of damping factor d is put to 0.85.

$$\begin{aligned} \boxed{PR(A)} &= (1-d) + d\,(\boxed{PR(B)/C(B)}+\boxed{PR(C)/C(C)}+\boxed{PR(D)/C(D)}) \\ &= (1-0.85) + 0.85(1/3+1/3+1/1) \\ &= \boxed{1.5666667}\checkmark \end{aligned} \tag{3}$$

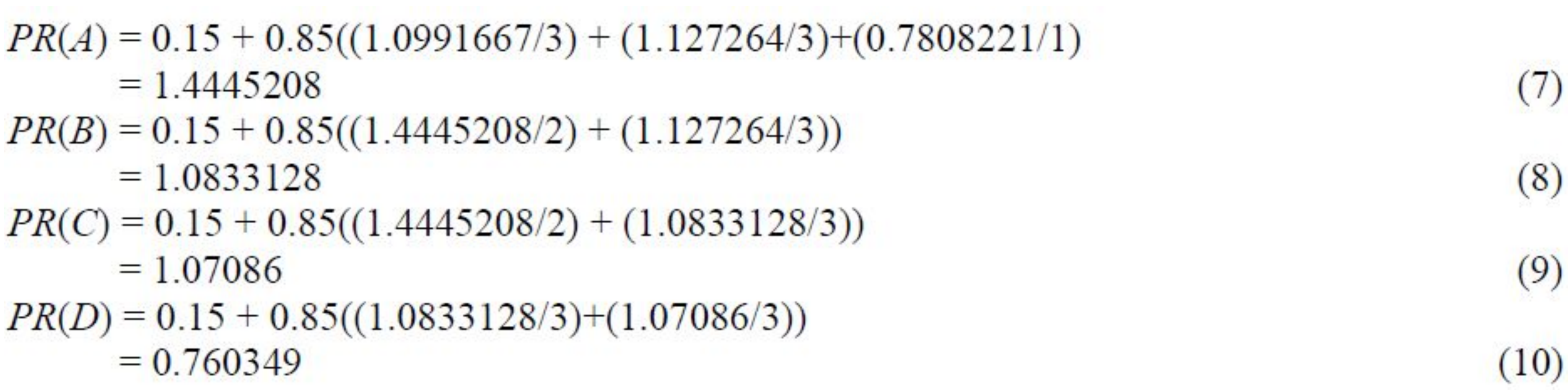
$$\begin{aligned} PR(B) &= (1-d) + d((\boxed{PR(A)}/C(A) + (PR(C)/C(C)) \\ &= (1-0.85) + 0.85(\boxed{1.5666667}/2+1/3) \\ &= \boxed{1.0991667} \end{aligned} \tag{4}$$

$$\begin{aligned} PR(C) &= (1-d) + d((PR(A)/C(A) + (PR(B)/C(B)) \\ &= (1-0.85) + 0.85(\boxed{1.5666667}/2+\boxed{1.0991667}/3) \\ &= 1.127264 \end{aligned} \tag{5}$$

$$\begin{aligned} PR(D) &= (1-d) + d((PR(B)/C(B) + (PR(C)/C(C)) \\ &= (1-.085) + 0.85(1.0991666/3+1.127264/3) \\ &= 0.7808221 \end{aligned} \tag{6}$$

Calculation of Second Iteration

For the second iteration by taking the above *PageRank* values from (3), (4), (5) and (6). The second iteration PageRank values are as following:

