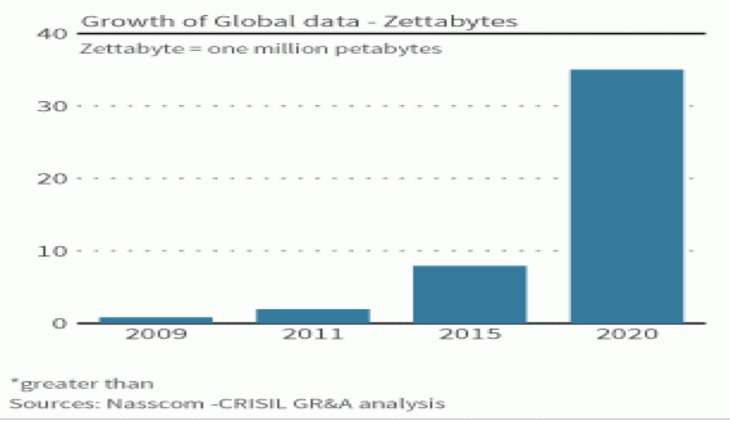
INTERROGATION DE DONNÉES DE TYPE GRAPHE

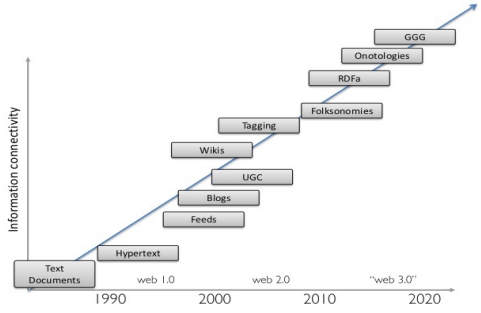
BDLE (BASES DE DONNÉES LARGE ECHELLE)
CAMELIA CONSTANTIN -- PRÉNOM.NOM@LIP6.FR

GRAPHES DE DONNÉES : EXEMPLES ET PROPRIÉTÉS

Croissance exponentielle du volume des données

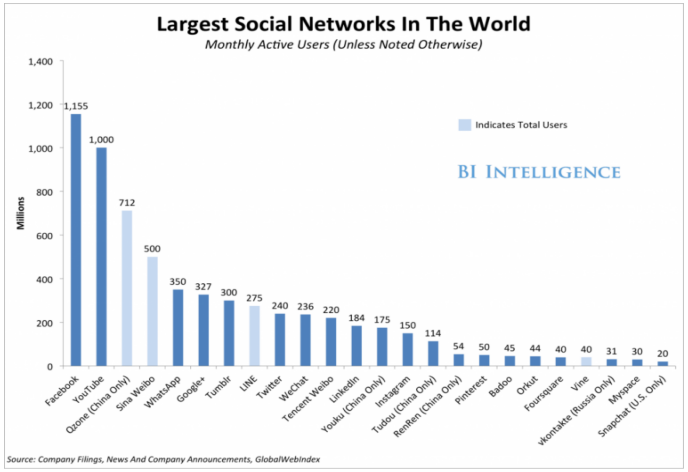


CONNECTIVITÉ DES DONNÉES



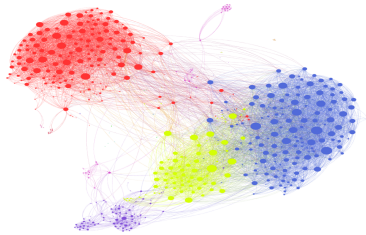
Volume x Connectivité = Complexité

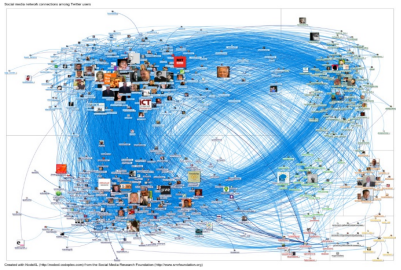
RÉSEAUX SOCIAUX: des graphes gigantesques

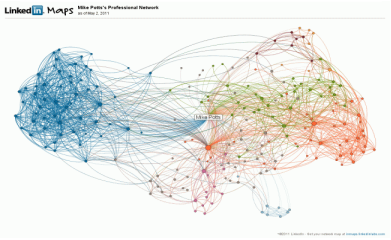


RÉSEAUX SOCIAUX (exemples 2014)

	Utilisateurs (10^6)	Arcs (10^9)
Facebook	1 300	400
LinkedIn	330	63
Twitter	650	130





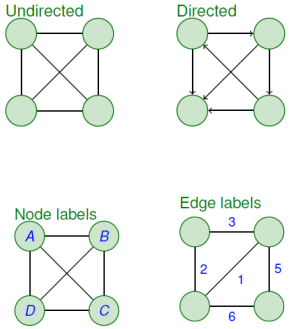


TYPES DE GRAPHE

- Graphes
  - dirigés : réseau social, citations bibliographiques web hypertexte, web sémantique, graphe de recommandation, graphe d'évolution…
  - non-dirigés : réseau routier, réseau des collaboration, graphes de cooccurrences, …

Graphes étiquetés

  - nœuds : nom, age, contenu
  - arcs : amitiés, coût, durée, ..



EXEMPLES DE REQUÊTES

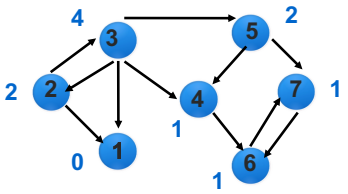
- Dans un réseau de type transport, alimentation, communication
  - Comment aller de l'adresse a à l'adresse b ?
  - Combien de chemins entre le nœud réseau a et le nœud réseau b ?
  - Le chemin le plus rapide entre l'usine a et le magasin b ?
- Dans un réseau de représentation de connaissance:
  - Une classe (élément) A est elle un sous-classe d'une classe B?
  - Existe-t'il un lien entre l'entité A et l'entité B?
  - Similarité entre l'entité A et l'entité B basé sur le graphe sémantique
- Réseaux de citations
  - Quels auteurs sont le plus cités directement et indirectement ?
  - Quels chercheurs / articles sont important dans un domaine (centralité) ?
- Réseaux sociaux
  - You Might Also Know de Facebook: si on partage beaucoup d'amis, on doit sans doute se connaître
  - Recommandation de lieux dans FourSquare d'après avis des amis: si des amis recommandent un lieu, alors de bonne chance que j'aime aussi
  - Degré de séparation (ex : nombre d'utilisateurs entre deux utilisateurs sur Facebook)

EXEMPLES DE REQUÊTES (SUITE)

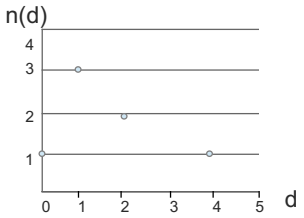
- Les voisins de A
- Voisins en commun entre A et B
- Existence / nombre de chemins entre A et B (et nombre, longueur, coût, etc.)
- Existence de chemins entre A et B correspondant à une expression régulière (A B)*(C|D)+
- Recherche du plus court chemin entre A et B, entre tous les nœuds
- Recherche de motifs de graphes : cycles, "motifs de graphes", arbre couvrant, circuit hamiltonien
- Calcul de propriétés : diamètre du graphe, centralité, ..

ANALYSE STATISTIQUE ET STRUCTURELLE

- Mesures de base :
  - Nombre de nœuds : |V|
  - Nombre d'arcs : |E| (plus important que |V|)
  - in/out-degree(n) : nombre d'arcs entrants/sortants
- L'histogramme d'un graphe (décrit le graphe)
  - diagramme de dispersion (d, n(d)) où n(d)=nombre de nœuds avec out-degree=d



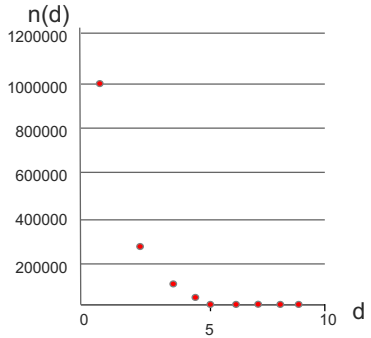
d	n(d)
0	1
1	3
2	2
3	0
4	1

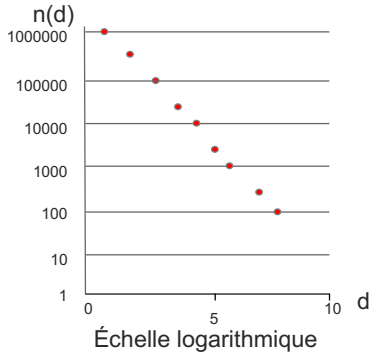


(en général cx^d pour x<1)

HISTOGRAMME : DISTRIBUTION EXPONENTIELLE

Graphe aléatoire $n(d) \simeq c(\frac{1}{2})^d$

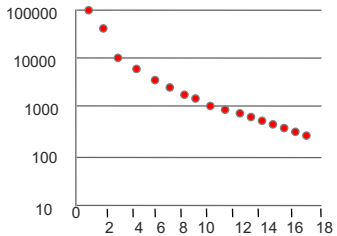


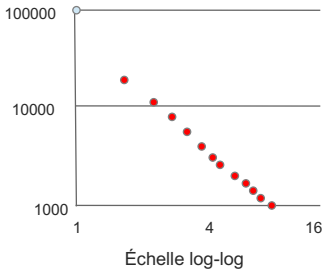


HISTOGRAMME : DISTRIBUTION ZIPF

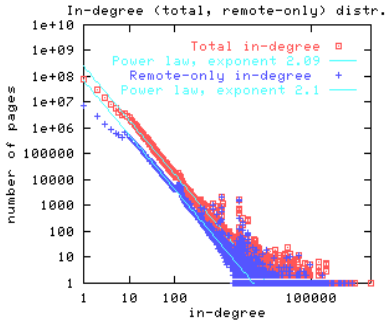
- Graphes générés par les applications
  - graphe occurrences mots / documents
  - graphe web

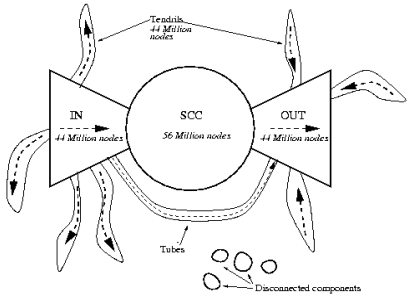
$n(d) \simeq \frac{1}{d^x}, x > 0$





GRAPHE DU WEB

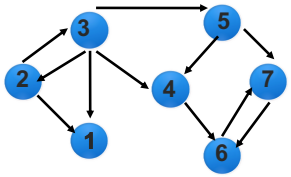




Pages Web (Broder et al.*, 2000) : A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener.Graph structure in the web. In WWW'00, pages 309-320. Crowl pages en 1999

CONNEXITÉ D'UN GRAPHE

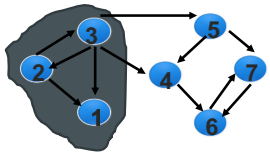
- Degré de connexité : nombre minimum de sommets qu'on doit enlever afin de déconnecter le graphe.
- Utile pour décider si un graphe est "intéressant" pour un certain type d'analyse.



