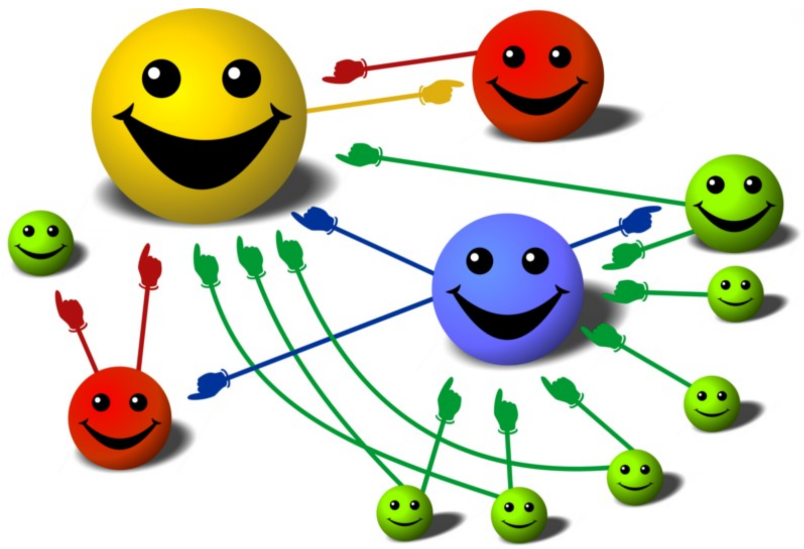
TIPE

Fiches rétroprojecteur

Le fonctionnement de Google

Etude de la structure du World Wide Web



Plan de la présentation

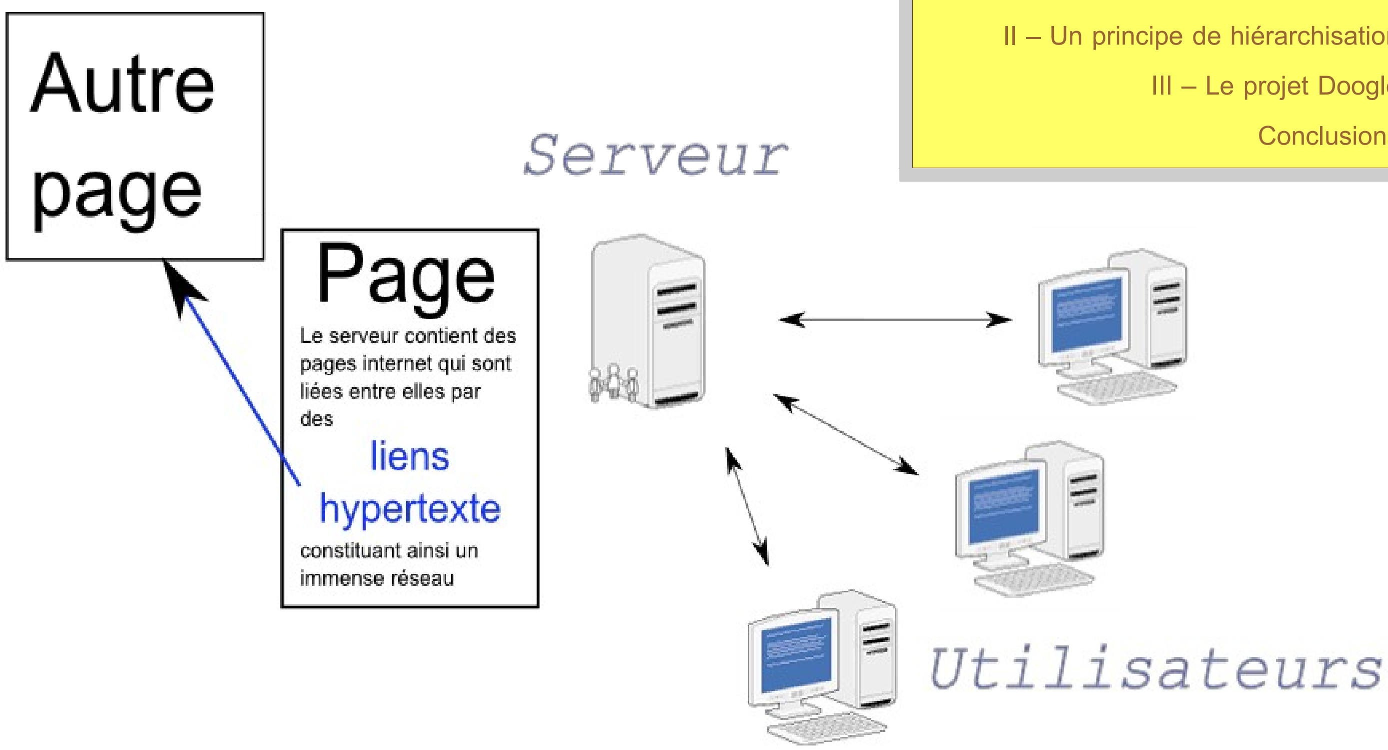
Introduction

I - Structure du World Wide Web

II - Un principe de hiérarchisation

III - Le projet Doogle

Conclusion



<u>Un immense réseau :</u> plus de 1000 milliards de pages

<u>Risques d'une requête aléatoire :</u>

- ✗ **Incohérence** et **impertinence** des résultats.
- ✗ **Mauvaise qualité** des sites obtenus (incomplets ou erronés).
