

Web mining

Shengrui Wang

22 Novembre 2024



Plan

Introduction

Forage de structure web

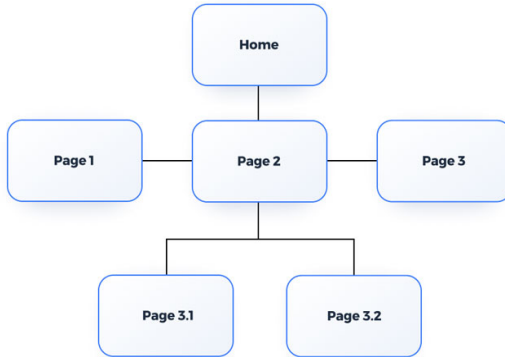
Forage de contenu web

Introduction

- Le Web pourrait être vu comme un repertoire de données,
- Il est caractérisé par :
 - ▶ **son grand volume de données** : en perpétuelle augmentation, la quantité de données disponibles est immense.
 - ▶ **l'hétérogénéité des données** : Les données présentes sont de types et natures diverses.
 - ▶ **l'interconnexion du contenu** : Les liens d'une page vers d'autres pages créent une toile facilitant la navigation et la découverte des contenus.
 - ▶ **le bruit** : Vu comme un média libre, toute personne est donc capable d'y poster une information qui pourrait être importante ou pas.
 - ▶ **sa vélocité** : En perpétuelle croissance dû à l'ajout perpétuel de contenus par les utilisateurs.
 - ▶ **sa sociabilité** : Les utilisateurs peuvent collaborer entre eux.

Introduction

- Bien qu'il soit vu comme un repertoire *fourre-tout*, où l'on peut incorporer tout type d'information, il n'en demeure pas moins que les liens entre les différentes pages de contenus constituent une structure topologique bien définie.



Introduction

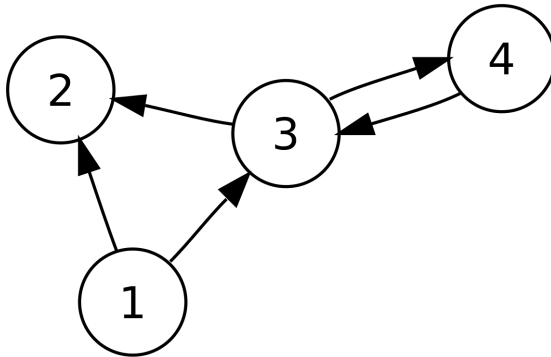
- Ainsi présenté, le Web n'est plus seulement vu comme un repertoire *fourre-tout*, mais aussi une *toile* organisée où l'on pourrait naviguer.
- La fouille de données a pour objectif l'extraction de connaissances à partir
 - ▶ **de la structure formée par les liens entre pages/objets/communautés** : Forage de structure. Ce type de fouille vise essentiellement à analyser la structure du Web.
 - ▶ **du contenu des pages Web** : Forage de contenu, Ce type de fouille vise essentiellement à analyser le contenu des pages.
 - ▶ **des données d'usages** (typiquement des logs de serveurs web) : forage de données d'usage.

Forage de structure web

- Les premiers moteurs de recherche retrouvaient les documents pertinents **seulement en fonction de la similarité** du contenu avec la requête de l'utilisateur.
- l'augmentation rapide du nombre de pages Web a rendu **très difficile l'ordonnement** des documents retrouvés.
- il est **très facile de corrompre** le contenu d'une page de telle sorte à la rendre similaire à des très nombreuses requêtes.
- Pour contourner ce problème, il s'est donc posé la question d'**exploiter les liens entre les pages**. En d'autres termes, exploiter la structure topologique du web.

- La structure topologique du Web peut-être représentée par une structure de graphe
- où les noeuds représenteraient les différentes pages, tandis que les liens seraient les hyperliens.
- De cette manière, on pourrait donc exploiter les propriétés statistiques du graphe et ainsi déduire de la pertinence de certaines pages Webs.

Forage de structure web



Forage de structure web

- Formellement, la structure topologique du Web ou d'un site Web pourrait être donnée par le graphe $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, où
- \mathbf{V} serait l'ensemble des noeuds représentant les différentes pages webs et
- $\mathbf{E} = \{(u, v) / u \in \mathbf{V}, v \in \mathbf{V}\}$ l'ensemble des liens representant les hyperliens entre les pages webs.
- Lorsque le graphe est non-dirigé on a, $(u, v) = (v, u)$ et lorsqu'il est dirigé on pourrait avoir $(u, v) \neq (v, u)$.

Forage de structure web – Importance d'une page web

- En fonction de sa position dans la structure topologique, on pourrait évaluer sa pertinence en utilisant les mesures de centralités
- Degré d'un noeud:

$$D(u) = |\{(u, v) \in \mathbf{E} / v \in \mathbf{V}\}|$$

- Prestige d'un noeud (fonctionnel uniquement dans les graphes dirigés), Il représente encore le degré entrant.

$$P(u) = |\{(v, u) \in \mathbf{E} / v \in \mathbf{V}\}|$$

- L'intermédierité:

$$I(u) = \sum_{v \in \mathbf{V} \setminus \{u, w\}} \sum_{w \in \mathbf{V} \setminus \{u, v\}} SP_{v,w}(u)$$

avec $SP_{v,w}(u)$ le nombre de plus court chemins entre v et w en passant par u



PageRank

Papier :

Lawrence Page and Sergey Brin

“The PageRank citation ranking: Bringing order to the Web”

<http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/7144/http:zSzzSzwww-db.stanford.eduzSz~backrubzSzpageranksub.pdf/page98pagerank.pdf>



PageRank

■ Définition

➤ Le PageRank est une valeur numérique (une note qui varie sur la barre d'outil Google de 0 à 10) que Google attribut a une page web pour représenter (refléter) son importance par rapport à d'autres pages du même site ou par rapport à n'importe quelle autre page sur le web.

➤ Le PageRank sera bénéfique à une page Web pour le positionnement sur google, ce n'est pas le seul facteur mais il a une grande importance. Google calcul cette note (le PR) en analysant principalement le nombre et la qualité des liens entrant et des liens sortant pour chaque page sur le web.

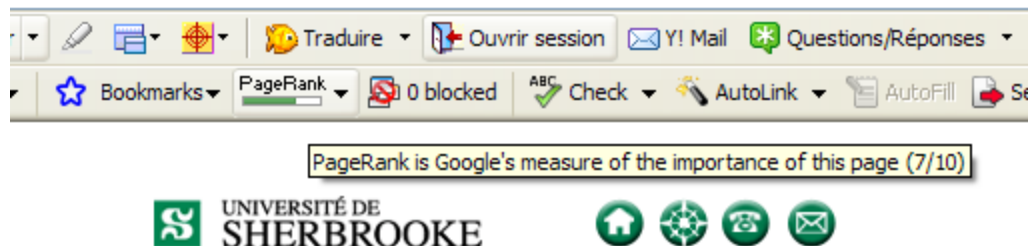
Chaque lien placé sur une page **A** vers une page **B** est interprété par Google comme un VOTE de **A** à **B**.

le plus de liens entrant une page, plus haute sa note sera (son PageRank).
Les liens sortant la feront baisser.

PageRank

■ Remarque

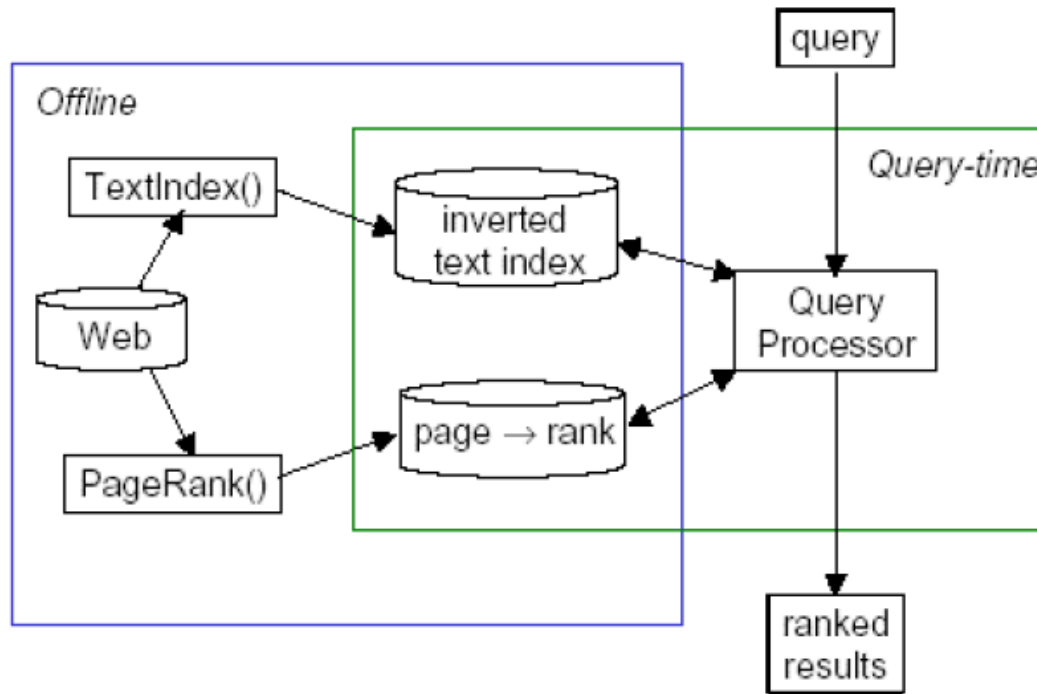
- Un vote émis par la page d'accueil d'un site majeur tel que Microsoft pèse beaucoup plus lourd qu'un vote émis par la page perso de votre site.
- Le PR est une mesure pour déterminer l'importance d'une page et non d'un site en entier
- Le PR d'une page peut être visualisé par les utilisateurs de « Google toolbar »