Exemple : degré de connexité du graphe ?
Degré 1 : enlever 3 on a deux composantes déconnectées

CENTRALITÉ D'UN NOEUD

- Mesure l'importance d'un sommet v
- Utile par exemple pour l'analyse de communautés
- Centralité basée sur le degré : in-degree(v)/|E|
- Centralité de proximité:
  - Distance moyenne des plus courts chemins vers ce nœud (graphe dirigé) → nœud central = a une faible distance des autre nœuds
- Centralité d'intermédiarité de v:
  - La proportion des plus courts chemins entre deux autres sommets qui passent par v



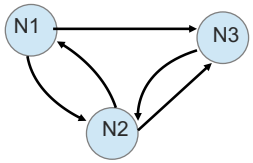
Exemple : centralité d'intermédiarité de 4 ?
(2,3,4,6), (3,4,6), (5,4,6)
3/nombre total de plus courts chemins
3/6 = 0.5

REPRÉSENTATION DES GRAPHES DE DONNÉES

MATRICE D'ADJACENCE

compte	nom	email	…
N1	Jean	…	
N2	Lucie	…	
N3	Marc	…	

Données



noeud	N1	N2	N3
N1	0	1	1
N2	1	0	1
N3	0	1	0

Matrice d'adjacence :

- Carrée (dimension n^2)
- Symétrique pour un graphe non-orienté

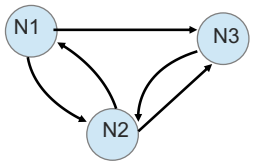
MATRICE D'ADJACENCE

- Avantages :
  - Utilisation de librairies d'algèbre linéaire si l'on dispose d'un stockage et d'un traitement efficaces de telles matrices (frameworks récents)
  - Temps constant pour trouver si un lien existe entre deux nœuds (en regardant l'entrée correspondante dans la matrice)
- Inconvénients :
  - Matrice de grande taille, creuse (beaucoup de valeurs zéro)
  - Recherche linéaire (inefficace) pour trouver les nœuds adjacents à un nœud donné (recherche linéaire sur toute la ligne correspondante au nœud i)

Liste d'arcs

compte	nom	email	…
N1	Jean	…	
N2	Lucie	…	
N3	Marc	…	

Données



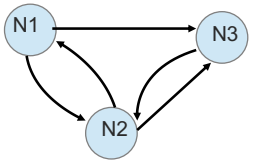
follower	followee
N1	N2
N1	N3
N2	N1
N2	N3
…	…

Liste d'arcs

LISTES D'ADJACENCE

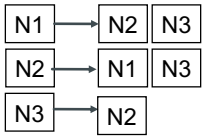
compte	nom	email	…
N1	Jean	…	
N2	Lucie	…	
N3	Marc	…	

Données



follower	followee
N1	N2
N1	N3
N2	N1
N2	N3
N3	N2

Liste d'arcs



Liste d'adjacence

COMPARAISON TAILLE

Exemple :LinkedIn (graphe non-dirigé, chaque entrée sur un bit)

- 330 millions de nœuds
- 63 milliards d'arcs
- Matrice d'adjacence

$330 * 10^6 * 330 * 10^6$ bits = $108{,}9*10^{15}$ /16 octets ≈

$7 . 10^{15}$ octets = 7 pétaoctets

- Table relationnelle / liste d'adjacence
  - 1 id de nœud = double = 4o
  - $63 * 10^9 * 2 * 4$ octets = $5 * 10^{11}$ octets = 0,5 téraoctets

COMPARAISON OPÉRATIONS

Mises-à-jour

- Matrice d'adjacence :
  - noeuds: insertion / suppression de lignes et colonnes (~ |N|)
  - arcs : maj cellule (constant)
- Liste d'adjacence
  - noeuds : creation de liste (const)
  - arcs: maj liste (~ out-degree)
- Liste d'arcs
  - arcs : insertion et suppression de nuplets (depend des index)

Accès

- Matrice d'adjacence
  - operations logiques (AND, OR)
  - operations matricielle : +, -, *, T
  - matrice = index "bitmap"
- Liste d'adjacence
  - parcours de listes

Liste d'arcs

- langages de requêtes
- stockage et indexation

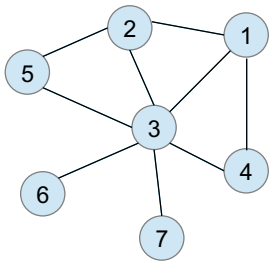
Comparaison des représentations

- Calcul des voisins communs dans Spark

  Comparaison liste d'arcs et liste d'adjacence

Liste d'arcs (DataFrame): N1 < N2

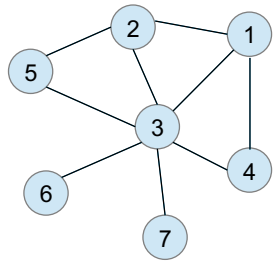
N1	N2
1	2
1	3
1	4
2	3
2	5
3	4
3	5
3	6
3	7



Comparaison des représentations

- Application: Calcul des voisins communs dans Spark
  - Calculer le nombre de voisins que l'utilisateur 1 a un commun avec tous les autres utilisateurs

Couple d'utilisateurs	Voisins communs
(1,2)	1
(1,3)	2
(1,4)	1
(1,5)	2
(1,6)	1
(1,7)	1



- Nombre de couples d'utilisateurs >> nombre de voisins directs de l'utilisateur -> Calcul efficace?

Comparaison des représentations

- Calcul des voisins communs dans Spark

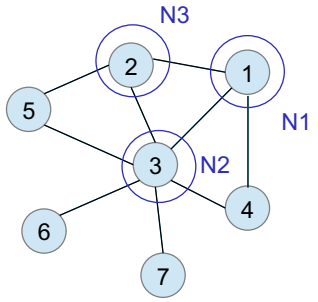
DataFrame (df1)

N1	N2
1	2
1	3
1	4
2	3
2	5
3	4
3	5
3	6
3	7

jointure

DataFrame (df2)

N2	N3
1	2
1	3
1	4
2	3
2	5
3	4
3	5
3	6
3	7



Calcul des voisins communs

joinDF = df1.join(df2, df1.N2 == df2.N2)

N1	N2
1	2
1	3
1	4
2	3
2	5
3	4
3	5
3	6
3	7

jointure

N2	N3
1	2
1	3
1	4
2	3
2	5
3	4
3	5
3	6
3	7

=

N1	N2	N2	N3
1	2	2	3
1	2	2	5
1	3	3	4
1	3	3	5
1	3	3	6
1	3	3	7
2	3	3	4
2	3	3	5
2	3	3	7

Calcul des voisins communs

joinDF

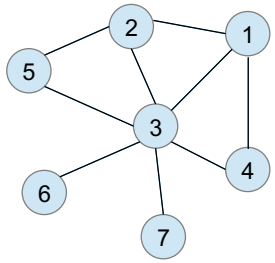
N1	N2	N2	N3
1	2	2	3
1	2	2	5
1	3	3	4
1	3	3	5
1	3	3	6
1	3	3	7
2	3	3	4
2	3	3	5
2	3	3	7

```
joinDF.drop(col('N2'))
      .groupBy('N1', 'N3')
      .count()
      .filter(col('N1')==1)
```

=

Résultat

N1	N3	count
1	3	1 ?
1	4	1
1	5	2
1	6	1
1	7	1



Manquent les voisins
Communs de 1 et de 2!

Calcul des voisins communs

joinDF = df1.join(df2, df1.N2 == df2.N2)

N1	N2
1	2
1	3
1	4
2	3
2	5
3	4
3	5
3	6
3	7
3	2

jointure

N2	N3
1	2
1	3
1	4
2	3
2	5
3	4
3	5
3	6
3	7
3	2

=

N1	N2	N2	N3
1	2	2	3
1	2	2	5
1	3	3	4
1	3	3	5
1	3	3	6
1	3	3	7
2	3	3	4
2	3	3	5
2	3	3	7
1	3	3	2

Calcul des voisins communs

N1	N2
1	2
2	1
1	3
3	1
1	4
4	1
2	3
3	2
2	5
5	2
3	4
4	3
3	5
5	3
3	6
6	3
3	7
7	3

jointure

N1	N2
1	2
2	1
1	3
3	1
1	4
4	1
2	3
3	2
2	5
5	2
3	4
4	3
3	5
5	3
3	6
6	3
3	7
7	3

=

N1	N2	N2	N3
1	2	2	1
1	2	2	3
1	2	2	5
1	3	3	1
1	3	3	2
1	3	3	4
1	3	3	5
1	3	3	6
1	3	3	7
1	4	4	1
1	4	4	3

joinDF .filter(col('N1')==1)

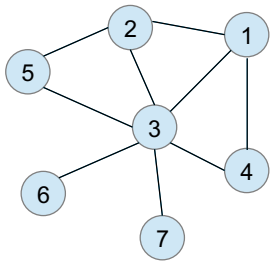
Calcul des voisins communs

N1	N2	N2	N3
1	2	2	1
1	2	2	3
1	2	2	5
1	3	3	1
1	3	3	2
1	3	3	4
1	3	3	5
1	3	3	6
1	3	3	7
1	4	4	1
1	4	4	3

```
joinDF.drop(col('N2'))
      .groupBy('N1', 'N3')
      .count()
```

=

N1	N3	count
1	3	2
1	4	1
1	5	2
1	6	1
1	7	1
1	1	3
1	2	1



Résumé du calcul basé sur les tables d'adjacence

Algorithme de calcul et problèmes:

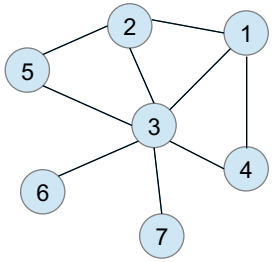
- Modifier le DF: pour chaque arc x->y ajouter l'arc y->x ← Problème: duplication des données
- Auto-jointure du DF construite à l'étape précédente ← Problème: shuffle des données
- Enlever les identifiants des amis communs
- Compter le nombre d'occurrences de chaque paire d'utilisateurs ← Problème: shuffle des données

Comparaison des représentations

- Calcul des voisins communs dans Spark
  Comparaison liste d'arcs et liste d'adjacence

