La recherche d'information sur le Web

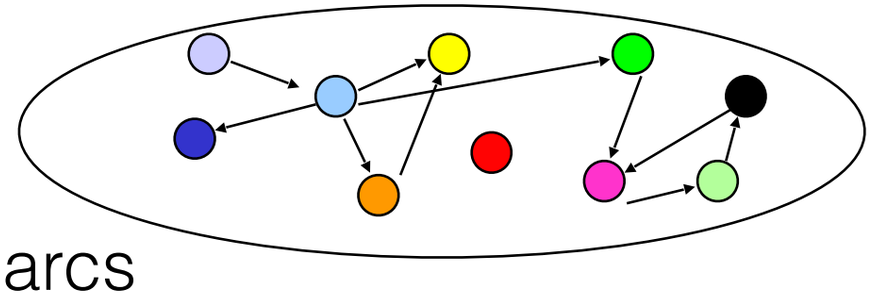
ISN 2 - 2019/2020

Lorraine Goeuriot (LIG, UGA)

lorraine.goeuriot@imag.fr
http://mrim.imag.fr/User/lorraine.goeuriot/isn2

(diapos tirées d'un cours de Philippe Mulhem - LIG)

Introduction - le Web



- Modélisation du Web : un graphe
  - les documents sont des nœuds
  - possibilités de parcours avec des liens, les arcs
- Présentation du Web : un système hypertexte
  - affichage des nœuds en indiquant les liens vers d'autres documents
  - lors d'un clic, affiche le nœud qui est la cible du lien dont l'ancre est la source.
- Avantages :
  - Généralité : tout type de structure, tout types de liens
  - Grande souplesse et simplicité d'utilisation (des clics souris)
- Inconvénients
  - Structure non explicite : perte de repères (séquentialité, hiérarchie)
  - Désorientation dans la navigation