- ✗ Sites **frauduleux** ou illégaux.
- ✗ **Temps de recherche** de l'utilisateur dans les résultats considérables.

Architecture des réseaux « Scale-Free »

Réseau : Ensemble de points (**nœuds**) connectés entre-eux par des **liens**.

Dans notre étude, un nœud représente une page, et un lien représente un lien hypertexte.

Degré d'un nœud : Nombre de liens partant du noeud.

Distribution de degré :
Probabilité pour un nœud choisi au hasard d'être de degré *k*.

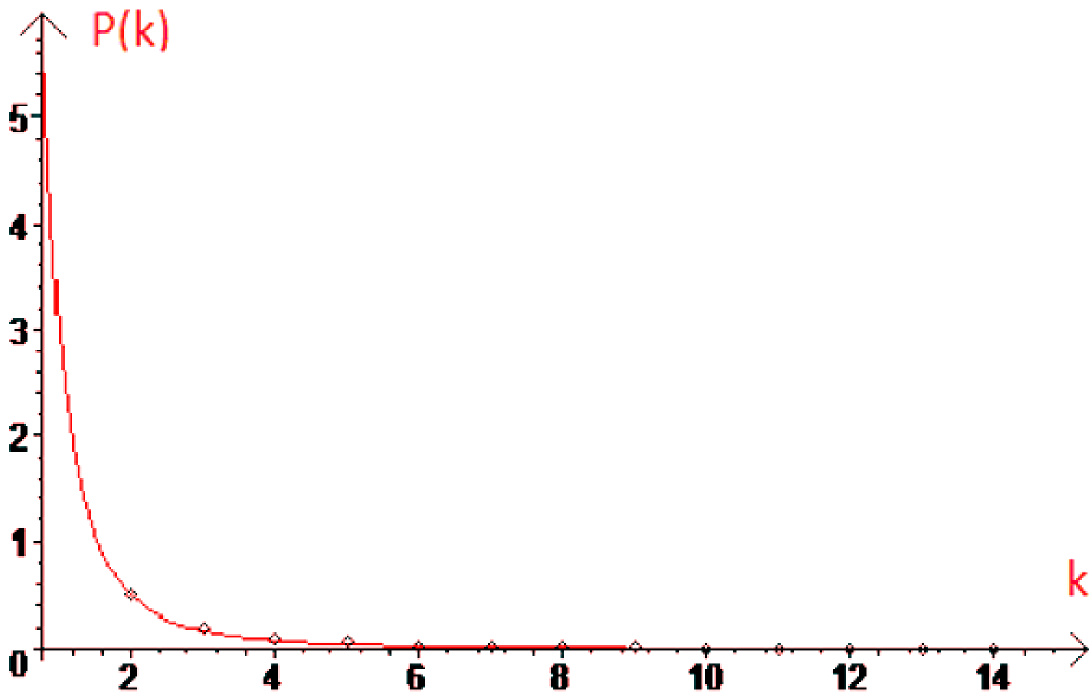
$$P(k) = \frac{Nk}{N}$$

Réseau de type « Scale-Free »