Liste d'adjacence (DataFrame): listes triées

N	Voisins
1	[2, 3, 4]
2	[1, 3, 5]
3	[1, 2, 4, 5, 6, 7]
4	[1, 3]
5	[2,3]
6	[3]
7	[3]



Les utilisateurs appartenant à la même liste d'adjacence ont un utilisateur en commun (N)

Calcul initial basé sur les listes d'adjacence

Algorithme de calcul et problèmes:

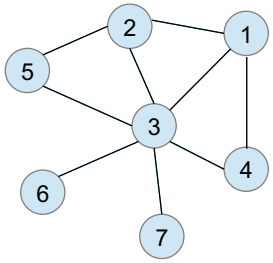
1. Pour chaque liste d'adjacence construire **tous les couples d'utilisateurs:**

   Exemple pour 1: [2,3], [3,2], [2,4], [4,2], [3,4], [4,3] ← Problème: duplication des données

2. Pour chaque couple compter le nombre d'occurrences (= nombre de voisins que les nœuds du couple ont en commun) ← Problème: shuffle des données pendant le groupBy

Réduction du volume de données générées à la première étape:

Constat : les couples [n1, n2] et [n2, n1] ont le même nombre de voisins → construire uniquement des couples ordonnés par ordre croissant des indentifiants



Comparaison des représentations

Liste d'adjacence: listes triées

N	Voisins
1	[2, 3, 4]
2	[1, 3, 5]
3	[1, 2, 4, 5, 6, 7]
4	[1, 3]
5	[2,3]
6	[3]
7	[3]

→

Couples contenant l'utilisateur 1

[N1, N2]
[1, 3]
[1, 5]
[1, 2]
[1, 4]
[1, 5]
[1, 6]
[1, 7]
[1, 3]

→

Nombre d'utilisateurs en commun

[N1, N2]	count
[1, 3]	2
[1, 5]	2
[1, 2]	1
[1, 4]	1
[1, 6]	1
[1, 7]	1

Nombre d'utilisateurs en commun correct

Pour [1,3] et [1,2] sans dupliquer les données!

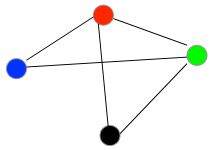
Compter les triangles

Pourquoi compter les triangles ?

Coefficient de clustérisation :
Pour un graphe non dirigé G=(V,E)
cc(v) = fraction des voisins de v qui sont eux-mêmes des voisins

$$= \frac{|\{(u,w) \in E | u \in \Gamma(v) \wedge w \in \Gamma(v)\}|}{\binom{d_v}{2}}$$

$\binom{d_v}{2}$ est le nombre total d'arcs possibles entre les voisins de v



cc(●) = 1/1

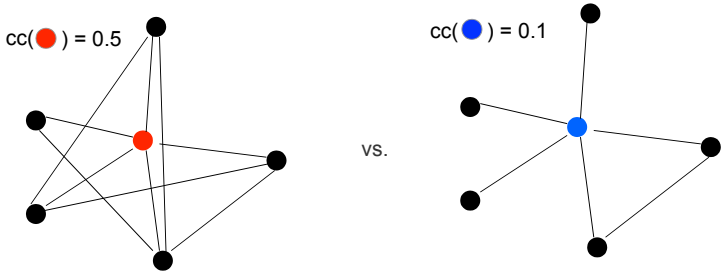
cc(●) = 2/3

cc(●) = 1/1

cc(●) = 2/3

Coefficient de clustérisation

Montre la densité de la connectivité autour d'un noeud



Comment compter les triangles ?

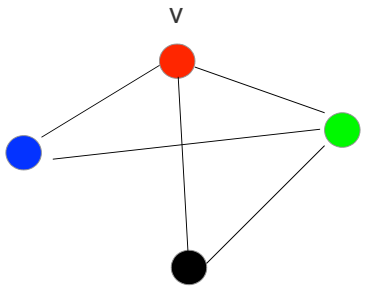
Algorithme séquentiel (graphe non dirigé) :

```
Triangles ← 0
foreach v in V
    foreach u,w in Adjacency(v)
        if (u,w) in E
            Triangles += 1/2
return (Triangles / 3)
```

Complexité de l'algorithme $O\left(\sum d_v^2\right)$

Chaque triangle est compté 3 fois (une fois par noeud)

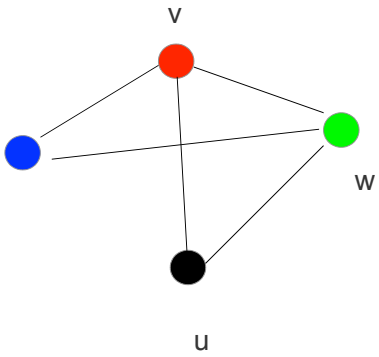
Exemple



À partir du nœud v

Triangles = 0

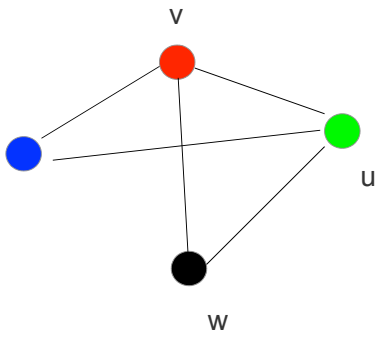
Exemple



À partir du nœud v

Triangles = 0.5

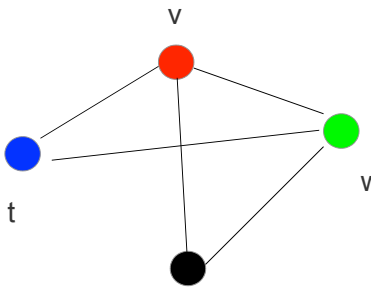
Exemple



À partir du nœud v

Triangles = 1

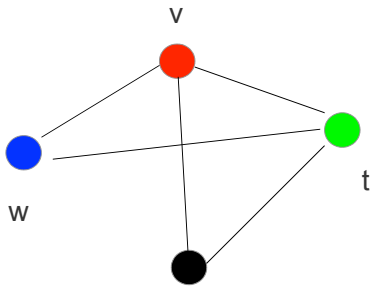
Exemple



À partir du nœud v

Triangles = 1.5

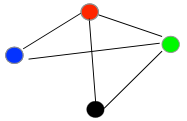
Exemple



À partir du nœud v

Triangles = 2

Algorithme parallèle



Map 1 : Input : $\{(v,u)|\ u\in\Gamma(v)\}$

foreach $u\in\Gamma(v)$ emit {(v,u)}

Reduce 1 : Input : $\{(v,u)\ |\ u\in\Gamma(v)\}$

foreach (u,w) : $u,w\in\Gamma(v)$

emit {((u,w), v)} //tous les arcs possibles dans le graphe

Exemple : ((●●)●) ((●●)●) ((●●)●),.....

Map 2 : emit { ((u,w), $)| $w\in\Gamma(u)$ }

Reduce 2 : Input : {(u,w)| v1,v2,...vk,$?}

foreach (u,w) if $ is part of the input, then :

Triangles[vi] +=1/2

((●●)● $) → Triangles(●)+1/2

(●●)● → ∅

Adaptation de l'algorithme

- On génère tous les chemins à vérifier en parallèle, le temps d'exécution est $max_{v \in V}(\sum d_v^2)$ => pour les nœuds avec beaucoup de voisins (millions) les reducer tasks correspondants peuvent être très lents

Amélioration :

- ordonner les nœuds par leur degré (pour ceux qui ont le même degré par leur identifiant)
- compter chaque triangle une seule fois, à partir du nœud minimum
- complexité : $O(m^{3/2})$ (m = nombre d'arcs dans le graphe)

Algorithme amélioré

Algorithme séquentiel :

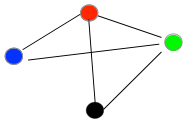
```
Triangles ← 0
foreach  v  in  V
    foreach u,w in Adjacency(v)
        if u > v && w > u
            if  (u,w)  in  E
                Triangles++
Return Triangles
```

Algorithme parallèle

Map 1 : Input : $\{(v,u) | u \in \Gamma(v)\}$

if u > v then emit {(v,u)}

Exemple : (●● ●) (●● ●) (●● ●) (●● ●)(●● ●)



Reduce 1 : Input : $\{(v,u) | u \in S \subset \Gamma(v)\}$

foreach (u,w) : $u,w \in S$ //u,v:voisins de v (v< u et v< w)

if w > u then emit {((u,w), v)}

Exemple : ((●●)●) ((●●)●) ((●●)●)

●< ●< ●< ●

Map 2 : emit { ((u,w), $)| $w \in \Gamma(u)$ }

Reduce 2 : Input : {(u,w)| v1,v2,...vk,$?}

foreach (u,w) if $ is part of the input, then : Triangles[vi] ++

Exemple : ((●●)● $) → Triangles(●)++

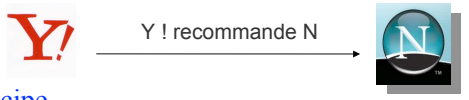
(●●)● → ∅

Algorithmes de calcul sur les graphes

PageRank

PageRank: principe

Liens hypertexte = recommandations



Principe

- Les pages avec beaucoup de recommandations sont plus importantes
- Importance aussi de *qui* donne la recommandation

  être recommandé par Yahoo! est mieux que par X

  la recommandation compte moins si Yahoo! recommande beaucoup de pages

  → l'importance d'une page dépend du nombre et de la qualité (importance de celui qui recommande) de ses liens entrants

PageRank simplifié

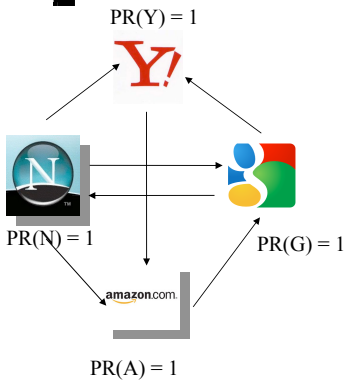
Recommandation donnée par **Y!** à **N** :

$$\frac{PR(Y!)}{|out(Y!)|} \quad \text{où} \quad \begin{cases} PR(Y!) = l'importance de Y! \\ |out(Y!)| = nombre de liens sortants de Y! \end{cases}$$

Importance de **N**etscape est la somme de ses recommandations

$$PR(N) = \sum \frac{PR(P)}{|out(P)|}$$ P = pages qui recommandent N

Exemple



PR(A) = PR(N) /3 + PR(Y)

PR(Y) = PR(N) /3 + PR(G)/2

PR(N) = PR(G)/2

PR(G) = PR(A) + PR(N)/3

Calcul des valeurs PR

Résolution système linéaire

- 4 équations avec 4 inconnues
- pas de solution unique

→ ajouter la contrainte PR(A)+PR(Y)+PR(N)+PR(G) = 1 pour assurer l'unicité

Observation:

- système linéaire de grandes dimensions, beaucoup de pages sans liens sortants => les méthodes de calcul directes (ex. méthode de Gauss) sont plus coûteuses que les *méthodes itératives*

Représentation matricielle

On considère *n* pages, pour chaque page i, on note:

- *out(i)* est l'ensemble de pages j référencées par i

$M(m_{ij})$ est la matrice d'adjacence associée au graphe du Web

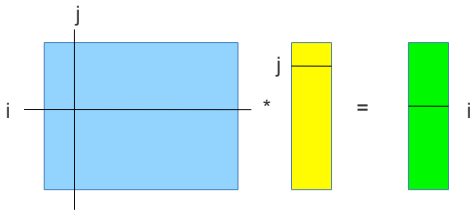
- m_{ij}: fraction de l'importance de j qui est donnée à i ($m_{ij} = 1/|out(j)|$, si j a des liens sortants, $p_{ij} = 0$ dans le cas contraire)
- ligne i = fractions d'importance reçues par i
- colonne j = distribution de l'importance de j (pour les pages j avec des liens sortants, la somme des éléments sur les colonnes est 1)

$PR(PR_1, PR_2,...,PR_n)$ est le vecteur des inconnues (importance)

PR_i est l'importance de la page i

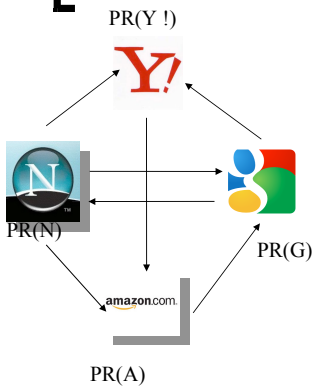
Exemple

Mise à jour de PR_i : $PR_i = \sum \frac{PR_j}{|out(j)|}$



$$PR_i = \sum m_{ij} * PR_j$$

Exemple



PR(A) = PR(N) /3 + PR(Y)

PR(Y) = PR(N) /3 + PR(G)/2

PR(N) = PR(G)/2

PR(G) = PR(A) + PR(N)/3

PR	=	Y	N	A	G	*	PR
PR(Y)	Y		1/3		1/2		PR(Y)
PR(N)	N				1/2		PR(N)
PR(A)	A	1	1/3				PR(A)
PR(G)	G		1/3	1			PR(G)

M

Algorithme de calcul itératif

- Un graphe avec n nœuds
- Initialisation : $PR^0 = [1,\ldots,1]$
- À chaque itération k, recalculer $PR^{(k)}$

$$PR^{(k)} = M * PR^{(k-1)}$$

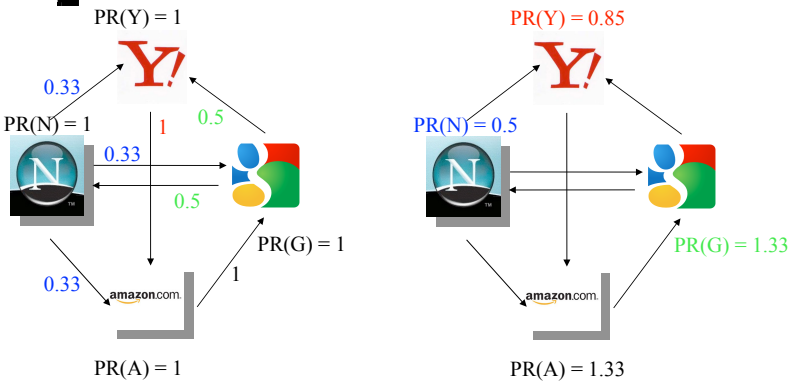
- Arrêt du calcul (convergence) :

$$\frac{\sum |PR_i^k - PR_i^{k-1}|}{\|PR^k\|_1} < \varepsilon, \varepsilon \in (0,1)$$

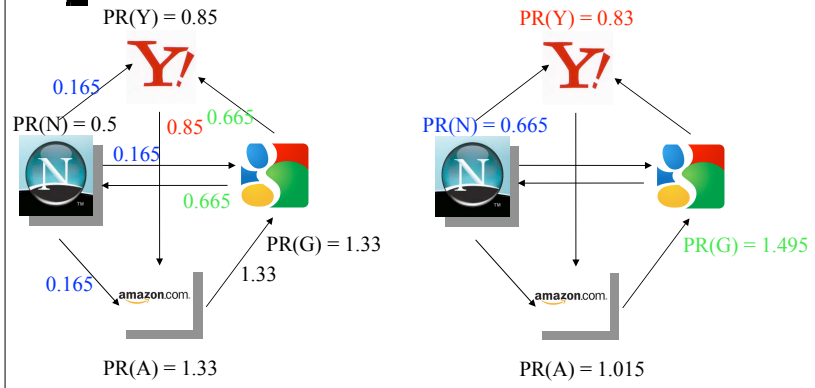
Le vecteur PR obtenu à la convergence satisfait la condition :

$$PR = M * PR$$

Exemple – Itération 1



Exemple – Itération 2



Algorithme itératif complet

Un graphe avec n nœuds

- Initialisation : $PR^0 = [1/N,\ldots,1/N]$
- Vecteur ajouté à chaque itération : $V = [1/N,\ldots,1/N]$
- À chaque itération k, recalculer $PR^{(k)}$

$$PR^{(k)} = d*M * PR^{(k-1)} + (1-d)*V$$

(ou *équivalent* : $\forall i : PR_i^k = d * \sum \frac{PR_j}{|out[j]|} + \frac{(1-d)}{N}$)

arrêt lorsque : $\frac{\sum |PR_i^k - PR_i^{k-1}|}{\|PR^k\|_1} < \varepsilon, \varepsilon \in (0,1)$

le facteur de décroissance d (valeur habituelle 0.85) est utilisé pour assurer l'unicité du vecteur PR calculé et la convergence du calcul itératif

PageRank en MapReduce

method **MAP** (nodeid n, vertex N)

$$p \leftarrow d * \frac{N.PAGERANK}{|N.OUT|}$$

EMIT(n, N)

for all nodeid m in N.OUT **do**

EMIT(m, p)

method **REDUCE** (nodeid m, [p_1, p_2…])

M ← null, s ← (1-d)/NUMVERTICES

for all p in [p_1, p_2…] **do**

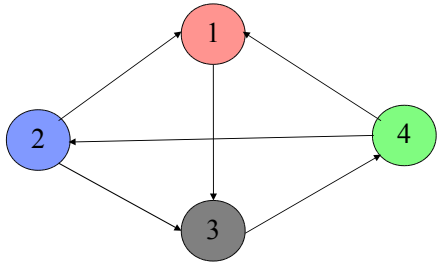
if ISVERTEX(p) **then** M ← p

else s ← s + p

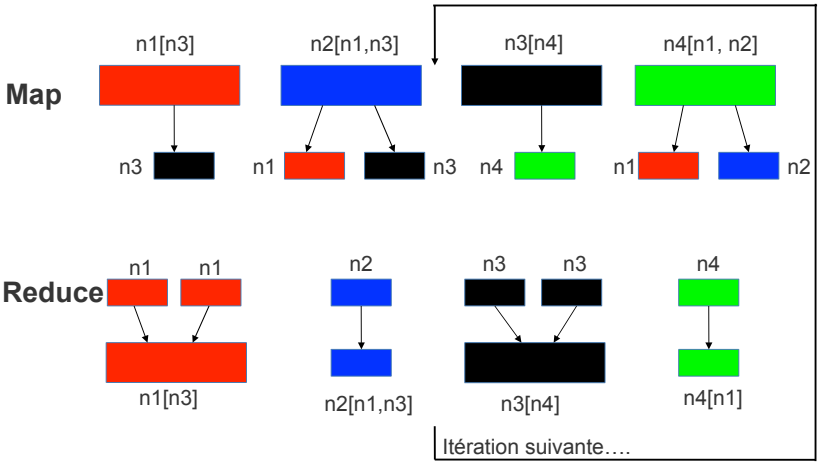
M.PAGERANK ← s

EMIT(m, M)

Exemple



PageRank en MapReduce



Calcul de PageRank en M/R

Map :

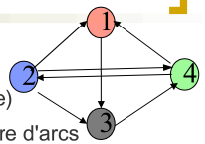
- Un Map task travaille sur une portion de la matrice M et du vecteur $PR^{(k-1)}$ et produit une partie de $PR^{(k)}$
- La fonction Map s'applique à un seul élément mij (1/|out(j)|) de la matrice M et produit la paire (i, d*mij*PRj) =>tous les termes de la somme qui permet de calculer le nouveau PRi auront la même clé
- On utilise également un combiner pour agréger localement les valeurs produites par le même Map task

=> Définir des **stratégies de partitionnement de la matrice et des vecteurs** qui tiennent compte de la capacité mémoire des nœuds de calcul exécutant les tasks

Reduce :

- La fonction Reduce aditionne les termes avec la même clé i, et produit la paire (i, PRi)

Encodage de la matrice



- Matrice creuse (beaucoup d'entrées sont 0)
- Stocker les entrées non nulles (listes d'adjacence)
- L'espace de stockage est proportionnel au nombre d'arcs
- Pour chaque nœud on stocke également le nombre de liens sortants
- Le vecteur PR est également de grandes dimensions (prop. au nombre de noeuds)

	1	2	3	4
1	0	1/3	0	1/2
2	0	0	0	1/2
3	1	1/3	0	0
4	0	1/3	1	0

→

Nœud source	*degré*	*Nœuds destination*
1	1	3
2	3	1, 3, 4
3	1	4
4	2	1, 2

Partitionnement en bandes

- **B** Map tasks

Map

- La matrice M est partitionnée en **B** bandes verticales (une bande correspond à un ensemble de nœuds source)
- Le vecteur $PR^{(k-1)}$ est partitionné en **B** bandes horizontales

=> chaque task reçoit une bande M_b de M et la bande correspondante de $PR_b^{(k-1)}$

- Chaque task produit une version locale du vecteur PR^k (de même taille que le vecteur PR^k) : $PR_b^k=((1,PR_b^k(1)), \ldots, (n,PR_b^k(n))$