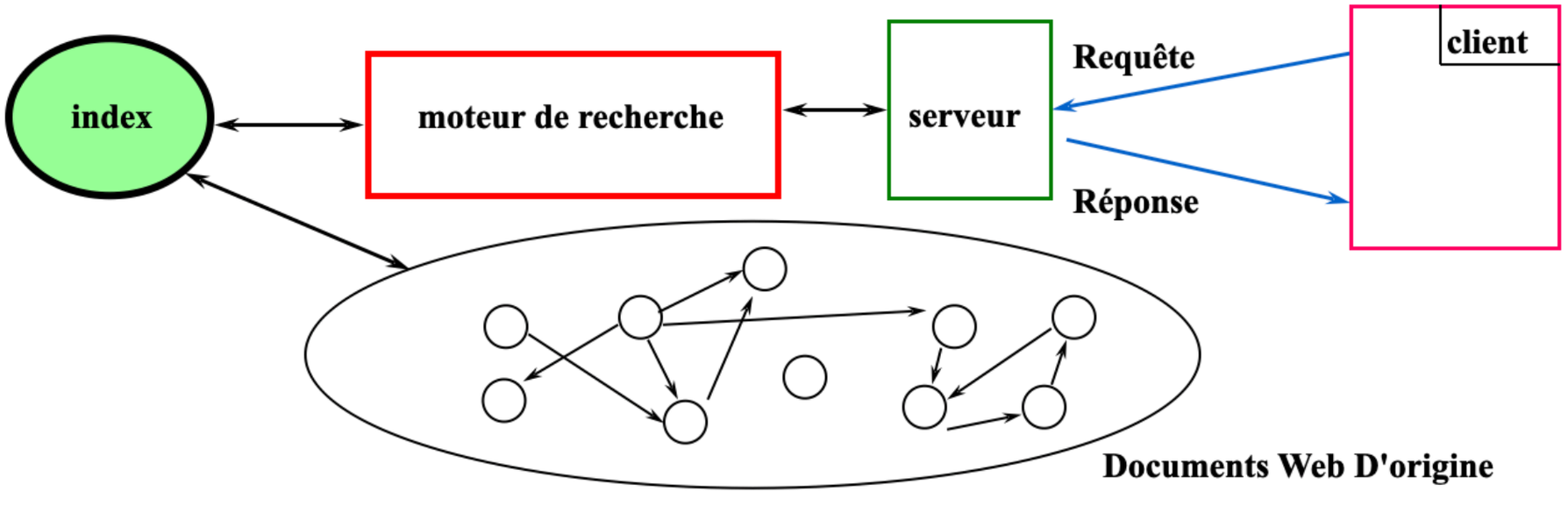
Introduction - le Web

- Rechercher sur le Web - problématique :
  - Quantité de données quasiment illimitées
  - On ne cherche pas à savoir s'il existe un document parlant d'un sujet, mais plutôt comment y accéder facilement
- Rechercher sur le Web - méthode :
  - Comme pour la RI classique :
    1. Extraction du sens des documents
    2. Représentation "sémantique" des documents utilisée comme base de la recherche

RI sur le Web par interrogation

Même principe que les SRI classiques :

- on indexe les pages Web (analyse du contenu)
- on stocke les index et les adresses (URL) des pages Web
- on fait correspondre le résultat de l'analyse de chaque requête avec les représentations des documents.



RI sur le Web - Indexation

Indexation classique par "robots" en deux phases :

1. Découverte dynamique du corpus
2. Indexation des pages trouvées

Exemple: robot utilisant un ensemble d'URL E

```
tant que E non vide :
    accéder à une page p d'URL e ∈ E
    retirer les marqueurs HTML de p
    indexer le contenu de p
    E = (E - e)  U cibles ℗
```

RI sur le Web - Matching

Calcul de correspondance classique :

- Correspondance entre les termes de la requête et ceux du document

Sur le Web :

- Correspondance classique
- ET prise en compte de la structure hypertexte

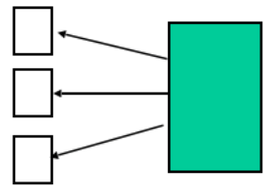
Deux algorithmes décrits :

- HITS
- PageRank

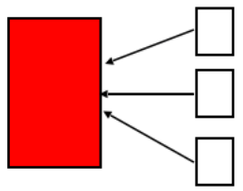
HITS - Kleinberg 98

Catégories de pages :

- Les bonnes sources de contenu



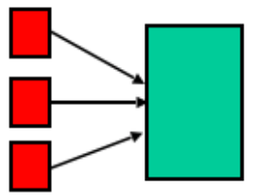
- Les bonnes sources de liens



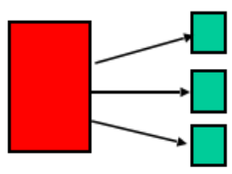
- Les autres

Intuitions :

- L’autorité vient des liens entrants
- Le fait d’être un bon hub vient des liens sortants
- Une meilleure autorité vient de liens entrants de bons hubs



- Un meilleur hub vient des liens entrants de bonnes autorités



HITS - Kleinberg 98

1. Initialisation (avec N pages)

$$\sum_i AUTH[i]^2 = \sum_i HUB[i]^2 = 1 \quad \Rightarrow \text{pour une page V} : AUTH[V] = HUB[V] = \frac{1}{\sqrt{N}}$$

2. Calcul itératif jusqu'à convergence

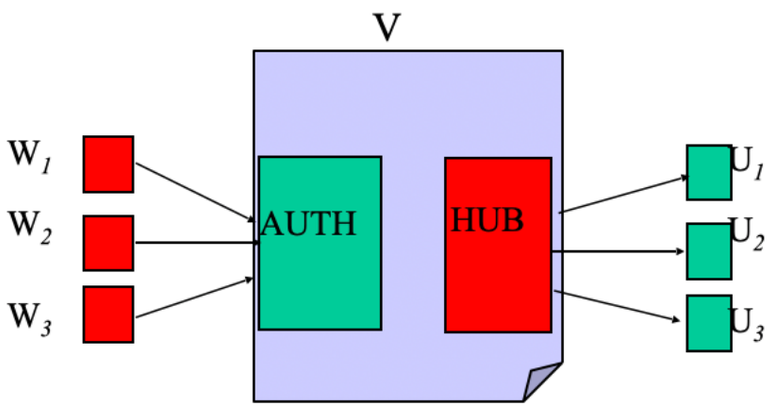
   i. Pour chaque page V

$$AUTH_{NN}[V] = \sum_{Lien(W_i,V)} HUB[W_i] \qquad HUB_{NN}[V] = \sum_{Lien(V,U_i)} AUTH[U_i]$$