■ En résumé ...

Le PageRank d'une page Web, c'est-à-dire son poids dans l'index de Google, dépend donc non seulement de la qualité du contenu d'un site, mais aussi, par extension, de celle des sites qui le lient.

Cycle de vie d'une requête



- Les calculs les plus complexes et les plus longs se font offline.
- Le moteur n'a plus que des calculs très simples à effectuer pour construire les pages de résultats

■ Calcule du PageRank

Une page A reçoit des liens émis par les page $T1, \dots, Tn$

$$PR(A) = (1 - d) + d * \left[\frac{PR(T1)}{C(T1)} + \dots + \frac{PR(Tn)}{C(Tn)} \right]$$

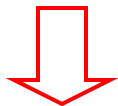
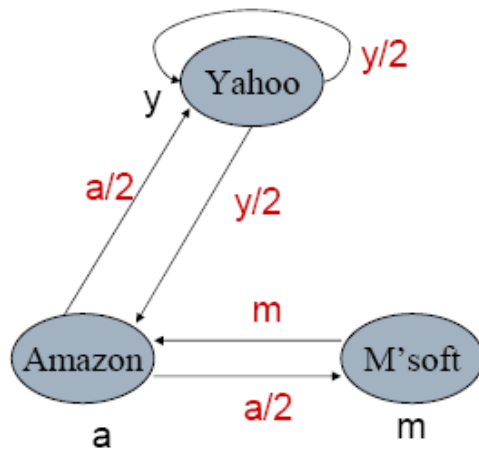
d : un facteur d'amortissement « damping factor » pouvant être ajusté entre 0 et 1.
Généralement d est entre 0.8 et 0.9, la valeur suggérée $d = 0.85$

$C(Ti)$: le nombre de liens émis par la page Ti (liens sortant) « Out-links »

- Le PageRank peut être calculé en utilisant un simple algorithme itératif
- On remarque que le PageRank d'une page ne peut être inférieur à 0.15 même s'il n'y a aucune page qui mène à elle. Donc une page aurait toujours un minimum de 0.15 à partager.

Intuition derrière la formule de PageRank

■ Estimation d'une matrice des liens Web



	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

➤ Les lignes et les colonnes de la matrice sont les page Web.
Soit X_{ij} est un élément de la matrice

➤ Construction de la matrice

Si la page j a n liens sortants **alors**

$X_{ij} = 1/n$ s'il y a un lien sortant de la page i vers j

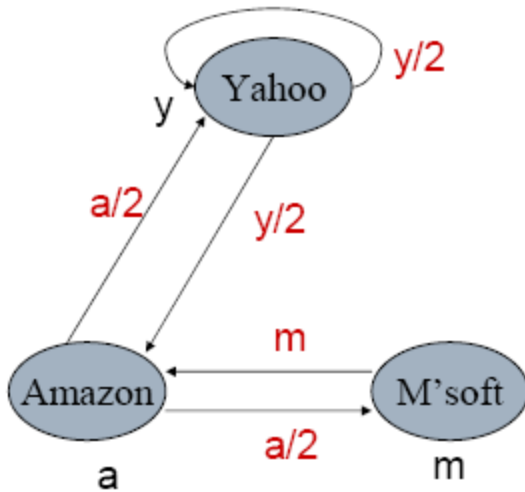
Sinon $X_{ij} = 0$

➤ les liens entre les page Web sont présentés dans une matrice stochastique (La somme des éléments de chaque colonne = 1)

➤ On remarque que Yahoo! divise son importance avec lui-même et Amazon. Microsoft donne toute son importance à Amazon

Intuition derrière la formule de PageRank

■ Estimation d'une matrice des liens Web



➤ Chaque vote est proportionnel à l'importance de sa page source.

➤ Soit P une page Web avec n liens sortant et une valeur d'importance $= x$

Chaque lien sortant de P aura un poids $= x/n$

➤ Résultat : 3 équations à 3 inconnus (en supposant $d=1$)

$$y = y/2 + a/2$$

$$a = y/2 + m$$

$$m = a/2$$

Estimation de l'importance des pages Web (en supposant $d=1$)

❑ Soit $[y, a, m]$ un vecteur qui représente l'importance de chaque page: Yahoo!, Amazon, Microsoft dans cet ordre.

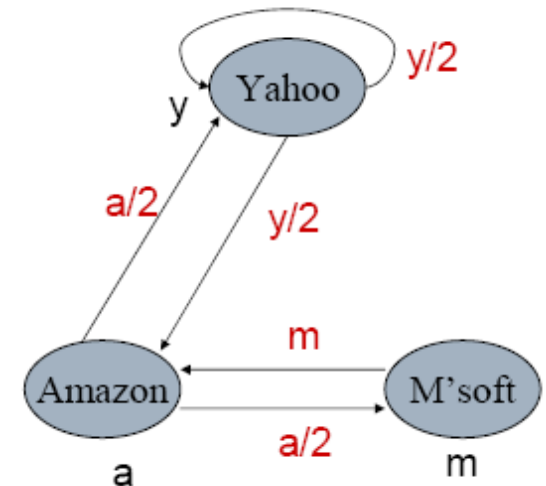
➤ L'équation qui décrit ces trois variables est définie par:

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

$$\begin{aligned} y &= y/2 + a/2 \\ a &= y/2 + m \\ m &= a/2 \end{aligned}$$

Solution de cette équation

y	=	1/3	1/3	5/12	3/8	2/5
a		1/3	1/2	1/3	11/24	2/5
m		1/3	1/6	1/4	1/6	1/5



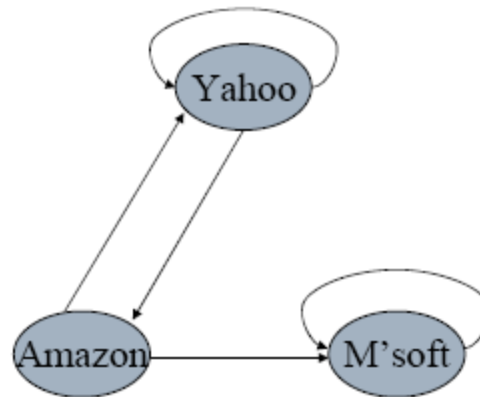
➤ Yahoo! et Amazon ont la même valeur d'importance, et le double de la valeur de l'importance de Microsoft

Problème ... Spider trap

Spider traps : un groupe d'une ou plusieurs pages qui n'ont aucun lien sortant vont accumuler toute les valeurs d'importance des pages web.

***Spider traps:** a group of one or more pages that have no links out of the group will eventually accumulate all the importance of the Web.*

■ Exemple



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	1

y	=	1	1	3/4	5/8	...	0
a		1	1/2	1/2	3/8		0
m		1	3/2	7/4	2		3

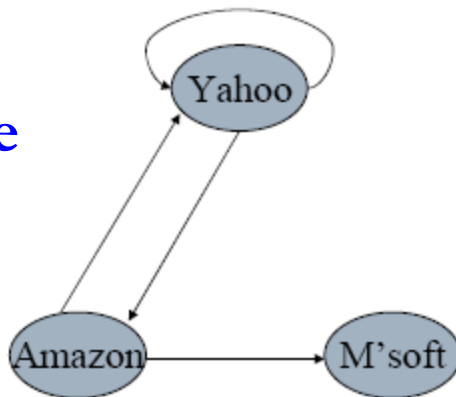
➤ Microsoft devient « a spider trap »

Problème ... Dead end

Dead ends : une page qui n'a aucun lien sortant.