Sa distribution de degré répond à une loi de puissance :

$$P(k) = \frac{a}{k^{\alpha}}$$

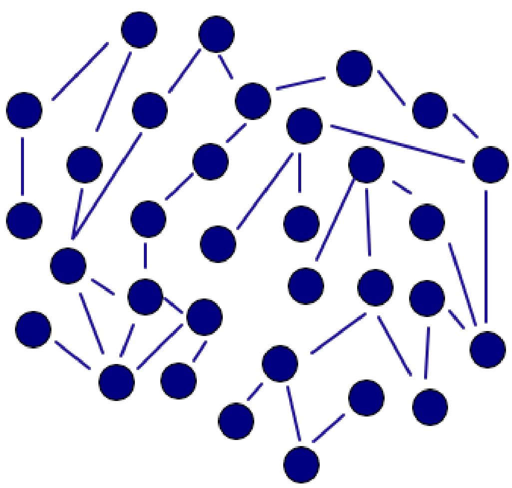
DISTRIBUTION DE DEGRÉ D'UN RÉSEAU SIMULÉ :



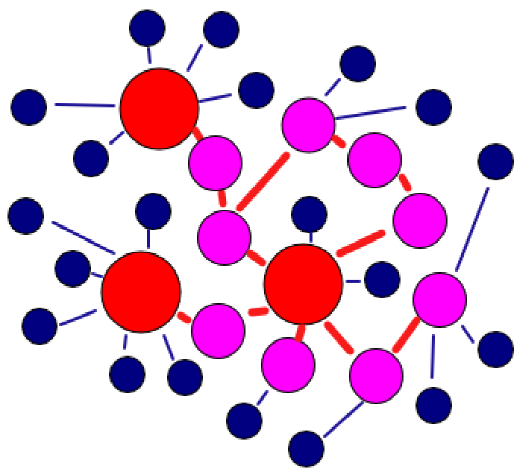
Application à des réseaux variés :

Réseau	Nœuds	Relations
Structure virtuelle d'internet	Pages web	Liens hypertextes
Publications scientifiques	Chercheurs	Citations
Industries	Firmes	Partenariats

De tels réseaux vérifient la présence de **hubs** (ou *moyeux*), des nœuds qui sont au cœur du réseau car ils concentrent les liens.