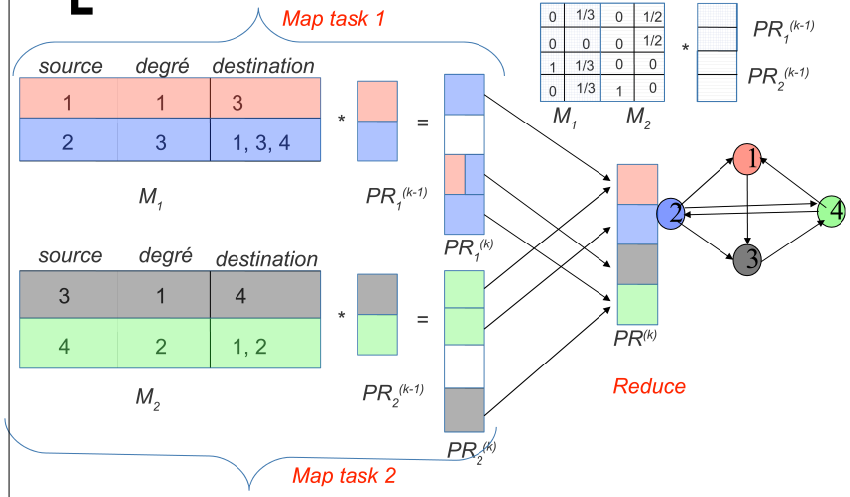
Reduce

- Aggrégation somme des vecteurs PR_b^k en fonction des clés

Avantage de cette méthode : on stocke seulement une partie(bande) de M et de $PR^{(k-1)}$ dans la mémoire locale d'un nœud de calcul

Inconvénient : on doit stocker le vecteur PR_b^k entièrement(peut avoir la même taille que PR^k) =>problème si pas assez de mémoire

Exemple



Partitionnement en blocks

B*B Map tasks

Map

M_{11}	….	M_{1B}
:	:	:
M_{B1}	….	M_{BB}

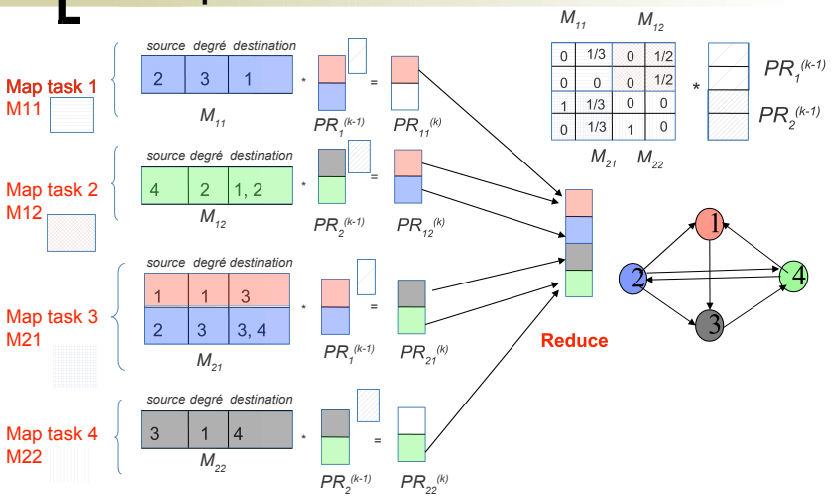
- La matrice M est partitionnée en ***B*B** blocks*
- Le vecteur $PR^{(k-1)}$ est partitionné en ***B** bandes horizontales*

  => chaque task reçoit un block M_{ib} de M et une bande de $PR_b^{(k-1)}$ (*$PR_b^{(k-1)}$ est transmise B fois :*à chaque task qui reçoit un block M_{ib}, pour i de 1 à B)
- Chaque task produit une version locale du vecteur PR^k_i :
- $PR_{ib}^k=((1,PR_{ib}^k(1)), \ldots, (i,PR_{ib}^k(i))$

  Reduce : somme des vecteurs PR_{ib}^k en fonction des clés
- **Avantage de cette méthode :** on stocke seulement une partie(block) de M et de $PR^{(k-1)}$ dans la mémoire locale d'un nœud de calcul, ainsi qu'une partie du vecteur final =>tout peut être stocké en mémoire

  Inconvénient : chaque bande $PR_b^{(k-1)}$ du vecteur $PR^{(k-1)}$ doit être répliquée plusieurs fois

Exemple



Calcul M/R : Problèmes

- Le graphe est transmis (shuffled) à chaque itération (objet **vertex N** transmis comme paramètre aux deux méthodes, il inclut la liste d'adjacence OUT)
- On veut pouvoir transmettre uniquement la nouvelle valeur d'importance et non pas la structure du graphe
- On doit contrôler les itérations en dehors de M/R (conditions de terminaison et logique du programme)

=> **Pregel** (calcul sur des graphes de grande taille)

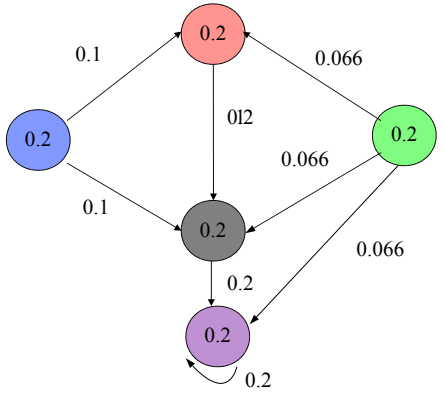
- Proposé par Google
- Implémentations open source : Apache Giraph, Stanford GPS, Apache Hama
- Calcul réalisé:
  - Pour chaque nœud en parallèle, tant qu'il y a des **nœuds actif**s (ou le nombre maximum d'itérations n'a pas été atteint) :
    - Traiter les messages reçus des la part de ses voisins à l'itération précédente
    - Envoyer des messages aux voisins
    - Devenir inactif si plus de calculs à réaliser

PageRank avec Pregel

```
class PageRankVertex
    : public Vertex<double, void, double> {
 public:
  virtual void Compute(MessageIterator* msgs) {
    if (superstep() >= 1) {
      double sum = 0;
      for (; !msgs->Done(); msgs->Next())
        sum += msgs->Value();
      *MutableValue() =
          0.15 / NumVertices() + 0.85 * sum;
    }

    if (superstep() < 30) {
      const int64 n = GetOutEdgeIterator().size();
      SendMessageToAllNeighbors(GetValue() / n);
    } else {
      VoteToHalt();
    }
  }
};
```

Si l'importance n'a pas changé sufisamment le noeud devient inactif (VoteToHalt())

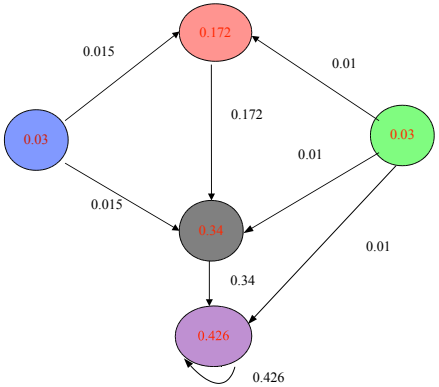
Exemple



sum = somme des valeurs reçues

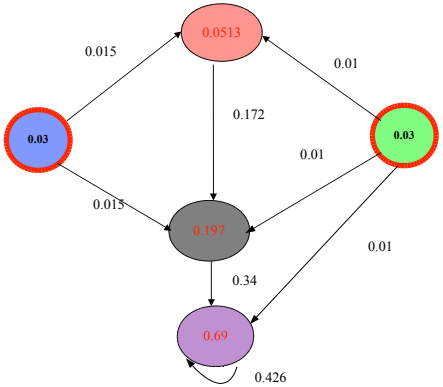
importance=0.85*sum+0.15/5

Exemple



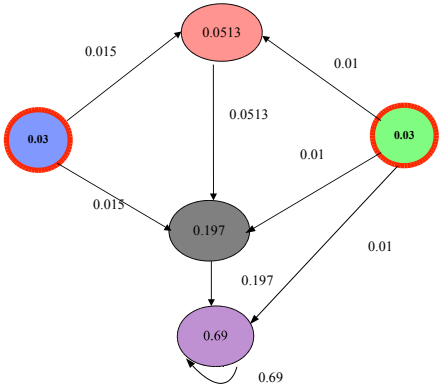
sum = somme des valeurs reçues
importance=0.85*sum+0.15/5

Exemple



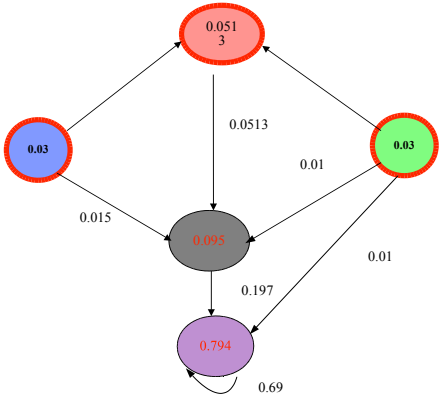
sum = somme des valeurs reçues
importance=0.85*sum+0.15/5

Exemple



sum = somme des valeurs reçues
importance=0.85*sum+0.15/5

Exemple



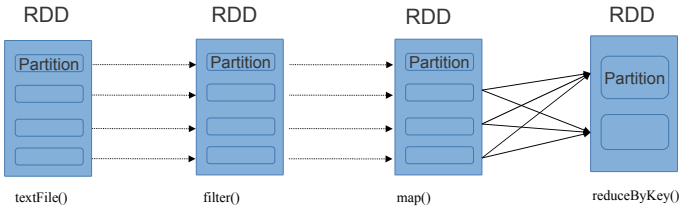
sum = somme des valeurs reçues
importance=0.85*sum+0.15/5

Modèle d'exécution Spark

Partitionnement de données

Exemple

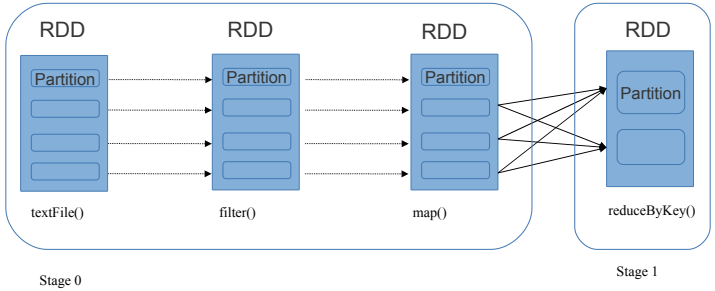
```
// Fichier avec des entrées [URL, idU, noms des pages ]
//  pour chaque utilisateur idU les adresses et les noms des pages visitées
val pages = sc.textFile("pages.txt")

// Créer une RDD avec des pairs (idU, nbVisites) :
// pour chaque utilisateurs le nombre de pages visitées qui ne contiennent pas 'database'
val visits = pages.filter(lambda x : "database" not in x)
                  .map(lambda x : (x.split("\t")[1], 1)).reduceByKey(lambda x, y : x+y)
```



Séquence de transformations

Exemple

- RDD : collection distribuée d'objets regroupés en partitions stockées et traitées par executors
- transformations : RDD en entrée et sortie
- Chaîne de traitement: séquence de RDD générées par les transformations, avec des dépendances entre elles
- Étape (Stage) : segment de la chaîne de traitement qui peut s'exécuter localement (sans échanges réseau)
- Les étapes sont séparées par des shuffle (redistribution de données)



Partitionnement de données

- RDD : collections de données de grande taille, ne peuvent pas être stockées dans un seul nœud, doivent être partitionnées sur plusieurs noeuds
- Coût de communication élevé dans un programme distribué : utiliser le partitionnement des RDD entre les nœuds

Partitionnement

- Défini pour des pairs (clé, valeur), divise les données en utilisant une fonction applicable sur la clé, définit quelles sont les données traitées par chaque tâche
- Le partitionnement a un coût
  - utile pour des données qui sont re-utilisées plusieurs fois dans des opérations basées sur des clés (eg. **Join**)
  - Inutile de partitionner à l'avance une RDD qui sera utilisée une seule fois

Partitions

- Sont évaluées en mode "lazy"(déclenché par une action, à chaque action Spark recalcule le DAG des transformations)
- Chaque nœud contient une ou plusieurs partitions, chaque partition est stockée sur une seule machine, on ne peut pas choisir le nœud pour une partition
- Une tâche par partition
- Le nombre de partitions peut être configuré(pas assez : pas assez de concurrence, mauvaise utilisation des ressources, trop de partitions : plus de temps pour affecter les tâches que pour leur exécution), par défaut : nb total de cœurs du cluster.