***Dead ends:** a page that has no successors has nowhere to send its importance. Eventually, all importance will “leak out of” the Web.*

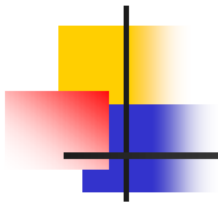
■ Exemple



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	0

	y	a	m				
y	1	1	3/4	5/8	1/2		0
a	1	1/2	1/2	3/8	5/16	...	0
m	1	1/2	1/4	1/4	3/16		0

➤ Microsoft devient « a dead end »



Solution de Google ... PageRank

■ L'idée

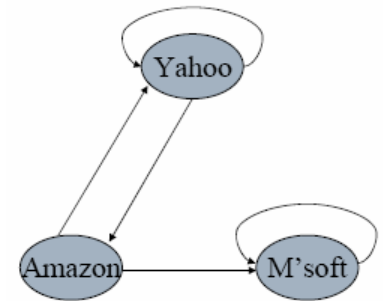
“Instead of applying the matrix directly, “tax” each page some fraction of its current importance, and distribute the taxed importance equally among all pages”

$$PR(A) = (1-d) + d * \left[\frac{PR(T1)}{C(T1)} + \dots + \frac{PR(Tn)}{C(Tn)} \right]$$

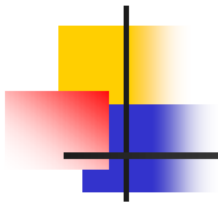
■ **Exemple** : tax = d = 20%

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = 0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$$

$$y = 7/11; \quad a = 5/11; \quad m = 21/11$$



Cette distribution de l'importance des pages Web est plus raisonnable que La première fois sans le d : (y = 0; a = 0; m = 3)



Exemple de calcul de PageRank

$$PR(A) = (1 - d) + d * \left[\frac{PR(T1)}{C(T1)} + \dots + \frac{PR(Tn)}{C(Tn)} \right]$$

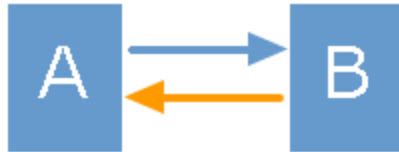
- Le PR d'une page A dépend du PR des pages T1, ..., Tn qui émettent un lien vers A et ne peut donc pas être déterminé sans connaître le PR de ces dernières et de toutes celles qui émettent un lien vers elles et ainsi de suite ...

- Selon Lawrence Page and Sergey Brin :
 - Le PR peut être calculé en utilisant un simple algorithme itératif.
 - Le PR correspond au vecteur propre principal de la matrice normalisée des liens du Web.

Exemple de calcul de PageRank

■ **Exemple 1:** 2 pages A et B pointant l'une vers l'autre

Chaque page a un lien sortant donc $C(A) = C(B) = 1$



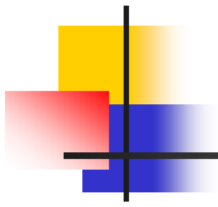
➤ Nous ne connaissons pas le PR des deux page, donc il ne faut une valeur de départ: 1 par exemple

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

Soit, avec un facteur d'amortissement de 0.85 :

PR(A)	=	0.15 + 0.85 * 1	=	1
PR(B)	=	0.15 + 0.85 * 1	=	1

$$PR(A) = PR(B) = 1$$



Exemple de calcul de PageRank

- Prenons une autre valeur de départ, cette fois $= 0$

Première itération

PR(A)	$= 0.15 + 0.85 * 0$	$= 0.15$
PR(B)	$= 0.15 + 0.85 * 0.15$	$= 0.2775$

Deuxième itération

PR(A)	$= 0.15 + 0.85 * 0.2775$	$= 0.385875$
PR(B)	$= 0.15 + 0.85 * 0.385875$	$= 0.47799375$

Troisième itération

PR(A)	$= 0.15 + 0.85 * 0.47799375$	$= 0.5562946875$
PR(B)	$= 0.15 + 0.85 * 0.5562946875$	$= 0.622850484375$

- On remarque bien que les valeurs de PR augmente à chaque itération
- Les valeurs vont continuer à augmenter jusqu'à ils seront proche ou égale à 1



Exemple de calcul de PageRank

- Prenons une autre valeur de départ, cette fois = 2

PR(A)	= 0.15 + 0.85 * 2	= 1.85
PR(B)	= 0.15 + 0.85 * 1.85	= 1.7225

PR(A)	= 0.15 + 0.85 * 1.7225	= 1.614125
PR(B)	= 0.15 + 0.85 * 1.614125	= 1.52200625

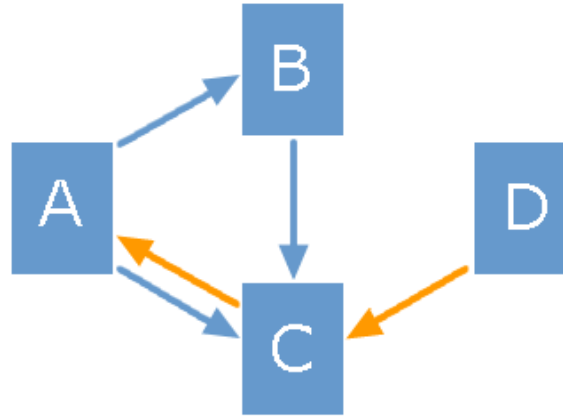
PR(A)	= 0.15 + 0.85 * 1.52200625	= 1.4437053125
PR(B)	= 0.15 + 0.85 * 1.4437053125	= 1.377149515625



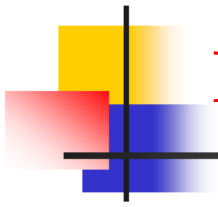
Les valeurs continuent à converger vers 1

Exemple de calcul de PageRank

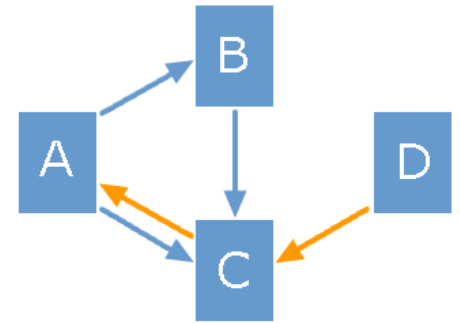
■ Exemple 2: Quatre pages liées



- Dans cet exemple nous avons une connexion entre 4 pages, dont une ne recevant aucun lien (la page D).
- Le PR de cette page sera donc de 0.15, grâce au premier terme de la formule du PageRank $(1-d)$.
- Bien qu'ayant un PR calculé, il est vraisemblable que cette page disparaîtra de l'index de Google très vite, n'ayant aucun lien entrant



Exemple de calcul de PageRank



➤ Au bout d'une vingtaine d'itérations, les valeurs de PR convergent vers les valeurs suivantes:

Page A	1.49
Page B	0.78
Page C	1.58
Page D	0.15
Somme des PageRank	4.0
Moyenne	1.0

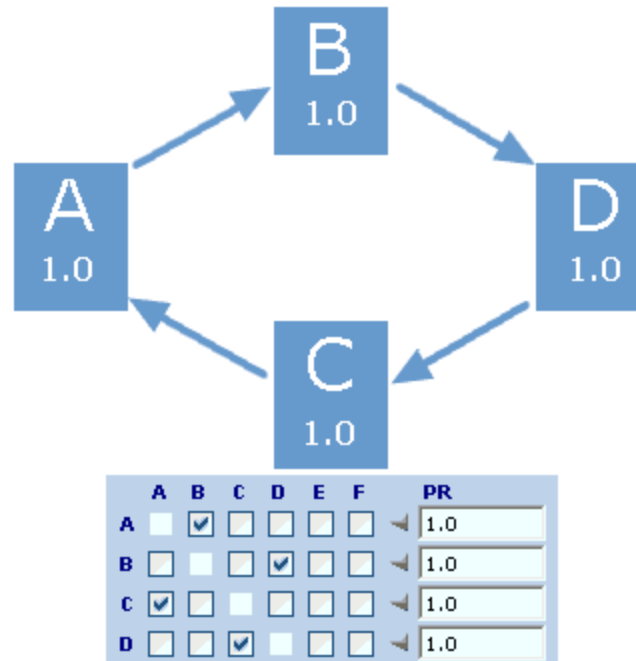
Nous voyons que dans notre exemple, la page C a le PR le plus élevé. C'était prévisible dès l'examen graphique, comme elle reçoit un lien entrant des pages A, B et D et n'en émet qu'un seul vers la page A.