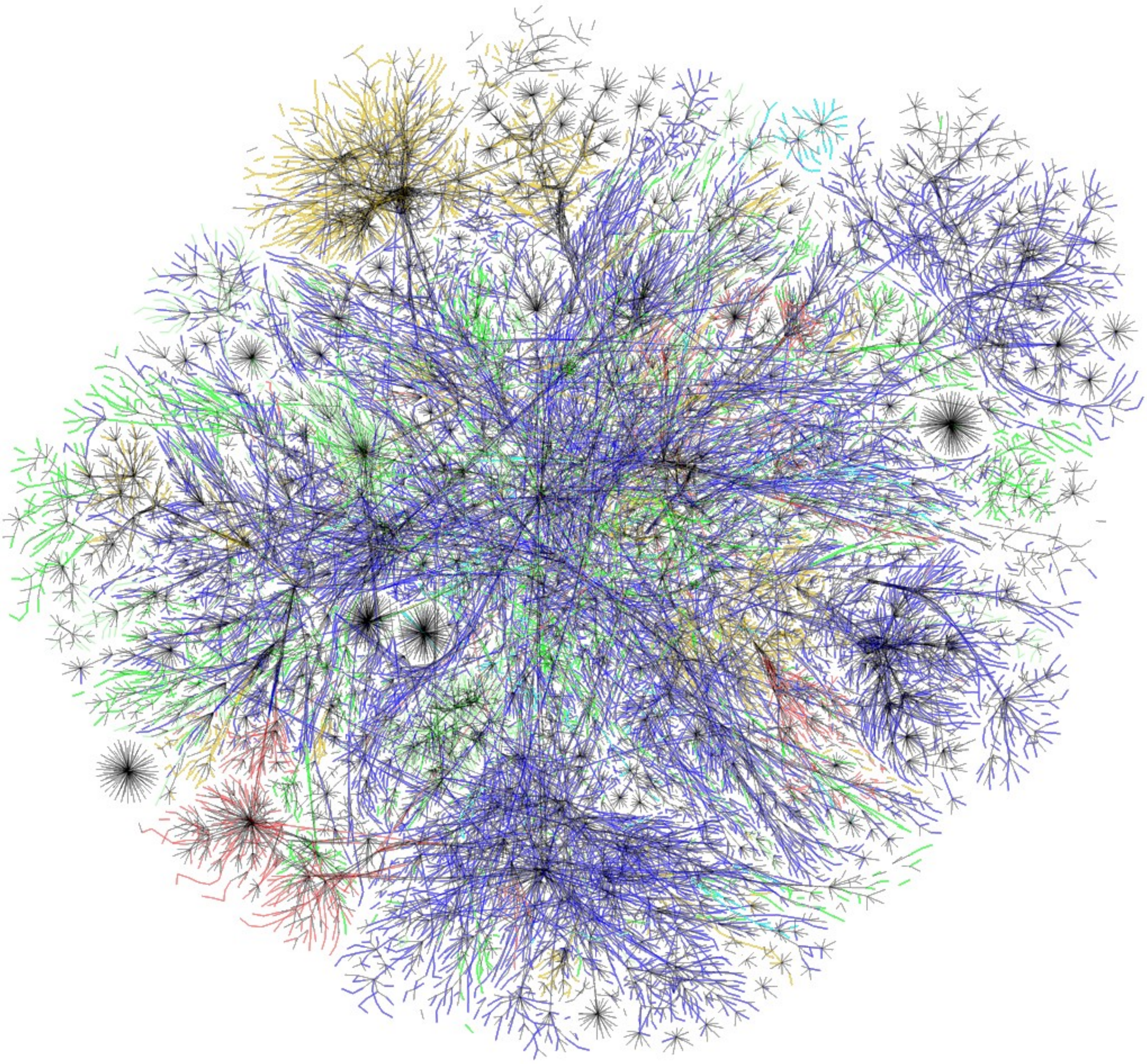


Réseau aléatoire



Réseau type "Scale-Free"

Carte du World Wide Web
(projet Opte)



Légende : Extensions des sites web (liée à leur position géographique) :

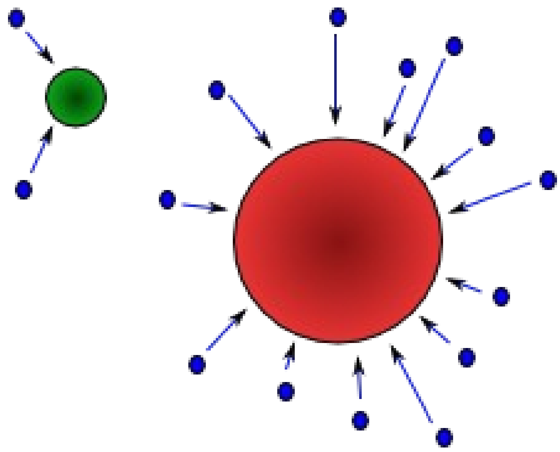
Bleu : .net, .ca, .us (Etats-Unis)
Vert : .com, .org (Internationnaux)
Jaune : .jp, .cn, .tw, .au (Asiatiques)
Rouge : .br, .kr, .nl (Brésil, Corée du sud, Pays-bas)
Noir : autres

Source : Projet Opte (2005)