   ii. Normalisation

$$AUTH[V] = \frac{AUTH_{NN}[V]}{\sqrt{\sum_i AUTH_{NN}[i]^2}}$$

$$HUB[V] = \frac{HUB_{NN}[V]}{\sqrt{\sum_i HUB_{NN}[i]^2}}$$



HITS - Exemples

1. Initialisation (étape 0)

$$AUTH[A] = HUB[B] = \frac{1}{\sqrt{2}} = 0.71$$

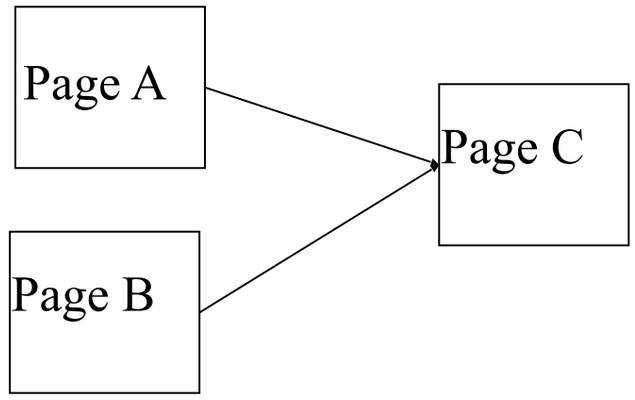
2. Calcul itératif jusqu'à convergence

	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	$\sqrt{\sum AUTH[U]^2}$	$\sqrt{\sum HUB[U]^2}$
0	0.71	0.71	0.71	0.71		
1.1 (NN)	0	0.71	0.71	0	0.71	0.71
	0	1	1	0		
1.2	0	1	1	0	1	1
2.1 (NN)	0	1	1	0		
2.2						

HITS - Exemples

1. Initialisation (étape 0)

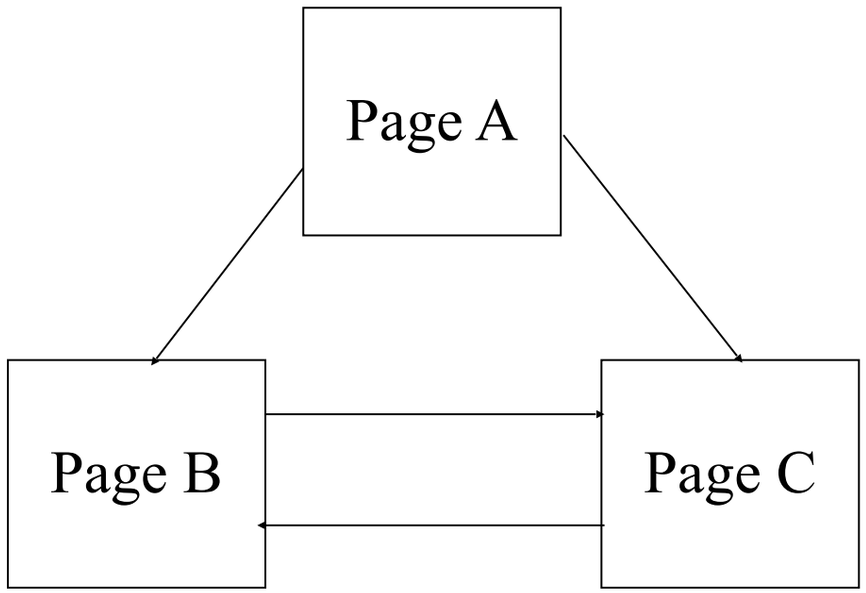
$$\text{AUTH}[A, B, C] = \text{HUB}[A, B, C] = \frac{1}{\sqrt{3}} = 0.58$$