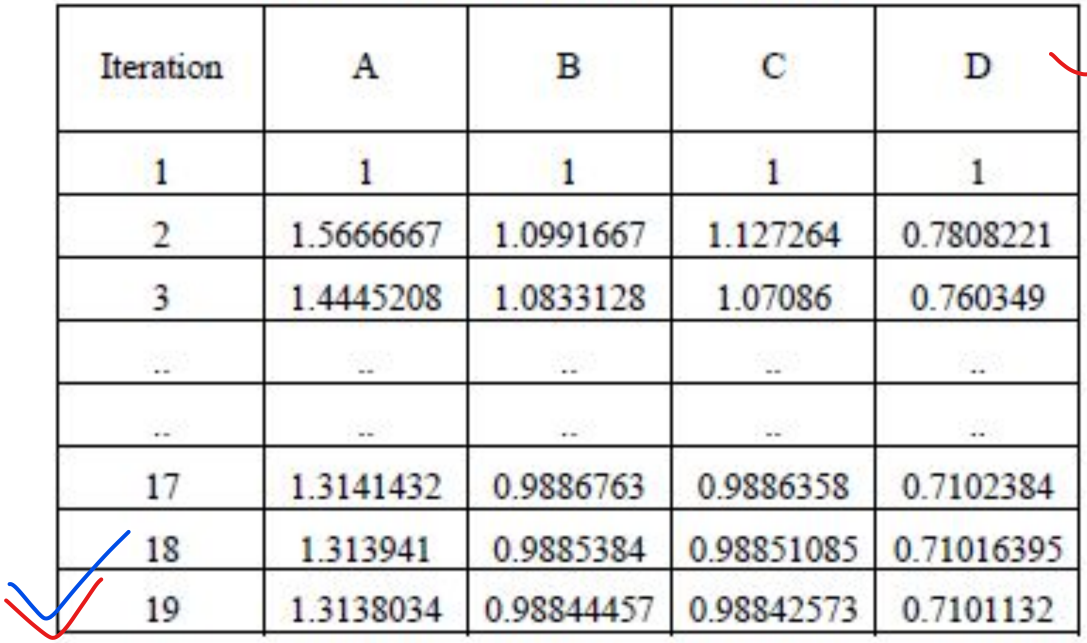


$$PR(A) = 0.15 + 0.85((1.0991667/3) + (1.127264/3)+(0.7808221/1)$$
$$= 1.4445208 \quad (7)$$
$$PR(B) = 0.15 + 0.85((1.4445208/2) + (1.127264/3))$$
$$= 1.0833128 \quad (8)$$
$$PR(C) = 0.15 + 0.85((1.4445208/2) + (1.0833128/3))$$
$$= 1.07086 \quad (9)$$
$$PR(D) = 0.15 + 0.85((1.0833128/3)+(1.07086/3))$$
$$= 0.760349 \quad (10)$$

Iteration (Conti…)

During the computation of 34th iteration, the average of the all web pages is 1. Some of the PageRank values are shown in Table 1. The table with the graph is shown in the simulation results section.

Table 1. Iterative Calculation for PageRank



Iteration	A	B	C	D
1	1	1	1	1
2	1.5666667	1.0991667	1.127264	0.7808221
3	1.4445208	1.0833128	1.07086	0.760349
..	..	..	..	..
..	..	..	..	..
17	1.3141432	0.9886763	0.9886358	0.7102384
18	1.313941	0.9885384	0.98851085	0.71016395
19	1.3138034	0.98844457	0.98842573	0.7101132

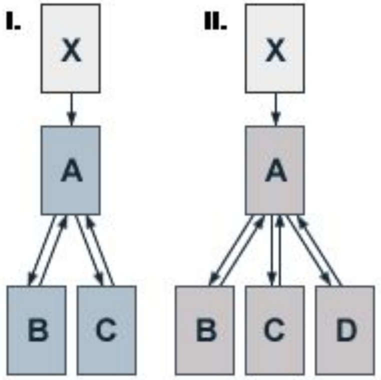
Extraction of Table-1

In the Table 1, we can notice that PageRank of *A* is higher than PageRank of *B*, *C* and *D*. It is because Page *A* has 3 incoming links, Page *B*, *C* and *D* have 2 incoming links as shown in Figure 2. Page B has 2 incoming links and 3 outgoing link, page C has 2 incoming links and 3outgoing links and page *D* has 1 incoming link and 2 outgoing links. From the Table 1, after the 34th iteration, the PageRank for the pages gets normalized.

Add new pages to your website

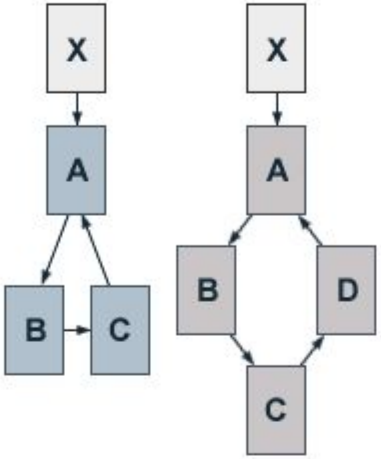
- When you add a new page to your site, be sure to link it to your front page and vice versa as it is shown on the picture.