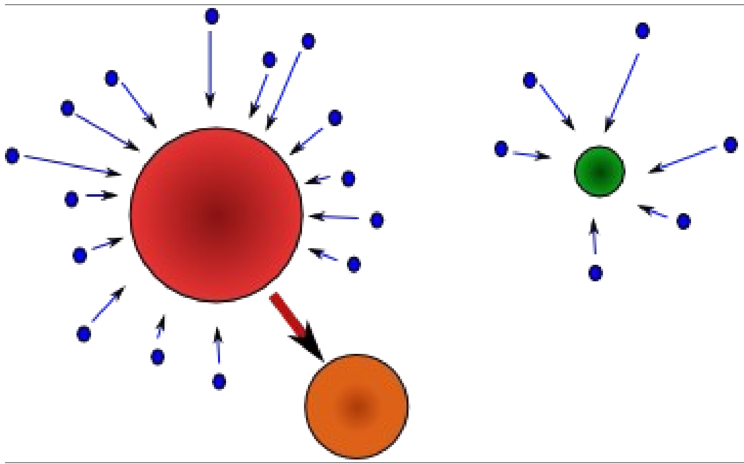
Le « Pagerank »

S'inspirant de l'idée de Google, nous avons utilisé un ***indicateur de confiance*** des pages web :

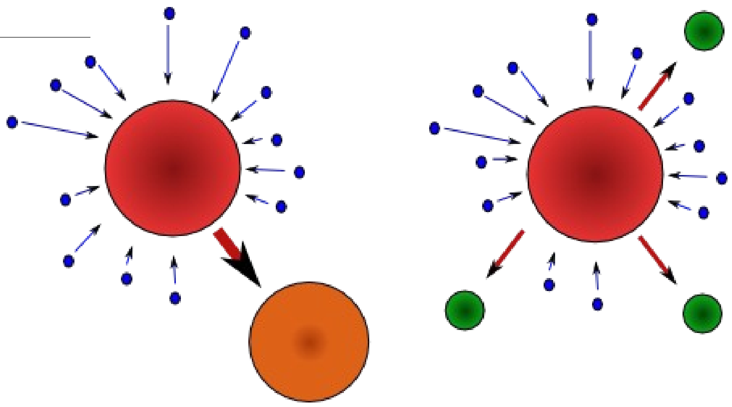
le Pagerank, proportionnel :



✔ **Au nombre de liens pointant vers le site (citations).**



✔ **Au Pagerank du site d'où vient le lien.**



✔ **A l'exclusivité du lien.**

D'où la formule mathématique :

$$P_a = [1 - c] + c\left(\frac{P_1}{\text{degré}(1)} + \frac{P_2}{\text{degré}(2)} + \dots + \frac{P_q}{\text{degré}(q)}\right)$$

Cet algorithme itératif converge vers une unique solution.

Le projet Doogle

Fonctionnement du moteur de recherche

● Recensement des pages du réseau.

Dooglebot : parcours de chaque page pour y repérer les liens menant vers d'autres pages internet.
C'est **l'indexation**.

● Calcul du Pagerank des pages.

Dooglerank : algorithme itératif.
C'est **l'index software**.

● Interface utilisateur.

La page internet sur laquelle vous arrivez en entrant http://www.google.fr dans votre navigateur internet. C'est le **query software** (engin de requête).

Base de données

C'est une structure permettant le **stockage d'information**, sous forme de tableaux composés de différentes colonnes.

Sites						Liens	
ID	URL	Titre	Pagerank	Keywords	Scan	Entrée	Sortie
1	.../index.html	Accueil	0.15	Fourier, ...	1	1	2
2	.../coursproba.html	Aucun	0.16		1	2	3
...	...	...	...	...	...	...	...

Ma base de données

● La table site

Liste des pages web contenant leurs informations (id, titre, adresse, pagerank...).

● La table liens

Liens entre les différentes pages (id du site entrant, id du site cible).