2. Calcul itératif jusqu'à convergence

	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	AUTH[C]	HUB[C]	$\sqrt{\sum A[U]^2}$	$\sqrt{\sum H[U]^2}$
0	0.58	0.58	0.58	0.58	0.58	0.58		
1.1	0	0.58	0	0.58	1.15	0	1.15	0.82
1.2	0	0.71	0	0.71	1	0		
2.1	0	1	0	1	1.41	0	1.41	1.41
2.2	0	0.71	0	0.71	1	0		

HITS - Exemples



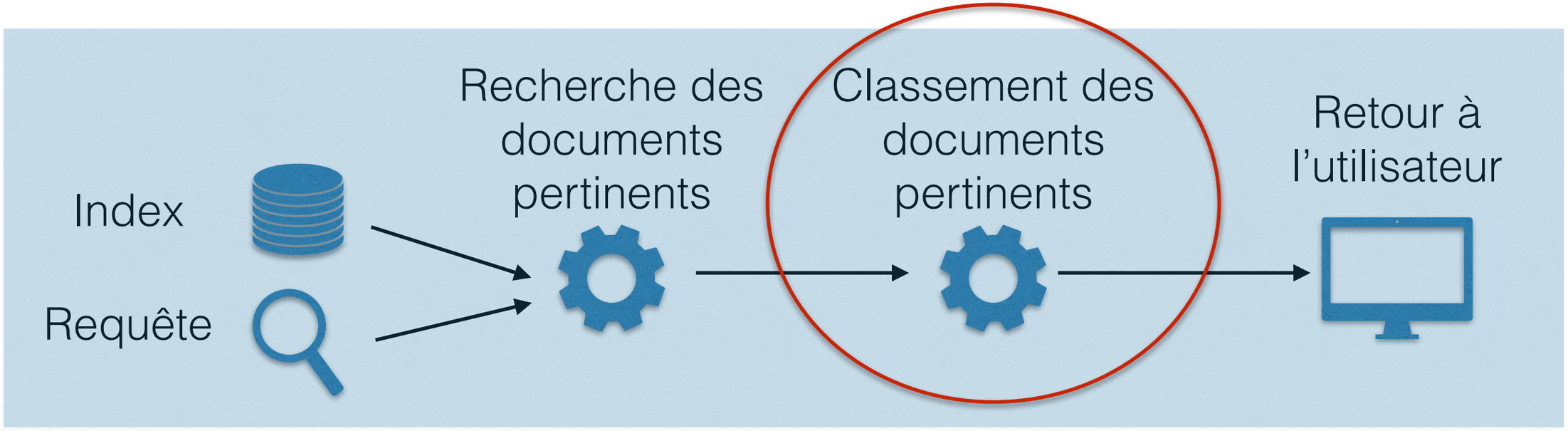
	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	AUTH[C]	HUB[C]	$\sqrt{\sum A[U]^2}$	$\sqrt{\sum H[U]^2}$
0	0.58	0.58	0.58	0.58	0.58	0.58		
1.1	0	0.58	0	0.58	1.15	0	1.15	0.82
1.2	0	0.71	0	0.71	1	0		
2.1	0	1	0	1	1.41	0	1.41	1.41
2.2	0	0.71	0	0.71	1	0		

HITS - Conclusion

1. Nécessite un calcul pour chaque requête
2. Calcul long (convergence après 20 boucles pour 200 pages Web)
3. Facile à utiliser à mauvais escient pour augmenter artificiellement les valeurs d'autorité et de hub en générant des liens automatiquement