Right



- If you want to reduce your front page's PageRank, then you can make circular references as you see on the second picture

Wrong



The effect of additional pages

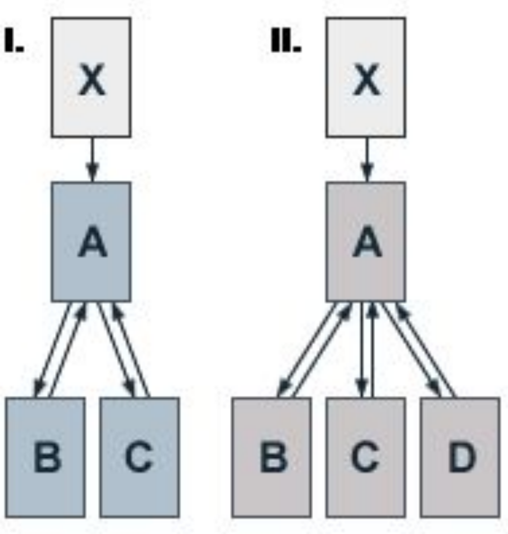
Sub-pages	PageRank of the front page
1	1.000000
2	1.428673
3	1.857347
4	2.286020
5	2.714694
10	4.858060
20	9.144795
50	22.005003
100	43.438648
250	107.739838
500	214.907135
700	300.642426
1000	429.246613

The effect of additional pages

- As you can see:

PageRank ≈ 1+0.428*NumberOfPages

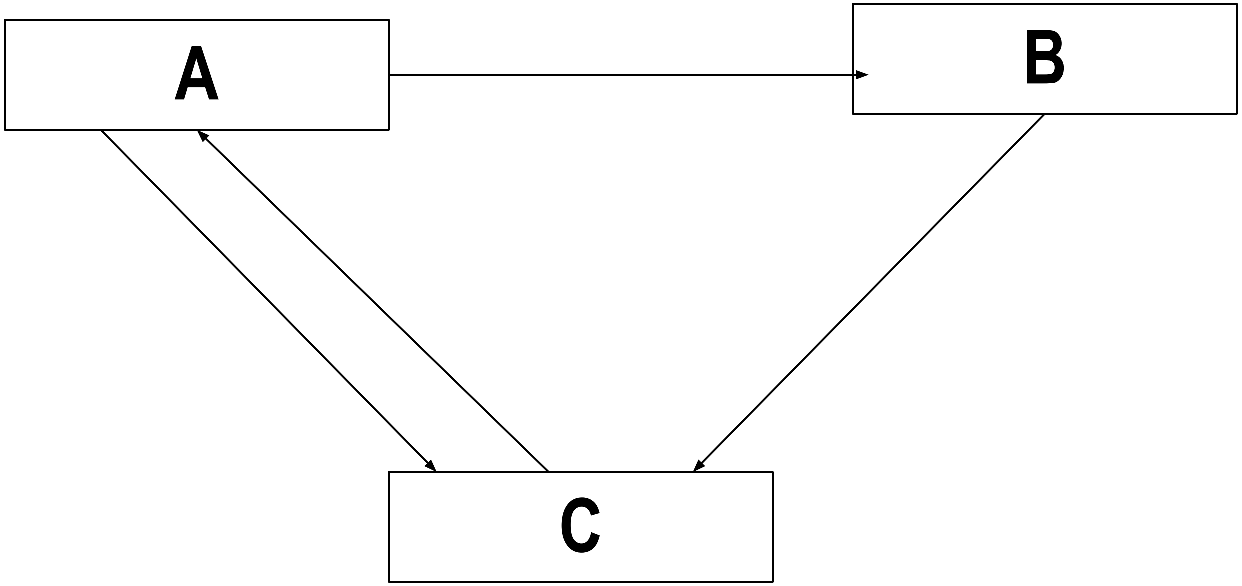
- So, if you add a web page to your website it will increase your page's rank by ≈0.428. Of course you need to do as it is shown on the picture



The effect of additional pages

- The problem with this method is that if you increase your front page’s PageRank by adding additional pages, than the rank of your other pages will go down.

Home Work



Results

Iterations	A	B	C
0	1	1	1
1	1	0.575	1.06375
2	1.0541875	0.5980296875	1.06354922

Assignment