✓ Pour les exemples qui suivent le calculateur du PageRank disponible dans le lien suivant sera utilisé

http://www.webworkshop.net/pagerank_calculator.php3

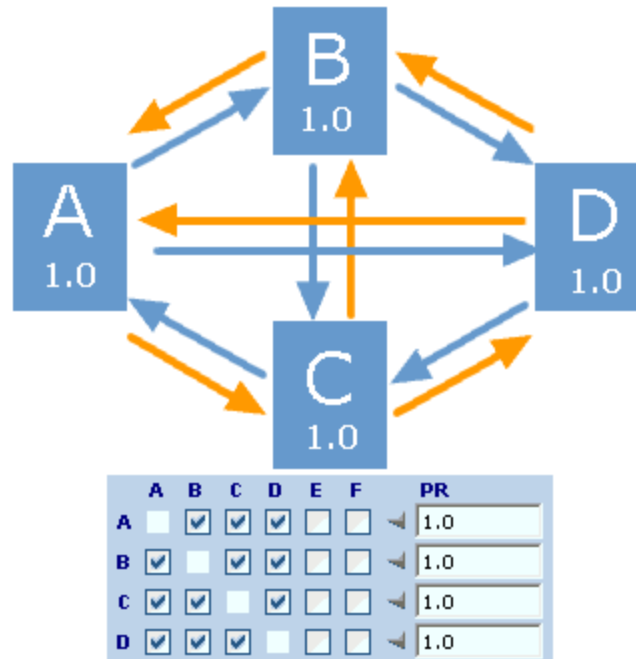
Exemple de calcul de PageRank

■ Exemple 3 : Liens circulaires



- Chaque page ayant exactement un lien entrant et un lien sortant.
- Les liens circulaires ne favorisent aucune page.
- Le PageRank de chaque page s'établira donc à 1.0

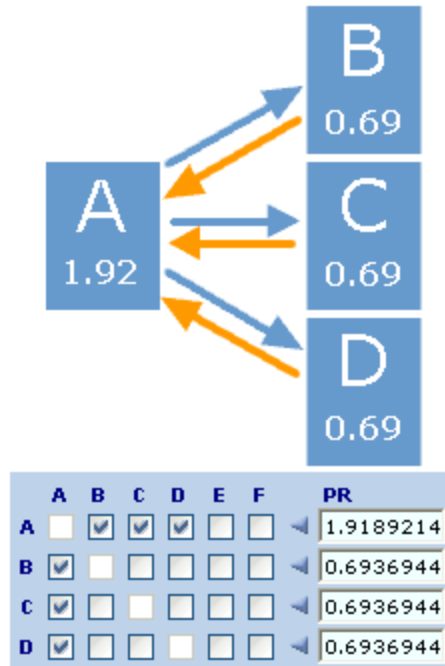
Exemple de calcul de PageRank



- Toutes les pages sont liées entre-elles.
- Aucune page n'est prépondérante.

Exemple de calcul de PageRank

■ Exemple 4 : Structure hiérarchique simple

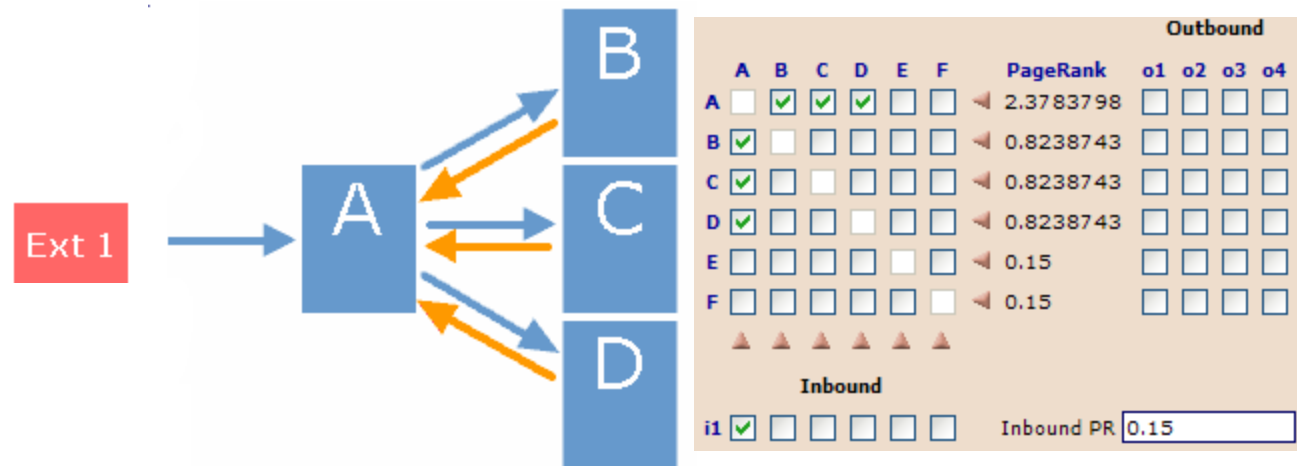


➤ Le PageRank de A est élevé car c'est elle qui reçoit le plus de vote.

Exemple de calcul de PageRank

■ Exemple 5: Effet des liens entrant

■ Exemple 5-1 : Structure hiérarchique avec lien entrant



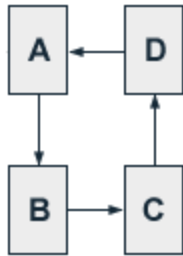
Un webmaster d'un site extérieur à émis un lien de sa page vers la page A.

➔ ceci va augmenter automatiquement le PageRank de A

Exemple de calcul de PageRank

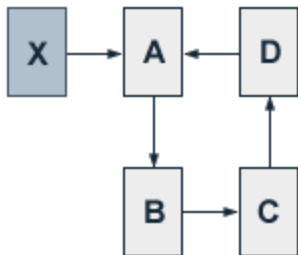
■ Exemple 5: Effet des liens entrant

■ Exemple 5-2 : Le cas d'une structure circulaire



					Outbound				
	A	B	C	D	PageRank	o1	o2	o3	o4
A	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1.0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
B	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	1.0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
C	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	1.0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
D	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1.0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Inbound									
i1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Inbound PR	<input type="text" value="0.15"/>			

$$PR(A) = PR(B) = PR(C) = PR(D) = 1$$



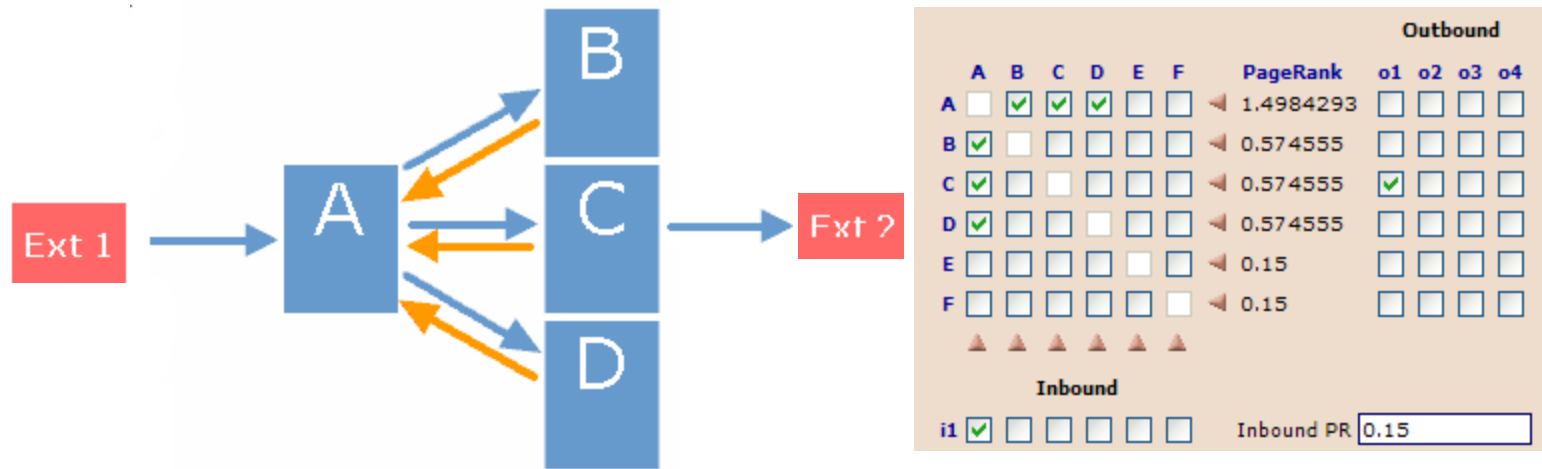
					Outbound				
	A	B	C	D	PageRank	o1	o2	o3	o4
A	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1.2667399	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
B	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	1.2267289	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
C	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	1.1927196	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
D	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1.1638116	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Inbound									
i1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Inbound PR	<input type="text" value="0.15"/>			