Créer des partitions

- *PartitionBy* : retourne une nouvelle RDD (ne modifie pas celle qui existe)
- Utiliser **persist()** après partitionBy pour ne pas exécuter le partitionnement à chaque accès
- Le nombre de partitions détermine le nombre de tasks qui vont exécuter les opérations en parallèle sur la nouvelle RDD (doit être ≥ au nombre de coeurs du cluster)

Déjà existants en Spark :

- HashPartitioner(nbPartitions: Int) :
  - les données dont les clés ont la même valeur % nbPartitions apparaissent dans la même partition
- RangePartitioner(nbPartitions: Int, rdd: RDD[_ <: Product2[K, V]], ascending: Boolean = true) :
  - les données avec les clés dans la même plage de valeurs apparaissent dans la même partition

Partitionnement

Chaque RDD a un objet `Partitioner` qui est optionnel, interface avec les méthodes :

- numPartitions() et getPartition()

Pour chaque opération qui implique un déplacement de données :

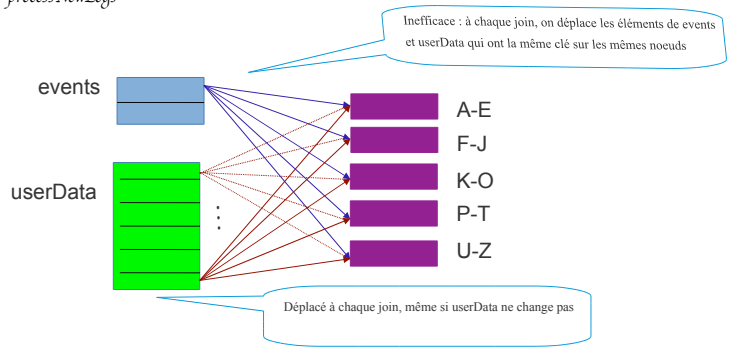
- Pour une RDD qui a un `Partitioner`, l'opération prend en compte ce `Partitioner`, les valeurs pour chaque clé sont calculées d'abord localement sur chaque machine et seulement le résultat est envoyé au master
- Appliquée à deux RDD, elle prend en compte le `Partitioner` d'une des deux RDD, s'il existe (les données de cette RDD ne seront pas déplacées). Aucune donnée n'est déplacée si :
  - Les deux RDD ont le même `Partitioner` et sont distribuées sur les mêmes machines ou l'une d'entre elles n'est pas encore calculée
- Spark utilise par défaut `HashPartitioner` (nb partitions=degré de parallélisme)

Exemple

```
val sc = new SparkContext(...)
//charger les informations une seule fois à partir d'un fichier Hadoop SequenceFile,
//fichier lu effectivement lors de l'exécution de count(), utiliser persist() pour éviter de le relire
//UserInfo : liste de sujets qui représentent l'intérêt de l'utilisateur
val userData = sc.sequenceFile[UserID, UserInfo]("hdfs://...").persist()

//méthode appelée périodiquement pour obtenir le log des actions de l'utilisateur pendant les 5 dernières minutes
//LinkInfo : information sur les liens visités par l'utilisateur
def processNewLogs(logFileName: String) {
  val events = sc.sequenceFile[UserID, LinkInfo](logFileName)
  val joined = userData.join(events)    // RDD  (UserID, (UserInfo, LinkInfo))
  val offTopicVisits = joined.filter {
    case (userId, (userInfo, linkInfo)) =>
      !userInfo.topics.contains(linkInfo.topic)
  }.count()  // Combien d'utilisateurs ont visité un lien qui n'appartient pas à un sujet de leur intérêt
  println("Number of visits to non-subscribed topics: " + offTopicVisits)
}
```

Exemple

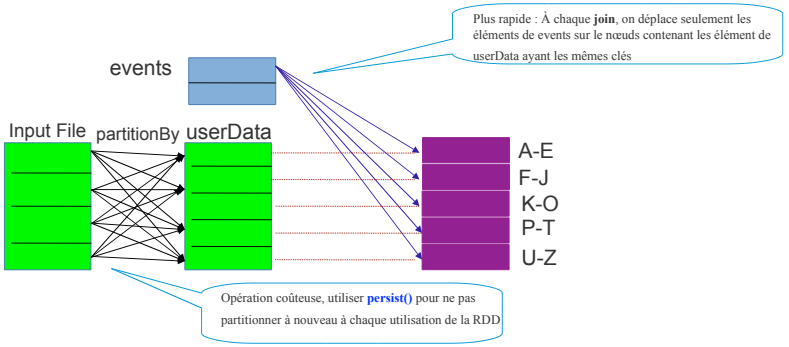
- les entrées de userData sont distribuées en fonction du bloc HDFS où elles ont été trouvées (Spark ne connaît pas l'emplacement de l'entrée correspondante à un certain UserID). Utilisée pour toutes les exécutions de *processNewLogs*, **ne change pas**
- events : RDD locale à *processNewLogs*, utilisée une seule fois, **change** à chaque exécution de *processNewLogs*



Solution : utiliser le partitionnement

- créer 5 partitions pour userData (partitionBy()= transformation, produit toujours une nouvelle RDD)
- pas besoin de partitionner events (locale à processNewLogs(), utilisée une seule fois)

```
val userData = sc.sequenceFile[UserID, UserInfo]("hdfs://...").partitionBy(new HashPartitioner(5)).persist()
```



Connaître le partitionnement existant

On utilise la méthode *.partitioner* sur une RDD

```
scala> val a = sc.parallelize(List((1, 1), (2, 2)))
scala> val b = sc.parallelize(List((1, 1), (2, 2)))
scala> val joined = a.join(b)

scala> a.partitioner
res0: Option[Partitioner] = None

scala> joined.partitioner
res1: Option[Partitioner] = Some(HashPartitioner@286d41c0)
```

Utilisation du partitionnement

Opérations qui utilisent le partitionnement existant

- cogroup, groupWith, join, leftOuterJoin, rightOuterJoin, groupByKey, reduceByKey, combineByKey, and lookup.
- Si l'opération utilise une seule RDD (e.g reduceByKey) partitionnée les valeurs pour chaque clés sont calculées localement, seulement la valeur finale sera envoyée au master
- Une opération binaire sur deux RDD partitionnées, les données d'au moins une des deux ne seront pas envoyées dans le réseau (aucune donnée ne sera envoyée si elles ont le même partitionnement et sont sur les mêmes machines)

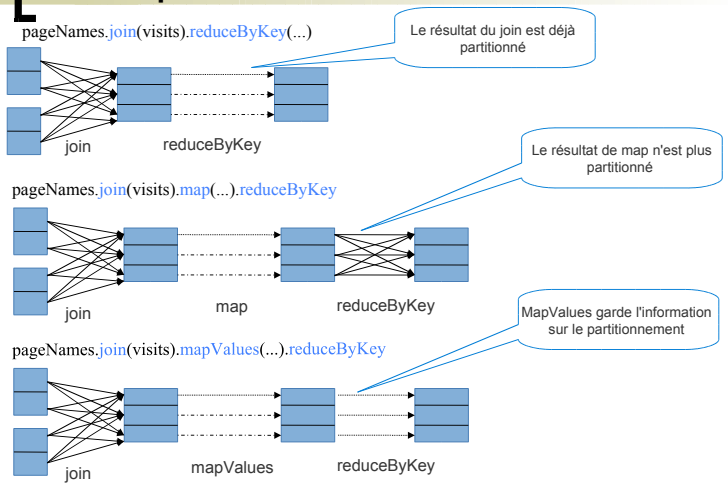
Utilisation du partitionnement

Opérations dont le résultat est partitionné

- cogroup, groupWith, join, leftOuterJoin, rightOuterJoin, groupByKey, reduceByKey, combineByKey, partitionBy, sort (e.g sortByKey produit un partitionnement range, groupByKey un partitionnement par hachage)
- mapValue, flatMapValue, filter (si la RDD à laquelle on les applique a un `Partitioner`)
- L'objet `Partitioner` est créé automatiquement sur les RDD créées par des opérations qui partitionnent les données. Pour les opérations binaires on obtient :
  - le même `Partitioner` que l'un des deux opérandes ayant un `Partitioner`
  - le `Partitioner` du premier opérande
  - si aucun opérande n'est partitionné : `HashPartitioner`

Toute autre opération produit des résultats sans `Partitionner` (e.g `map()`)

Exemples



Définir son propre partitionnement

Extension de la classe *Partitioner,* écrire les méthodes :

- *NumPartitions :* retourne le nombre de partitions
- *getPartition :* retourne la partition d'une clé donnée (entre 0 et NumPartitions)
- *equals :* utile pour décider si deux RDD sont partitionnées de la même manière

Exemple : regrouper les pages du même domaine dans la même partition

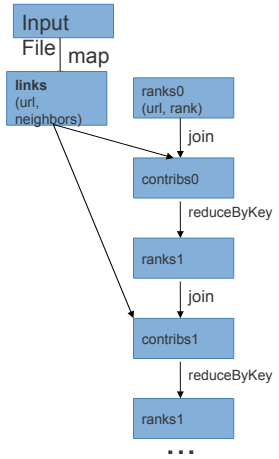
```
class DomainPartitioner(numParts: Int) extends Partitioner {
    override def numPartitions: Int = numParts
    override def getPartition(key: Any): Int = {
            val domain = new java.net.URL(key.toString).getHost()
            val code = (domain.hashCode % numPartitions)
            if (code < 0) {code + numPartitions
            } else {code}
    }
    override def equals(other: Any): Boolean = other match {
                                                case dnp: DomainNamePartitioner
                                                        => dnp.numPartitions == numPartitions
                                                case _ => false  }
}
```

