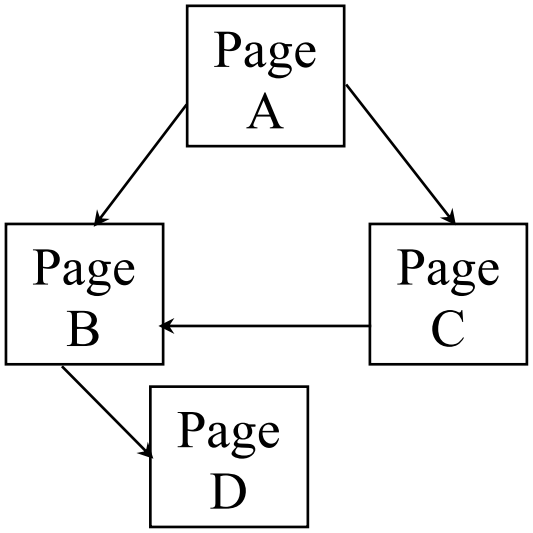
Exercices - RI sur le Web - correction

Exercice 1 – Calculs de HITS

Calculer les valeurs d'HUB et d'AUTHORITY des pages web des configurations suivantes :



Commentez vos résultats au bout de 4 boucles en plus de l'initialisation.

Les formules sont :
$AUTH_{NN}[A] = 0$
$AUTH_{NN}[B] = HUB[A] + HUB[C]$
$AUTH_{NN}[C] = HUB[A]$
$AUTH_{NN}[D] = HUB[B]$

$HUB_{NN}[A] = AUTH[B] + AUTH[C]$
$HUB_{NN}[B] = AUTH[D]$
$HUB_{NN}[C] = AUTH[B]$
$HUB_{NN}[D] = 0$

Comme demandé, on s'arrête <u>après 4 boucles</u> (les lignes en jaunes sont avec normalisation, étape 2 vue en cours).

	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	AUTH[C]	HUB[C]	AUTH[D]	HUB[D]	Rac(AUT HS^2)	Rac(HU BS^2)
Init	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50		
1.1	0	1,00	1,00	0,50	0,50	0,50	0,50	0	1,224	1,224
1.2	0	0,82	0,82	0,41	0,41	0,41	0,41	0,00		
2.1	0,00	1,22	1,22	0,41	0,82	0,82	0,41	0,00	1,53	1,53
2.2	0,00	0,80	0,80	0,27	0,53	0,53	0,27	0,00		
3.1	0,00	1,34	1,34	0,27	0,80	0,80	0,27	0,00	1,58	1,58
3.2	0,00	0,85	0,85	0,17	0,51	0,51	0,17	0,00		
4.1	0,00	1,35	1,35	0,17	0,85	0,85	0,17	0,00	1,60	1,60
4.2	0,00	0,84	0,84	0,11	0,53	0,53	0,11	0,00		

On remarque que B a la plus haute autorité (normal car 2 pages pointent sur elle) et que A a le plus haut hub car pointe sur 2 pages. La page C a une haute autorité car la page B et la page A pointent dessus. La page A a la plus grande valeur de HUB car elle point sur 2 pages, et en particulier sur B qui a la meilleure autorité.

Exercice 2 – Calculs de Pagerank

Reprendre l'exercice 1 avec le calcul de 5 boucles de pagerank.

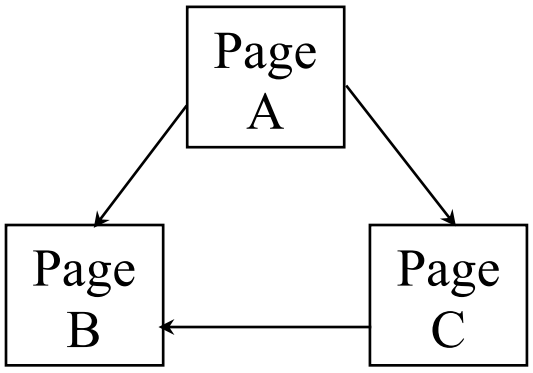
On utilise d = 0.85 comme vu en cours, et on initialise avec 1.

Les formules sont :
PR[A] = 0.15
PR[B] = 0.15 + 0.85 * (PR[A]/2 + PR[C]/1)
PR[C] = 0.15 + 0.85 * (PR[A]/2)
PR[D] = 0.15 + 0.85 * (PR[B]/1)

	A	B	C	D
init	1	1	1	1
1	0,15	1,425	0,575	1
2	0,15	0,7025	0,21375	1,361
3	0,15	0,3954	0,21375	0,7471
4	0,15	0,3954	0,21375	0,48612
5	0,15	0,3954	0,21375	0,48612

D est la page la plus populaire car toutes les pages lui pointent dessus directement ou indirectement, alors que A est la moins populaire. B est plus populaire que C car elle est pointée par C et A.

Exercice 3 – Intégration de Pagerank dans une correspondance vectorielle
Prenons la configuration suivante :



Reprendre les documents de l'exercice 3 de première feuille de TD. On considère que le document 1 est la page A, le document 2 la page B et le document 3 la page C.
Si nous posons que la valeur de pertinence finale est :

Pagerank * valeur de correspondance vectorielle

Donner le résultat de la requête q1 du même exercice 3.

1. On commence par faire les calculs de Pagerank :
Les formules :
PR(A) = 0.15
PR(B) = 0.15 + 0.85*(PR(A)/2 + PR(C)/1)
PR(C) = 0.15 + 0.85*(PR(A)/2)

Le tableau des valeurs :

PR(A)	PR(B)	PR(C)
1	1	1
0,150	1,425	0,575
0,150	0,703	0,214
0,150	0,395	0,214
0,150	0,395	0,214

Elements tirés de l'exercice 3 du TD1. (cf. correction)
--- début
Tableau des vecteurs avec les normes :

	t1	t2	t3	t4	t5	t6		norme
d1	1,00	0,00	1,00	0,00	0,00	1,00		1,73
d2	3,00	0,00	2,00	1,00	0,00	1,00		3,87
d3	1,00	2,00	3,00	0,00	1,00	0,00		3,87
q1	2,00	0,00	2,00	0,00	0,00	0,00		2,83

Calcul de la similarité Sim par cosinus de l'ex3 :

cos	d1	d2	d3
q1	1,00	0,94	0,73

--- fin
Si on prend en compte les PR qu'on vient de calculer, et en posant que la Page A est d1, la Page B est d2 et la Page C est d3, on a :

	Sim(q1, d.)	Pagerank	**Valeur finale calculée**
d1	1.00	0,15	**0,15**
d2	0,94	0,395	**0,371**
d3	0,73	0,214	**0,156**

Résultat de la requête q1 est donc :
d2, puis d3, puis d1.
On voit que l'ordre est très différent avec ou sans Pagerank. Avec Pagerank le document d1 n'est plus la meilleure réponse, car il n'est pas populaire du point de vue de PageRank.