$$PR(A) > PR(B) > PR(C) > PR(D)$$

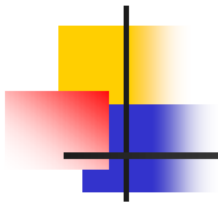
Exemple de calcul de PageRank

■ Exemple 6: Effet des liens sortant

■ Exemple 6-1 : Structure hiérarchique avec lien entrant et lien sortant

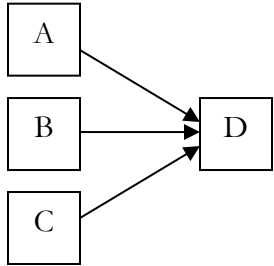


- Le simple fait d'émettre un lien vers une page extérieure au départ de la page C fait chuter le PR de manière conséquente sur toutes les pages du site.
- La seule cause de la chute de PR vient du fait que la page C qui redistribuait dans l'exemple 5-1 tout son PR à la page A, ne lui renvoie plus que la moitié de celui-ci.

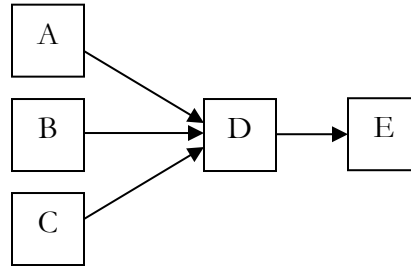


Exemple de calcul de PageRank

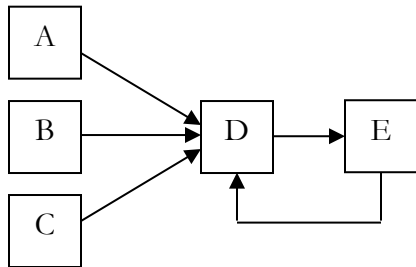
■ Exemple 6-2



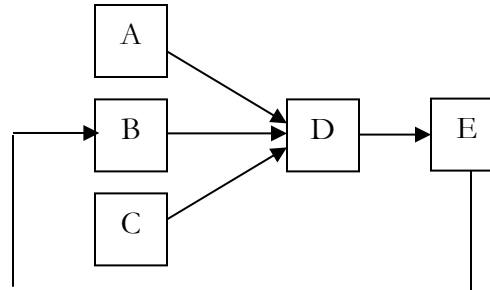
	A	B	C	D	PageRank
A	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0.15
B	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0.15
C	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0.15
D	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.5325



	A	B	C	D	E	PageRank
A	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
B	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
C	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
D	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	0.5325
E	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0.602625



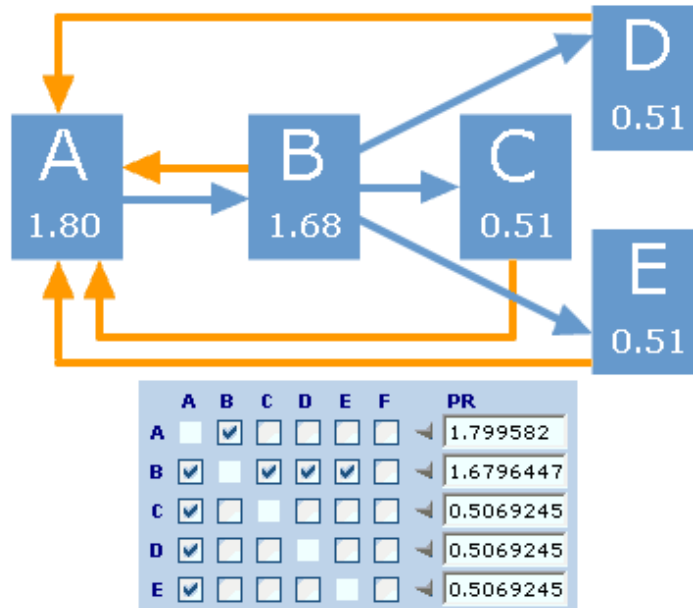
	A	B	C	D	E	PageRank
A	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
B	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
C	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
D	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	2.3783753
E	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	2.171619



	A	B	C	D	E	PageRank
A	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
B	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	1.4774538
C	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	0.15
D	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	1.6608358
E	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1.5617104

Exemple de calcul de PageRank

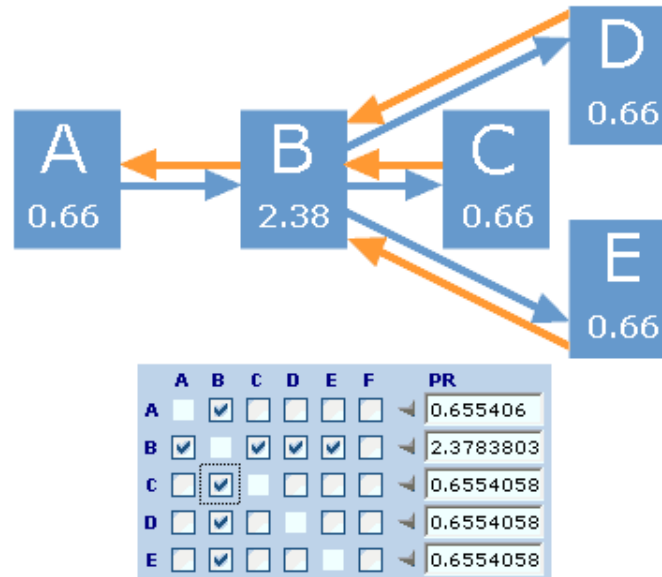
■ Exemple 7 : Un plan de site



➤ Dans cet exemple, la page d'accueil (A) émet un lien vers le plan du site (B). Celui-ci, en plus du lien retour vers la page d'accueil émet un lien vers chacune des pages du site (C, D et E). Pour éviter les "fuites" de PageRank, celles-ci émettent un lien en retour vers la page d'accueil.

Exemple de calcul de PageRank

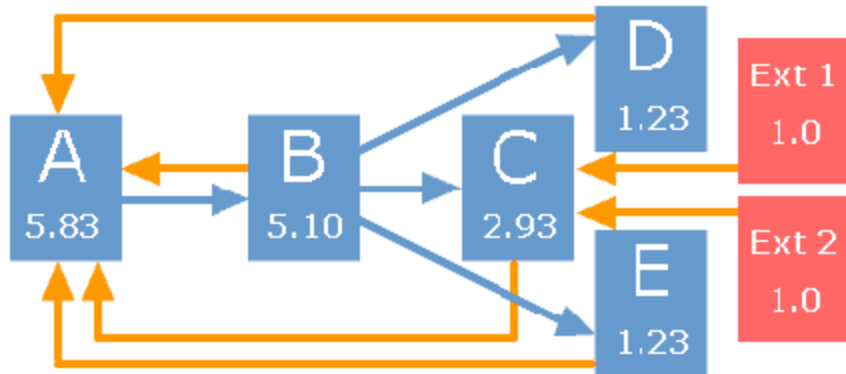
■ Exemple 8 : Un plan de site lié différemment



- ❑ La seule différence avec l'exemple précédent vient du lien en retour des pages internes du site. Plutôt que d'émettre un lien vers la page d'accueil, elles émettent le lien en retour vers le plan du site (B).
- ❑ Ceci a pour effet de favoriser le PageRank de la page B, ce qui peut être souhaité si elle est riche en mots clés.
- ❑ Par contre, la page d'accueil A voit son PageRank diminuer fortement comme elle n'a plus qu'un lien entrant.

Exemple de calcul de PageRank

■ Exemple 9 : Un plan de site avec des liens entrants sur une page interne









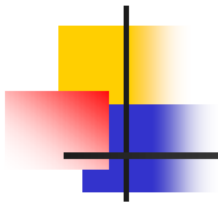
➤ Le contenu intéressant de la page C lui vaut des liens spontanés de la part de deux sites extérieurs.

➤ On suppose le PR des pages entrantes = 1
La page C voit tout naturellement son PR augmenté.

➤ Grâce au chainage interne la page d'accueil et le plan du site font toutes deux un bond vers le haut.

➤ Dans une moindre mesure les pages D et E profitent du gain en PR du plan du site B.