Exemple : PageRank sous Spark

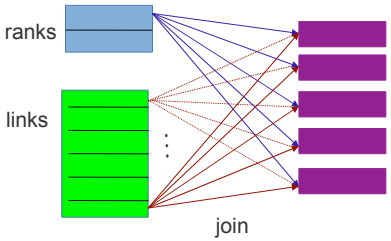
- Le score de chaque page est initialisé à 1
- À chaque itération, une page p envoie le score rank(p)/|numVoisins(p)| à ses voisins
- Le score de chaque page q est calculé comme étant 0.15+0.85*scoresReçus(q)

```
val lines = sc.textFile("links.txt")
val links =lines.map{s=>val parts = s.split(",")
                        (parts(0), parts(1))
     }.distinct().groupByKey().cache() //statique et utilisé à chaque itération
var ranks = links.map(v => (v._1, 1.0))

for (i <- 1 to ITERATIONS) {
        val contribs = links.join(ranks).values.flatMap {
                        case (urls,  rank) =>
                                val size = urls.size
                                urls.map(url => (url, rank/size))
        }
        ranks = contribs.reduceByKey(_ + _).mapValues(.15 + .85*_)//réduit le volume du shuffle
}
```

Exécution sans partitionnement



À chaque join on déplace links et ranks (regroupés en fonction de leurs clés avec HashPartitioner)



Exécution sans partitionnement

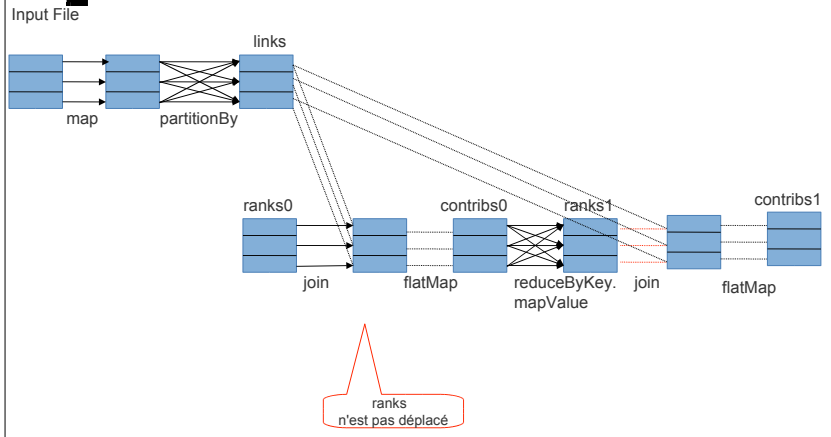
- Shuffles produits par la jointure : links et ranks ont les mêmes clés qui ne changent pas mais pas le même partitionnement

```
val links =lines.map{s=>val parts = s.split(",")
                (parts(0), parts(1))
}.distinct().groupByKey().cache() //statique et utilisé à chaque itération
var ranks = links.map(v => (v._1, 1.0))

ranks.partitioner
res0: Option[Partitioner] = None

links.partitioner
res1: Option[Partitioner] = Some(HashPartitioner@286d41c0)

ranks.partitioner == links.partitioner
False
```

Exécution avec le partitionnement existant dans Spark

```
val links = sc.textFile(...).map(...).....partitionBy(new HashPartitioner(8)).persist()

//links beaucoup plus grand que ranks, ne sera plus envoyé, gardé en RAM pour toutes les itérations

var ranks = links.mapValues(v=>1.0) //utiliser mapValues pour garder le //partitionnement
for (i <- 1 to ITERATIONS) {
        val contribs = links.join(ranks).values.flatMap {
                        case (urls,  rank) =>
                                val size = urls.size
                                urls.map(url => (url, rank/size))
        }

        ranks = contribs.reduceByKey(add, numPartitions=links.getNumPartitions())
                        .mapValues(.15 + .85*_)


    //utiliser mapValues à la place de map, garde le partitionnement de reduceByKey et optimise join()
    suivant

}
```

Nouvelle exécution



Autre optimisation : Variables Broadcast

Les opérations sur les RDD prennent comme arguments des fonctions (fermetures)

→ les fermetures et les variables qu'elles utilisent sont représentées par des objets Java qui sont sérialisés et envoyés avec tes tâches aux workers

Dans certains cas, des variables de grande taille doivent être accessibles et partagées entre plusieurs tâches ou entre plusieurs opérations

=> ***Variables broadcast***

- on garde une copie d'une variable en lecture seule sur chaque machine
- on n'envoie pas une copie de la variable avec chaque tâche
- algorithmes efficaces de distribution des variables broadcast afin de réduire le coût de communication

=> peuvent être utilisées pour donner à chaque nœud une copie des données de grande taille

Exemple

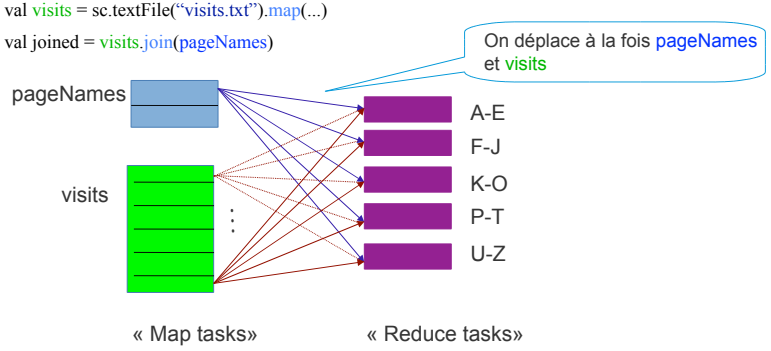
```
scala> val broadcastVar = sc.broadcast(Array(1, 2, 3))
broadcastVar: org.apache.spark.broadcast.Broadcast[Array[Int]] = Broadcast(0)

scala> broadcastVar.value
res0: Array[Int] = Array(1, 2, 3)
```

Variables broadcast

- créées avec SparkContext.broadcast(valeurInitiale)
- *broadcastVar* doit être utilisée à la place de Array(1, 2, 3) (la variable sera envoyée aux nœuds une seule fois).
- accessibles à l'intérieur des tâches avec la méthode *.value* (la première tâche à accéder la variable obtient sa valeur)
- la variable ne doit pas être modifiée après broadcast (la variable serait modifiée sur un seul nœud) afin que tous les nœuds aient la même valeur

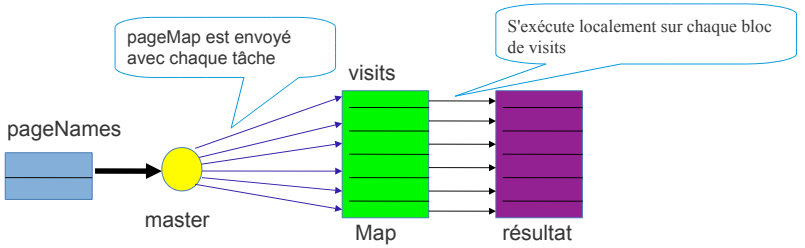
Exemple : Join

```
// Créer une RDD avec des pairs (URL, noms)
val pageNames = sc.textFile("pages.txt").map(...)
// Créer une RDD avec des pairs (URL, visites)
val visits = sc.textFile("visits.txt").map(...)
val joined = visits.join(pageNames)
```



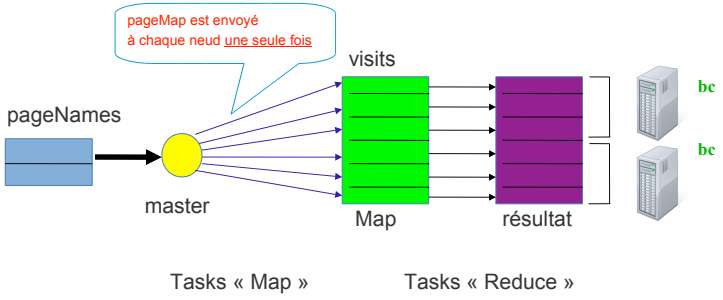
Si l'une des deux tables est de petite taille

```
val pageNames = sc.textFile("pages.txt").map(...)
val pageMap = pageNames.collect().toMap()//les stocker comme un tableau associatif d'objets sur le driver
val visits = sc.textFile("visits.txt").map(...)
val joined = visits.map(v => (v._1, (pageMap(v._1), v._2)))
```



Meilleure version avec broadcast

```
val pageNames = sc.textFile("pages.txt").map(...)
val pageMap = pageNames.collect().toMap()
val bc = sc.broadcast(pageMap)  //est de type Broadcast[Map[...]]
val visits = sc.textFile("visits.txt").map(...)
val joined = visits.map(v => (v._1, (bc.value(v._1), v._2)))
```



HITS

Quelques problèmes liés au score PR

Mesure de popularité générique

- ne tient pas compte de la sémantique des informations associées aux entités classées (documents, utilisateurs ..), le sujet de la requête, les préférences de l'utilisateur → versions de PR améliorées

Une seule mesure d'importance

- autres méthodes : *HITS*

Influencé par des liens de spam

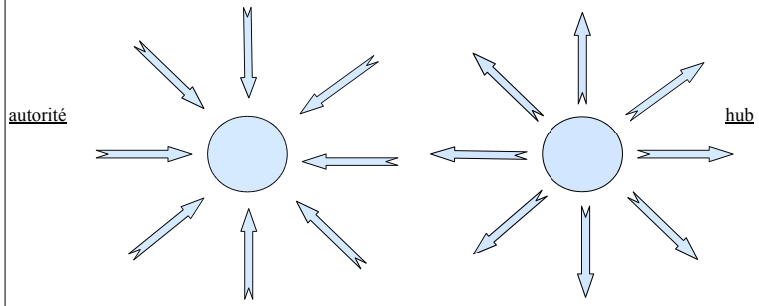
- Liens artificiels crées pour influencer le score PR → solutions pour détecter ces liens et corriger les scores

HITS

- HITS = Hyperlink Induced Topic Search
- Une autre technique d'utilisation du graphe du Web pour le classement
- Part de l'idée qu'il existe 2 rôles pour une page: hub et autorité
- Chaque page peut être un hub, une autorité, ou les 2
- On donne par conséquent 2 scores (scores de hub et d'autorité) à chaque page au lieu d'un seul
- Les scores sont calculés *au moment* de la requête