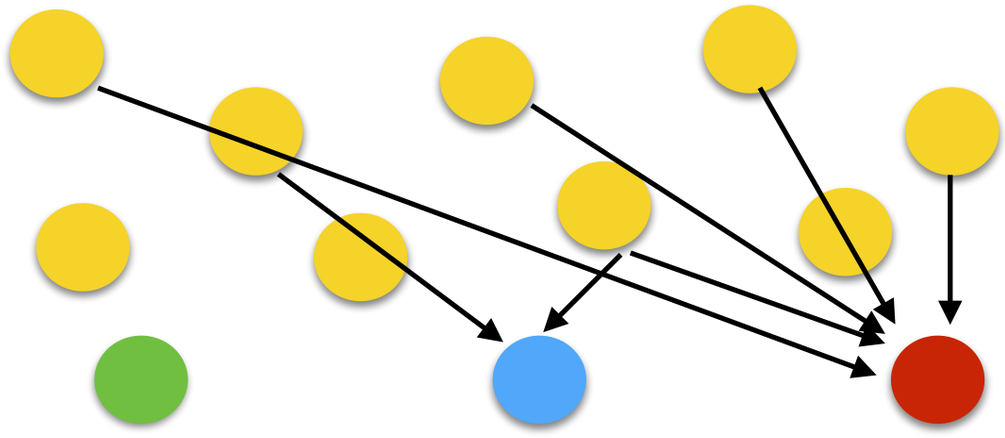
PageRank

- Système de classement des pages Web utilisé par Google



- Méthode d'évaluation de la "popularité" d'une page

- Utilisée parmi d'autres critères pour classer les documents résultats



PageRank

- Mesure de la popularité d'une page :
  - Lien vers la page : +1
  - Absence de lien : 0
- Calcul du PageRank d'une page A:

$$PR(A) = (1-d) + d * (\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

  - T_i : Les pages du web qui pointent sur A
  - $C(T_i)$: nombre de liens sortant de la page T_i
  - d : facteur d'équilibrage dans [0, 1]

PageRank

importance minimale d'une page (somme PR de toutes les pages vaut 1)

chaque page a une notion de sa propre importance

$$PR(A) = (1-d) + d * (\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

ajout équilibré des "votes" de toutes les pages (couramment d=0.85)

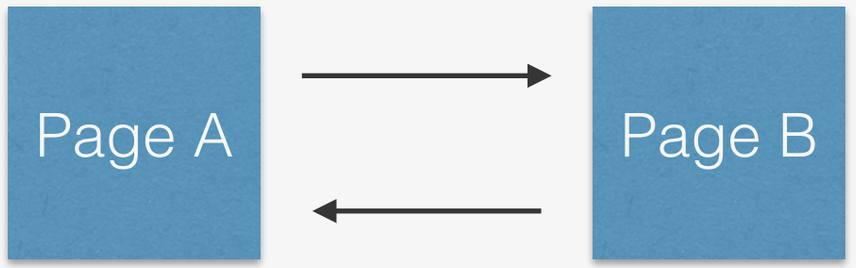
équilibre les votes de manière égale à tous les liens sortants

PageRank

- Formule basée sur une boucle
  - Pour calculer PR(A) il faut connaître PR(T1), mais si A pointe sur T1 il faut connaître PR(A) (!!)
  - On itère avec des valeurs initiales fixées