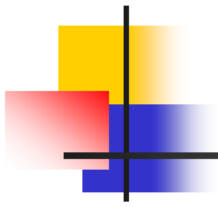
	A	B	C	D	E	F	PR	s1	s2	s3	s4
A	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	5.8271056	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
B	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	5.1030397	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
C	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2.9343959	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
D	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1.2343959	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
E	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	1.2343959	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
F	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
											
Liens entrants											
e1	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	PR	<input type="text" value="1"/>			
e2	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	PR	<input type="text" value="1"/>			



PageRank - Discussion

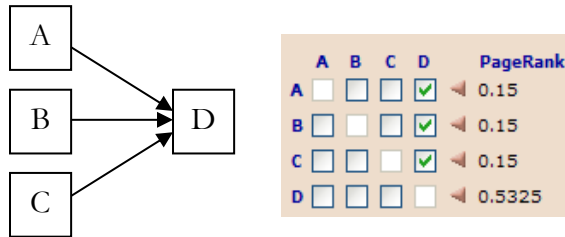
❑ Améliorer la pertinence des résultats constitue un objectif technologique majeur pour la plupart des moteurs de recherche. Le problème, c'est que cette « pertinence » est une notion subjective, qu'il est donc difficile de manier avec des algorithmes purement mathématiques.

❑ La valeur du PageRank part du principe intuitif que l'importance d'une page dépend du nombre de liens entrants pointant vers cette page, mais aussi de l'importance des pages d'où partent ces liens.

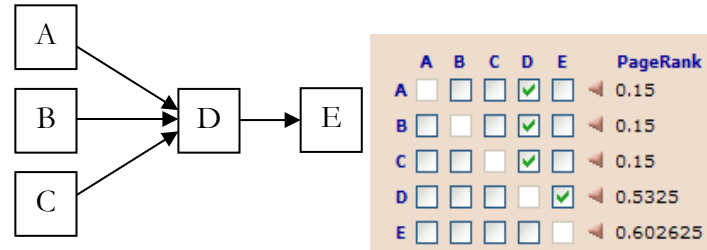


PageRank - Discussion

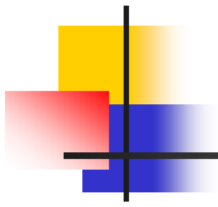
➤ Problème : propagation du PageRank



Plusieurs pages pointent vers la page D
La page D a un bon contenu
Le PR de D va augmenter



Il est toujours possible de voir un lien sortant d'une page avec un bon contenu vers une page non importante (ex. D → E). Conséquence, cette page (E) va avoir son PR augmenté alors qu'elle ne contient aucune information pertinente.

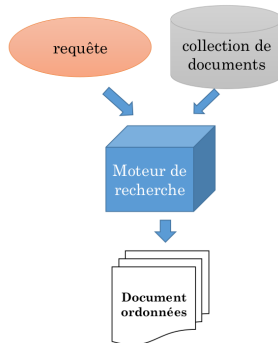


PageRank - Discussion

- ❑ Le calcul du PageRank pour un index de la taille de Google nécessite du temps et une puissance de calcul phénoménale.
- Il est possible que ce calcul est géré avec une architecture basée sur un grand nombre de machines calculant chacune les valeurs pour une petite zone de la colossale matrice du Web.
- L'algorithme itératif de calcul du PageRank converge au bout d'un certain nombre d'itérations vers une valeur fixe.
 - ❑ La convergence de l'algorithme n'est pas prouvée théoriquement. Cependant, les expérimentations montrent que la convergence est assurée en pratique.

Forage de contenu web

- Ici le contenu de la page web est prise en considération
- Un classement des pages web est effectué en fonction de leurs scores de pertinence par rapport à la requête.



- Étant donné D l'ensemble des contenus des pages webs.
- La sélection des documents peut se faire suivant une représentation booléenne:
 - ▶ Ici la requête de l'utilisateur est prise comme une suite logique.
 - ▶ Exemple: $((x \text{AND} y) \text{AND} (\text{NOT} z))$ signifie les documents contenant les termes x et y et non z
 - ▶ le système récupère chaque document qui rend la requête logiquement vraie
 - ▶ Ici l'appariement est exact, on ne tient pas compte de la pertinence d'un mot dans le document recherché.

- La sélection des documents peut se faire suivant une représentation vectorielle.
 - ▶ Étant donnée la famille de termes $\mathbf{W} = \{W_1, W_2, \dots, W_n\}$ que l'on pourrait rencontrer sur une page web P ,
 - ▶ On pourrait représenter la page web P par le vecteur $P = (\omega_1, \omega_2, \dots, \omega_n)$ où ω_i , $1 \leq i \leq n$, est le nombre d'occurrence du mot W_i dans le contenu de la page P .
 - ▶ La requête de l'utilisateur est aussi représenté sous le même format
 - ▶ En fonction d'une mesure de similarité on peut donc faire le classement des pages en fonction de celles qui sont les plus similaires de la requête.



HITS

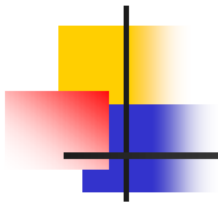
Papier :

J. M. Kleinberg

“Authoritative Source in a Hyperlinked Environment”

Journal of ACM, vol. 46, no. 5, pp. 604-632, 1999.

<http://www.cs.cornell.edu/home/kleinber/auth.pdf>



HITS : Hyperlinked Induced Topic Search

- ❑ HITS est, avec le PageRank, l'un des plus célèbres algorithmes de classement des pages Web.
- ❑ HITS est resté plus longtemps dans les laboratoires que le Pagerank, mais a eu aussi des applications pratiques. Des applications parfois méconnues, mais pourtant intégrées dans des moteurs de recherche grand public comme Ask Jeeves !
- ❑ HITS a permis de nombreux développements destinés à étudier des portions **limitées** du web, ou de repérer des communautés dans le web.

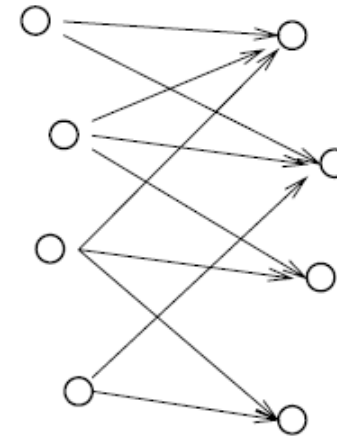
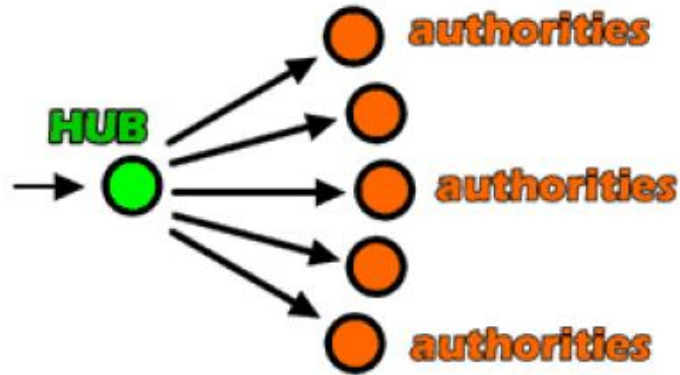
■ Principe de l'algorithme HITS : HUBS et AUTHORITIES

❑ L'algorithme HITS s'appuie sur un principe simple : tous les sites web n'ont pas la même importance, et ne jouent pas le même rôle.

➤ Certains sites sont des "sites de référence", leurs pages sont souvent citées dans d'autres sites. Ces sites de référence sont appelés "authorities".

➤ Alors que les "authorities" sont les véritables sites qui contiennent de l'information, d'autres sites appelés " hubs" jouent un rôle tout aussi important, bien qu'ils ne contiennent pas de contenu informatif... Il s'agit des sites qui contiennent des liens vers les "authorities", et qui permettent de "structurer" le Web en indiquant où sont les pages intéressantes pour un sujet donné.

Hubs & Authorities



hubs

authorities

- Si l'on observe la structure des liens, un hub se caractérise par la présence de nombreux liens sortants pointant vers des autorités, tandis que les autorités montrent surtout des liens entrants émanant des hubs
- L'analyse des hubs et autorités permet aussi de distinguer, sur le Web, l'existence de "communautés", c'est à dire de groupes de sites fortement liés entre eux.
C'est l'un des avantages de l'algorithme HITS (le Pagerank ne le permet pas aussi directement).