● La table mots

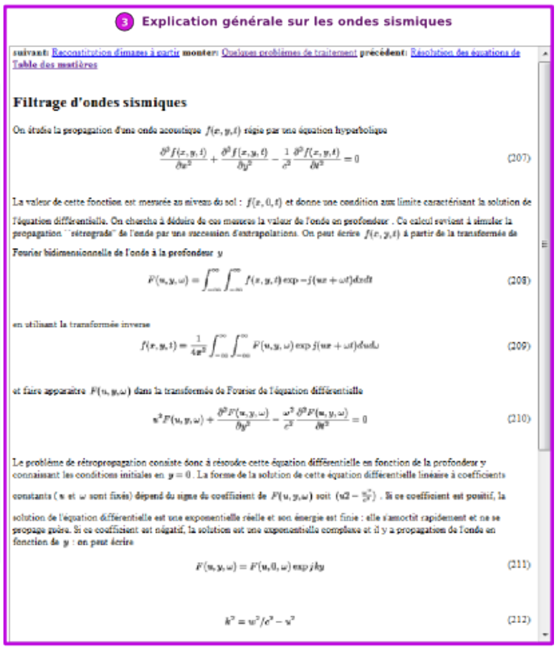
Mots de taille significative apparaissant sur une page.

Avec un **aspirateur de sites** j'ai téléchargé en réseau local près de 700 pages.

Doogle | ondes | Recherche Doogle

1. Table des matières
2. Quelques problèmes de traitement de signaux multidimensionnels
3. Filtrage d'ondes sismiques

Transformée de Fourier et propagation d'ondes en optique cohérente ou en électromagnétisme
Résolution des équations de Maxwell utilisant la transformée de Fourier en compatibilté avec le théorème d'échantillonnage
Reconstitution d'images à partir de projections
presentationfourier
Schéma d'un système de transmission
La modulation des signaux
Les bandes de fréquences
Les supports de transmission
Le câble coaxial
Ondes herziennes
Liaisons satellites



<u>Les résultats sont beaucoup plus pratiques pour l'utilisateur :</u>

1-> Sommaires très généraux
2-> Sommaires spécialisés
3-> Pages populaires

Web [Images] [Maps] [Actualités] [Vidéo] [Gmail] [plus ▼]

racines site:www.polytech.unice.fr/~leroux/ | Rechercher | Recherche avancée | Préférences

Rechercher dans : ◉ Web ○ Pages francophones ○ Pages : France

Web Résultats **1** - **32** sur **32** provenant de **www.polytech.unice.fr/~leroux** pour **racines**. (**0,13** secondes)

1 [Interprétation en termes de lieu des **racines**]
Ce lieu des **racine** se présente sous la forme de une ou plusieurs boucles fermées ou **...**
Figure 26: Quatre configurations de lieux des **racines**: le filtre est **...**
www.polytech.unice.fr/~leroux/crim2/node70.html - 9k - En cache - Pages similaires

[Analyse en fréquence d'un filtre non récursif du deuxième ordre]
Nous prendrons un exemple où les **racines** du polynôme $B(z)$ **...** L'atténuation est d'autant plus importante que les **racines** sont proches du cercle de rayon **...**
www.polytech.unice.fr/~leroux/courssignal/node55.html - 6k - En cache - Pages similaires

. . . .

[Diapositive 1]
Minimum de phase : **racines** de z2B(z) situées à l'intérieur. du cercle unité. Dé phasage nul : **racines** par quadruplets (G à coefficients réels). H(z)=G(z). **...**
www.polytech.unice.fr/~leroux/DIAPOS%20COURS%20SIGNAL/
diaposCours%20SignalChap5.../slide000... - 17k - En cache - Pages similaires

2 [Stabilité des filtres causaux]
Sous-sections. Le théorème de Rudin et ses corollaires · Interprétation en termes de lieu des **racines** · Stabilisation d'un filtre récursif instable **...**
www.polytech.unice.fr/~leroux/crim2/node68.html - 4k - En cache - Pages similaires