Quelques définitions

- Une autorité est une page qui fournit des informations importantes et fiables sur un sujet donné (comme dans PageRank)
- Un hub est une page qui contient une collection de liens vers des autorités sur un sujet donné



Idée de l'algorithme

Les bonnes pages de hubs ont des liens vers les pages d'autorité les plus intéressantes

(ex.: le fan de Star Wars aura de nombreux liens vers des pages sur des sites importants dédiés au film)

Les pages d'autorité importantes sont pointées par des hubs importants

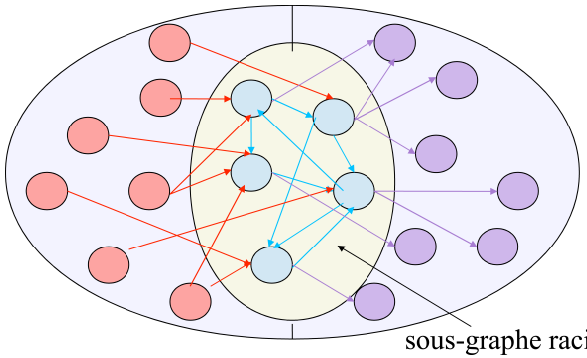
(ex: le site officiel de l'université est « pointé » par les sites des différents masters et des universités collaborant)

=> relation de renforcement mutuel

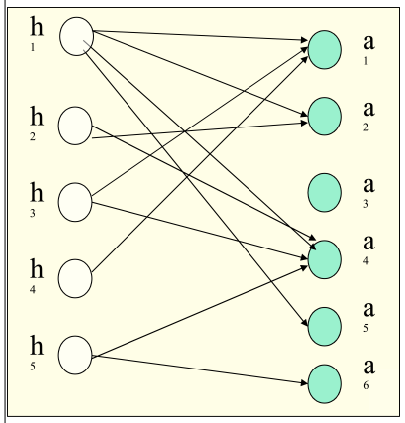
Trouver les autorités et hubs

on construit d'abord un sous-graphe du web centré sur le sujet souhaité (en fonction de la requête)

ensuite on calcule les scores de hubs et d'autorités dans ce sous-graphe



Exemple



Score de hub:
h1 = a1 + a2 + a4 +a5
h4 = a1
donc h1 meilleur hub

Score d'autorité:
a2 = h1 + h2
a4 = h1 + h2 + h3 + h5
donc a4 meilleure autorité

Calcul des scores

Algorithme itératif :

- pour chaque page p on maintient un score de hub h(p) et un score d'autorité a(p)
- à chaque itération on calcule d'abord les scores d'autorité à partir des scores de hub précédents:

$$a(p)^k = \sum_{q \text{ pointe vers } p} h(q)^{k-1}$$

- avec ce nouveau score on calcule les scores de hub :

$$h(p)^k = \sum_{p \text{ pointe vers } q} a(q)^k$$

Calcul des scores (suite)

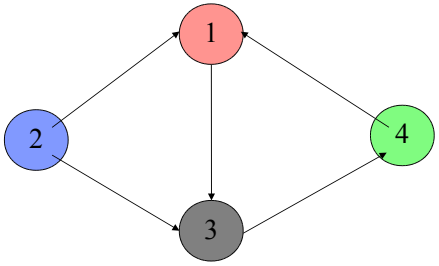
- à chaque itération, lorsque tous les scores ont été calculés, on les normalise (on peut utiliser n'importe quelle norme, on veut les valeurs relatives):

$$a(p)^k = \frac{a(p)^k}{\sqrt{\sum (a(p')^k)^2}} \qquad h(p)^k = \frac{h(p)^k}{\sqrt{\sum (h(p')^k)^2}}$$

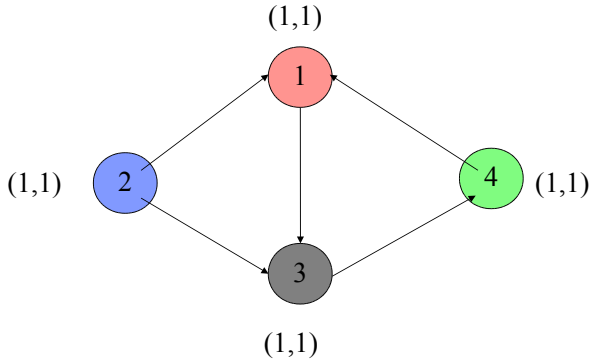
- test de convergence pour un seuil de tolérance ε :

$$\sum \left(h_i^k - h_i^{(k-1)}\right)^2 < \varepsilon \qquad \sum \left(a_i^k - a_i^{(k-1)}\right)^2 < \varepsilon$$

Exemple



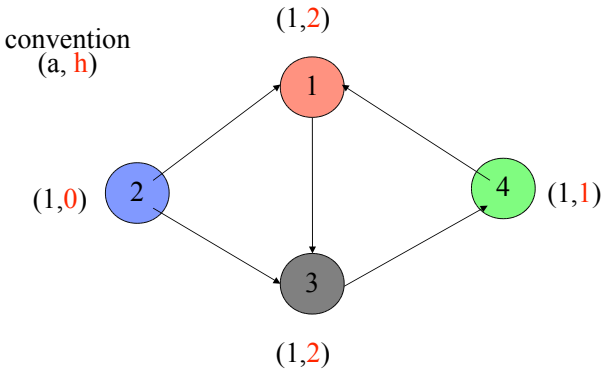
Exemple



Initialisation

Exemple

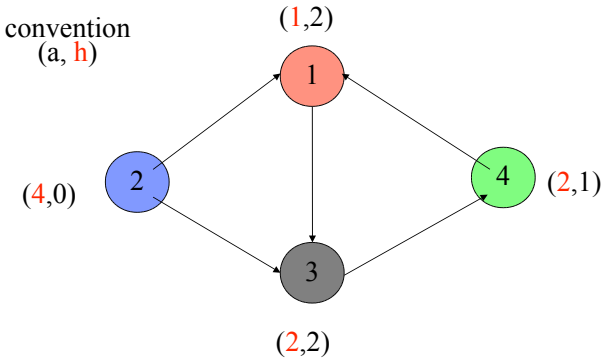
convention
(a, h)



Calcul des scores d'autorité

Exemple

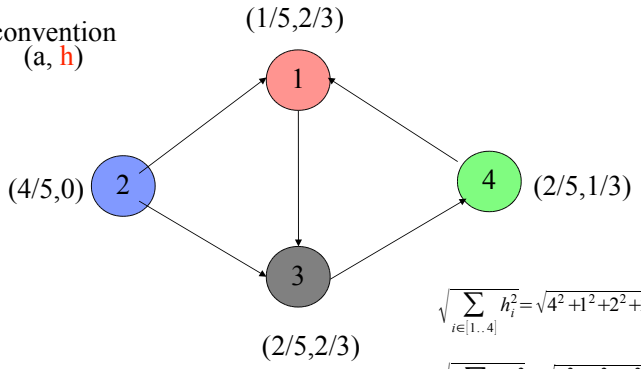
convention
(a, h)



Calcul des scores de hub

Exemple

convention
(a, h)



$$\sqrt{\sum_{i\in[1..4]} h_i^2}=\sqrt{4^2+1^2+2^2+2^2}=\sqrt{25}=5$$

$$\sqrt{\sum_{i\in[1..4]} a_i^2}=\sqrt{0^2+2^2+2^2+1^2}=\sqrt{9}=3$$

Normalisation des scores

Algorithme séquentiel

$a^0:=\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)\in\mathbb{R}^n$

$h^0=\left(\frac{1}{\sqrt{N}},\ldots,\frac{1}{\sqrt{N}}\right)\in\mathbb{R}^n$

$k:=1$

do $\forall\ p$:

$$a(p)^k=\sum_{q \text{ pointe vers } p} h(q)^{k-1}$$

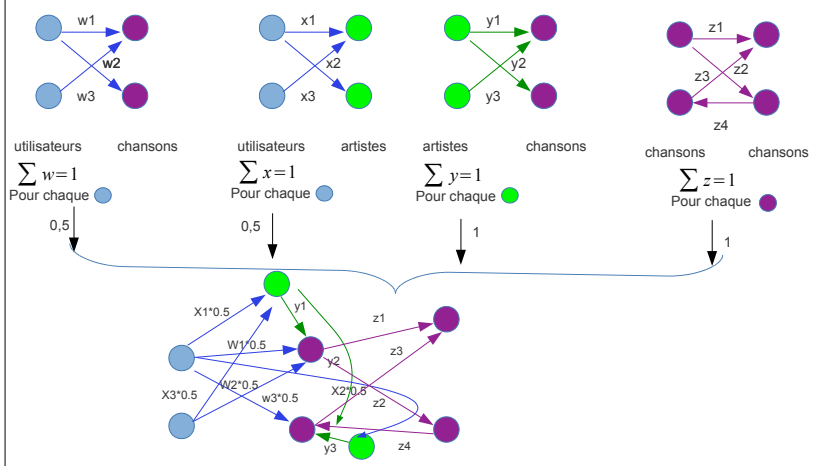
$$h(p)^k=\sum_{p \text{ pointe vers } q} a(q)^k$$

$$a(p)^k=\frac{a(p)^k}{\sqrt{\sum\left(a(p'_k)^k\right)^2}}$$

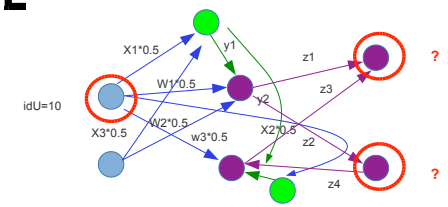
$$h(p)^k=\frac{h(p)^k}{\sqrt{\sum\left(h(p')^k\right)^2}}$$

while $\sum\left(h_i^k-h_i^{(k-1)}\right)^2>\varepsilon$

TME: Graphe utilisateurs, artistes, chansons



TME: PPR



Itération k : $si\ i=10 : PR_i^k = d * \sum PR_j * poids_{j \to i} + (1-d)$

$si\ i \neq 10 : PR_i^k = d * \sum PR_j * poids_{j \to i}$

Initialisation : $PR_{10}^0 = 1, PR_i^0 = 0\ si\ i \neq 10$

Références

https://www.safaribooksonline.com/library/view/learning-spark/9781449359034/ach04.html

http://ampcamp.berkeley.edu/wp-content/uploads/2012/06/matei-zaharia-amp-camp-2012-advanced-spark.pdf

https://www.cs.berkeley.edu/~matei/papers/2012/nsdi_spark.pdf

Mining of Massive Datasets (Chapitre 5) : http://infolab.stanford.edu/~ullman/mmds/bookL.pdf