d=0.85, PR(A) = 1 valeur initiale

PR(A) = (1-d) = 0.15

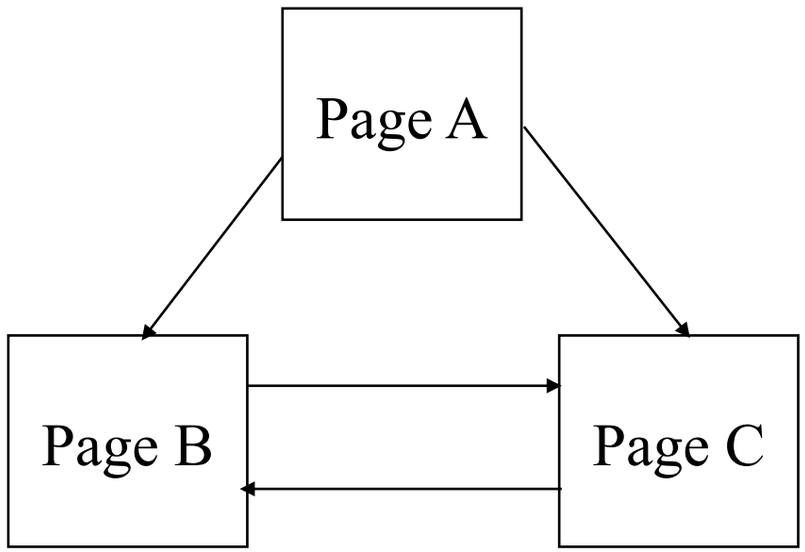


d=0.85, PR(A) = 1 PR(B) = 1

PR(A) = (1-d) + d*(PR(B)/1) = 0.15 + 0.85/1 = 1

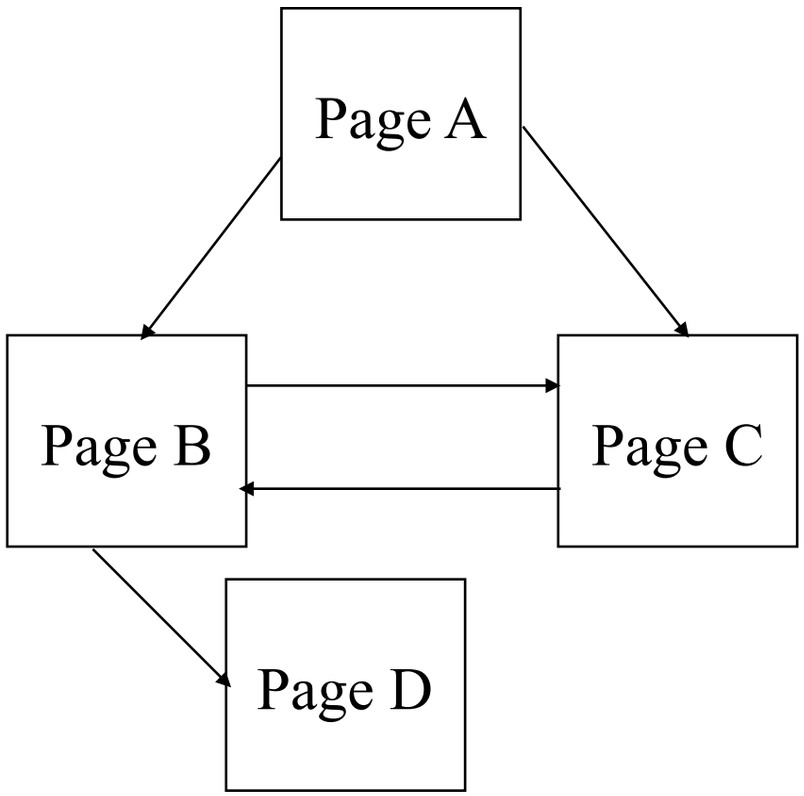
PR(B) = (1-d) + d*(PR(B)/1) = 1

- Google – Calculs de Pagerank
  - Avec d=0.85 et initialement PR(A) = 1, PR(B) = 1 et PR(C)=1



	PR(A)	PR(B)	PR(C)
1	1	1	1
2	0,15	1,425	1,425
3	0,15	1,425	1,425

- Google – Calculs de Pagerank



	PR(A)	PR(B)	PR(C)	PR(D)
1	1	1	1	1
2	0.15	1.425	1	0.575
3	0,15	1,06	0,82	0,76
4	0,15	0,91	0,67	0,60
5	0,15	0,78	0,60	0,54
6	0,15	0,72	0,55	0,48
7	0,15	0,68	0,52	0,46
8	0,15	0,66	0,50	0,44
9				

Note : on s'arrête quand la moyenne des différences est inférieure au seuil 0.02.

37

PageRank - conclusion

- + Principe très intéressant et qui a prouvé son utilité
- + Difficile à spammer
- - Complexité de calcul sur les milliards de pages du web
- Défavorise la nouveauté
- Pas de liens typés : quels sens donner aux liens??