Soit :

- Une requête : Q
- Un moteur de recherche : MO
- Des constantes: n, m, k

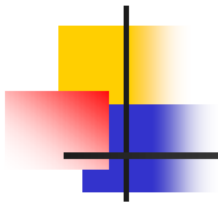
□ Suite à une requête Q , on obtient un ensemble de pages « Root Set »
 $R = \{ R_1, \dots, R_n \}$ qui contiennent les mots de la requête.

➤ **But** : On veut classer ces pages par rapport à leur importance sur cette requête.

Prendre note que : PageRank donne une importance globale, indépendante de la requête.

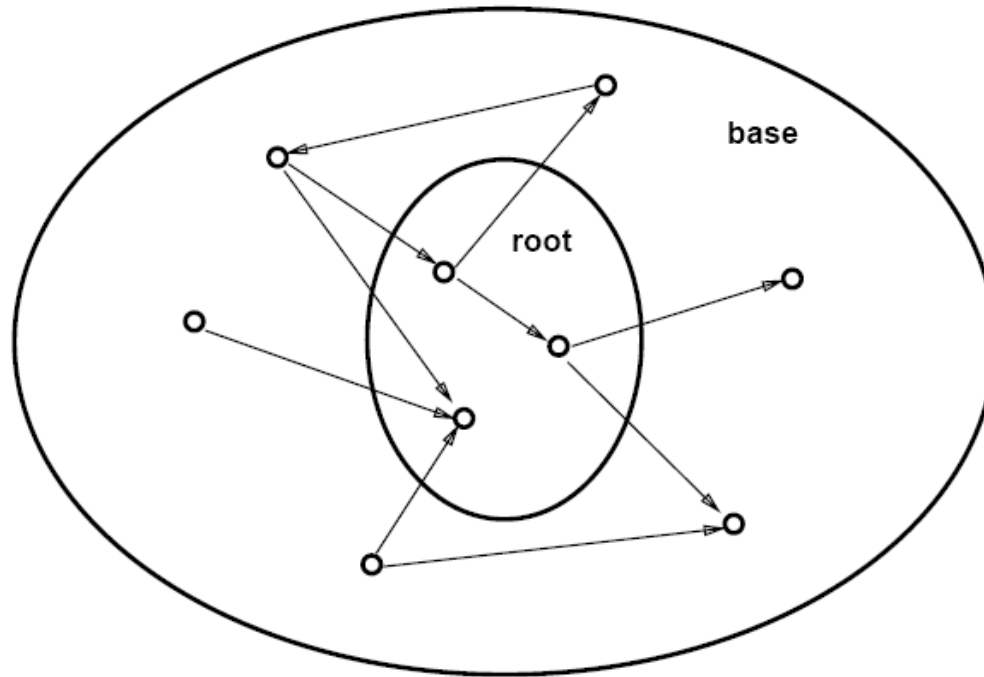
□ Pour cela, on considère deux ensembles supplémentaires de pages :

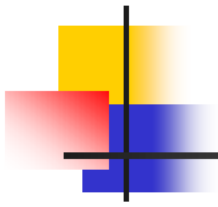
- l'ensemble "origine" $O = \{O_1, \dots, O_m\}$ qui sont les pages qui ont des liens vers les pages R_1, \dots, R_n
- l'ensemble "cible" $C = \{C_1, \dots, C_k\}$ qui sont les pages vers lesquelles les pages R_1, \dots, R_n ont une référence (lien).



Exemple

- En considérant les liens comme les arcs d'un graphe et chaque élément des ensembles R, O et C comme des nœuds, on obtient un graphe.





Hubs & Authorities

❑ HITS associe à chaque page deux scores : « Hub score » et « Authority score »

❑ **Intuition :**

➤ *A good Hub links to many good Authorities*

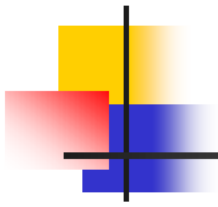
Une bonne page de liens pointe vers de bonnes pages de contenu

➤ *A good Authority is linked by many Hubs*

Une bonne page de contenu obtiendra beaucoup de liens de la part de bonnes pages de liens

❑ Les définitions des « Hubs » et des « Authorities » sont donc mutuellement récursives.

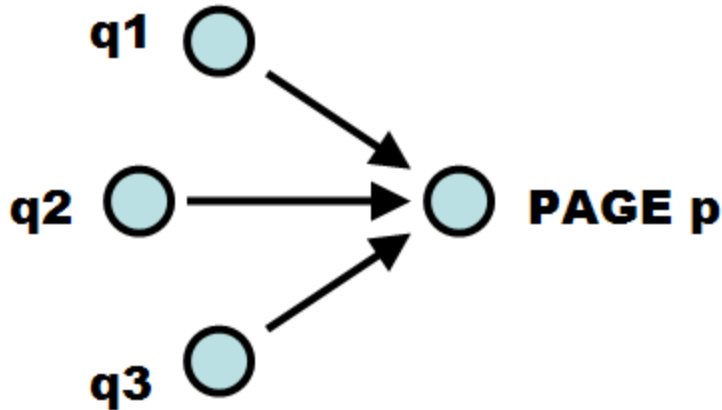
❑ Donc l'algorithme de calcul du score HITS est itératif avec comme limite la stabilisation des valeurs de scores Hubs et Authorities



Hubs & Authorities

Authorities:

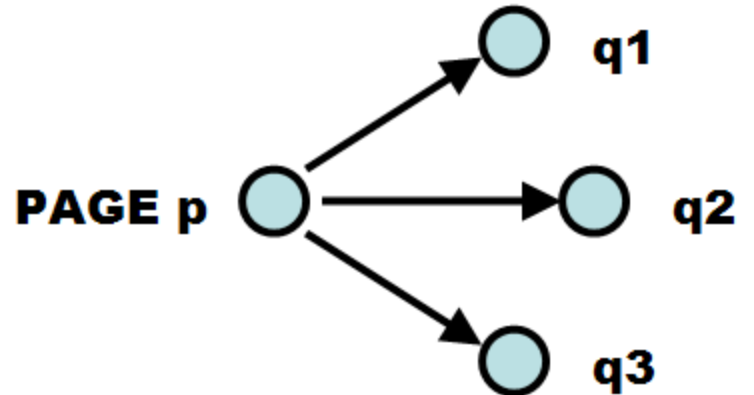
$$a(p) = \sum_{q \rightarrow p} h(q)$$



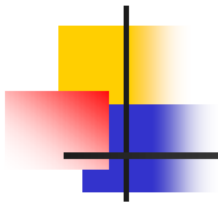
$$a(p) = h(q1) + h(q2) + h(q3)$$

Hubs:

$$h(p) = \sum_{p \rightarrow q} a(q)$$



$$h(p) = a(q1) + a(q2) + a(q3)$$



HITS – Implémentation

Initialisation : Pour toute page p appartenant à l'ensemble $G : a(p) = h(p) = 1$
avec $G = R \cup O \cup C$

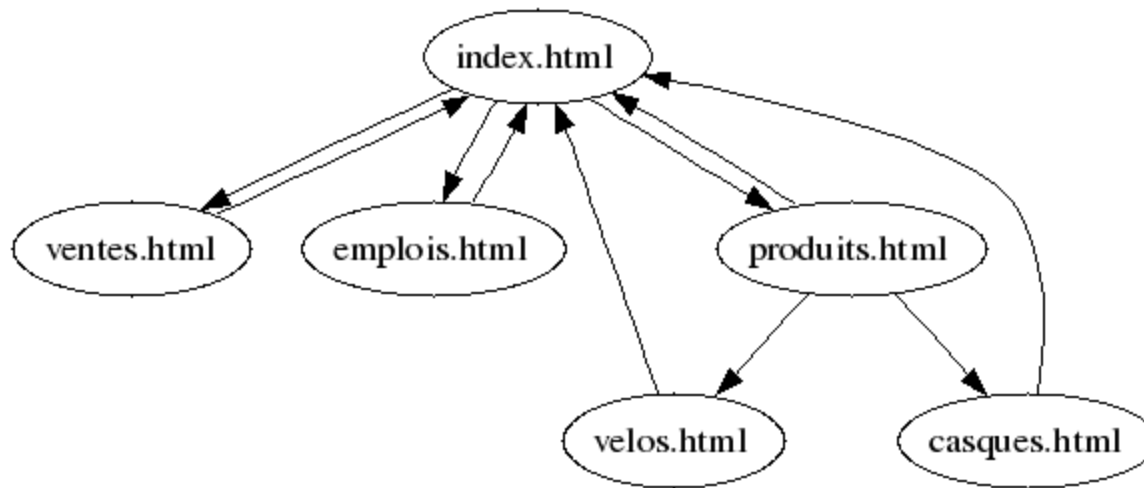
Itération : Répéter les trois opérations suivantes jusqu'à ce que les valeurs $a(p)$ et $h(p)$ convergent vers des valeurs stables.

1 - $a(p) = \sum_{q \rightarrow p} h(q)$ // une bonne page de contenu obtiendra beaucoup de liens de la part de bonnes pages de liens

2 - $h(p) = \sum_{p \rightarrow q} a(q)$ // une bonne page de liens pointe vers de bonnes pages de contenu

3 - $a(p) = \frac{a(p)}{\sqrt{\sum_{q \in G} a(q)^2}}$, $h(p) = \frac{h(p)}{\sqrt{\sum_{q \in G} h(q)^2}}$ // Normalisation des scores $a(p)$ et $h(p)$ de manière à ce que la somme de leurs valeurs au carré soit unitaire.

Exemple 1



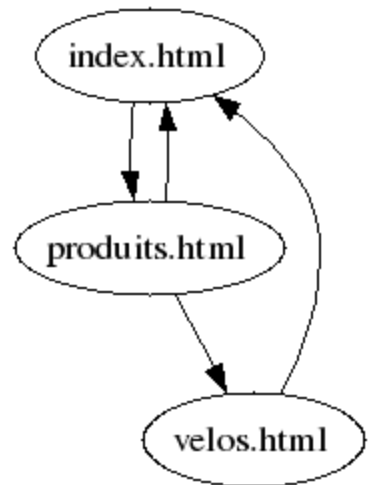
Le web comme graphe dirigé

Soit les pages avec leur description :

1. index.html pointe vers 3 pages (ventes.html, emplois.html, produits.html) ;
2. produits.html pointe vers 3 pages (velos.html, casques.html, index.html) ;
3. emplois.html pointe vers une page (index.html) ;
4. ventes.html pointe vers une page (index.html) ;
5. velos.html pointe vers une page (index.html) ;
6. casques.html pointe vers une page (index.html).

Exemple 1

- Supposons que, à la suite d'une recherche par mots clés : « rabais postal sur vélo », on obtient comme pages correspondantes seulement la page « velos.html ».
- Il faut alors ajouter toutes les pages qui font des liens vers « velos.html » et les pages qui reçoivent des liens en provenance de « velos.html ».



Ensemble-racine de la page velos.html



Exemple 1

Définissons alors la matrice A par $A_{ij} = 1$ si $j \rightarrow i$ et par $A_{ij} = 0$ sinon.

La première colonne correspond à la page « index.html », la seconde colonne correspond à la page « produits.html », et la troisième colonne correspond à la page « velos.html ».

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

On fixe $h(1) = h(2) = h(3) = 1$

$$a(p) = \sum_{q \rightarrow p} h(q) = A \begin{bmatrix} h(q_1) \\ h(q_2) \\ h(q_3) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$



Exemple 1

Pour calculer h à partir de a il suffit de prendre la transposée de A

$$\begin{bmatrix} h(p_1) & h(p_2) & h(p_3) \end{bmatrix} = A^T \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

Normalisation

$$a = \begin{bmatrix} 2 & 1 & 1 \end{bmatrix} / \sqrt{6}, \quad h = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix} / \sqrt{14}$$

On répète ensuite une seconde fois.

Estimation de a à partir de h

$$a = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{14} \\ 3/\sqrt{14} \\ 2/\sqrt{14} \end{bmatrix} = \begin{bmatrix} 5/\sqrt{14} \\ 1/\sqrt{14} \\ 3/\sqrt{14} \end{bmatrix}$$

Estimation de h à partir de a

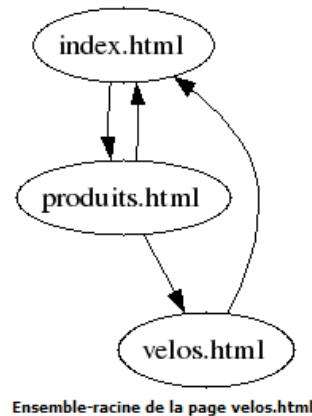
$$h = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5/\sqrt{14} \\ 1/\sqrt{14} \\ 3/\sqrt{14} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{14} \\ 8/\sqrt{14} \\ 5/\sqrt{14} \end{bmatrix}$$

Exemple 1

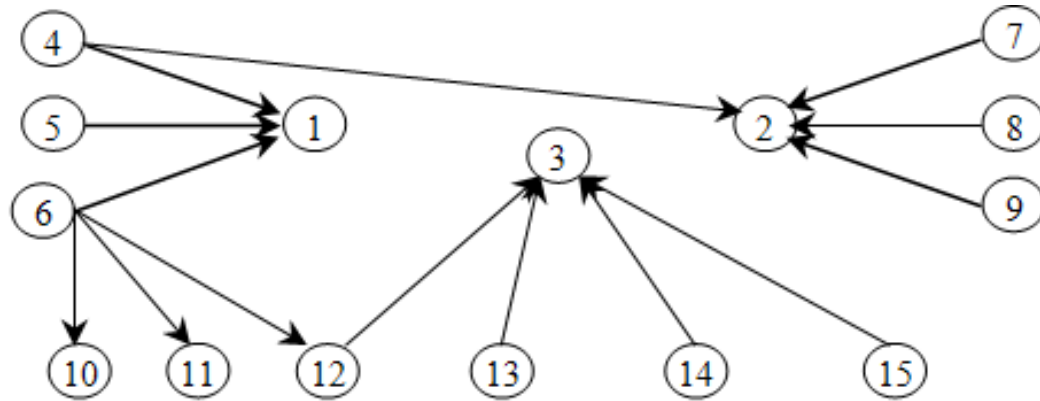
$$a = [0.85, 0, 0.53]$$

$$h = [0, 0.85, 0.53]$$

L'algorithme HITS recommande de visiter d'abord la page « index.html » et ensuite, seulement, la page « velos.html ». La page « produits.html » ne serait pas recommandée.

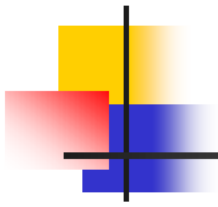


Exemple 2



Cette représentation graphique suggère que :

- Les nœuds qui doivent recevoir un grand « authority scores » sont : 1, 2 et 3.
- Les nœuds 1, 2 et 3 doivent avoir un « authority score » très élevé par rapport aux nœuds 10, 11 et 12.



Exemple 2

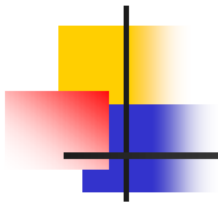
Avec HITS on trouve :

Les valeurs de « authority score » pour les nœuds 1, 2, 10, 11, 12 sont très élevés alors que le nœud 3 a un « authority score » très proche de zéro.

Cette performance de HITS peut être expliquée par le fait que

- ✓ Le « hub score » du nœud 6 augmente puisque il pointe vers plusieurs nœuds. Parmi eux il y a le nœud 1 qui a un « authority score » élevé.
- ✓ Le fait que le nœud 6 pointe vers 1 contribue à augmenter son « hub score » par conséquent « authority score » des nœuds 10, 11 et 12 va augmenter.
- ✓ Les nœuds qui pointent vers 3 ont relativement un faible « hub score » en comparaison avec les autres nœuds qui pointent vers 1 et 2.

Résultat, le nœud 3 va avoir un faible « authority score ».

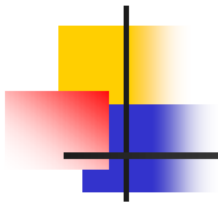


Exemple 2

Ce comportement de HITS est attribué au fait qu'il se base essentiellement sur la notion de renforcement mutuelle des relations entre les hubs et autorités

Question :

est ce la répartition des authorities score par HITS, pour le graphe de l'exemple 2 est appropriée?



HITS vs PageRank

➤ **HITS** est basé sur l'idée du renforcement mutuelle entre les hubs et authorities. A chaque page donc il estime un hub score et un authority score.

PageRank ne fait pas de distinction entre hubs et authorities. Il estime juste un authority score pour chaque page.

➤ **HITS** est appliqué sur un ensemble de page relativement petit qui dépend d'une requête bien spécifique.

PageRank est appliqué sur le Web en entier et il est indépendant des requêtes.

Note :

HITS et PageRank ne sont pas seulement utilisés dans le contexte de classement des pages Web seulement. Il y a plusieurs applications dont lesquelles HITS et PageRank peuvent êtres utiles. En classe, je parlerais de certains d'entre eux.