3 [Filtrage des signaux bidimensionnels]
Le théorème de Rudin et ses corollaires · Interprétation en termes de lieu des **racines** · Stabilisation d'un filtre récursif instable **...**
www.polytech.unice.fr/~leroux/crim2/node56.html - 6k - En cache - Pages similaires

racines site:www.polytech.unice.fr/~leroux/ | Rechercher

Rechercher dans ces résultats | Outils linguistiques | Conseils de recherche

Accueil Google - Programmes de publicité - Solutions d'entreprise - Confidentialité - À propos de Google

---

racines | Recherche Doogle

1 [Interprétation en termes de lieu des racines]
2 [Stabilité des filtres causaux]
3 [Filtrage des signaux bidimensionnels]

Temps moyen pour l'analyse d'une page	
pas de table MOTS	**avec table MOTS**
0,203 secondes	0,603 secondes

Temps moyen d'une requête	
pas de table MOTS	**avec table MOTS**
0,688 secondes par résultat Recherche sans résultats 43 secondes	quasi instantané Recherche sans résultats 0 seconde mesurée

Temps moyen d'une itération	
requêtes SQL	**dans un tableau**
7,286 secondes	5,286 secondes

Améliorations

Filtres sémantiques

Le Blockrank

Estimation du Pagerank sur des blocs de pages.

L'algorithme HITS

– Sites classiques
– Sites de **référence** (« authorities ») : spécialistes.
– Sites **moyeux** (« hubs ») : annuaires de liens.

Le Pagerank thématique

Sites de **référence** sélectionnés comme base pour un **calcul du Pagerank en fonction de chaque requête**.

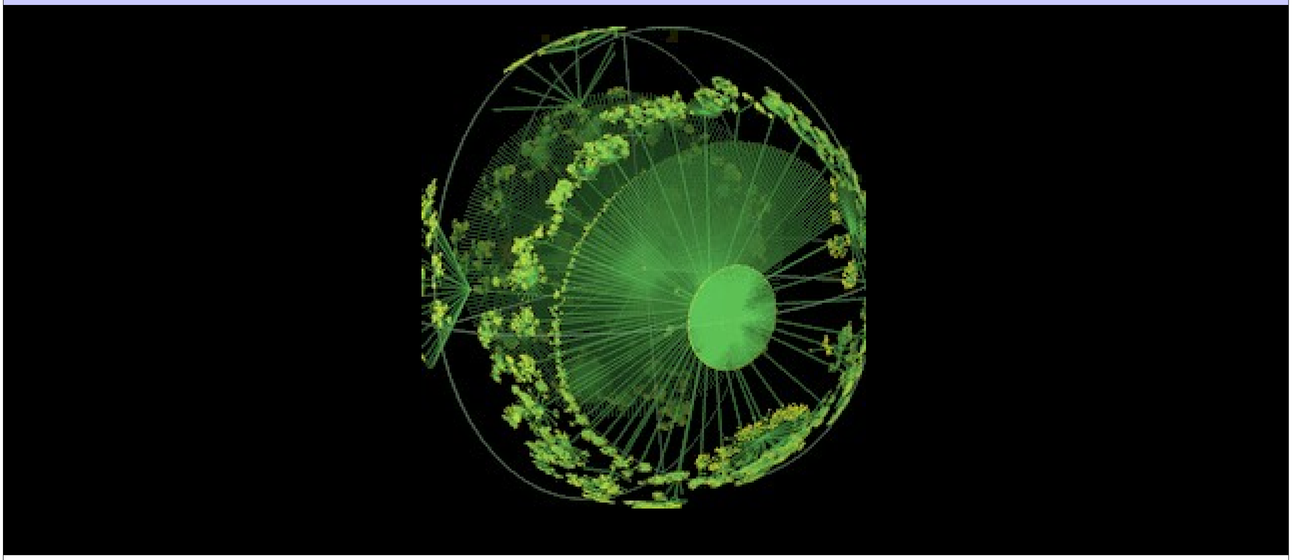
A l'avenir, les moteurs de recherche devront prendre en compte :

- sens des mots
- contexte de leur apparition
- circonstances de la recherche
- personnalité de l'utilisateur

TIPE

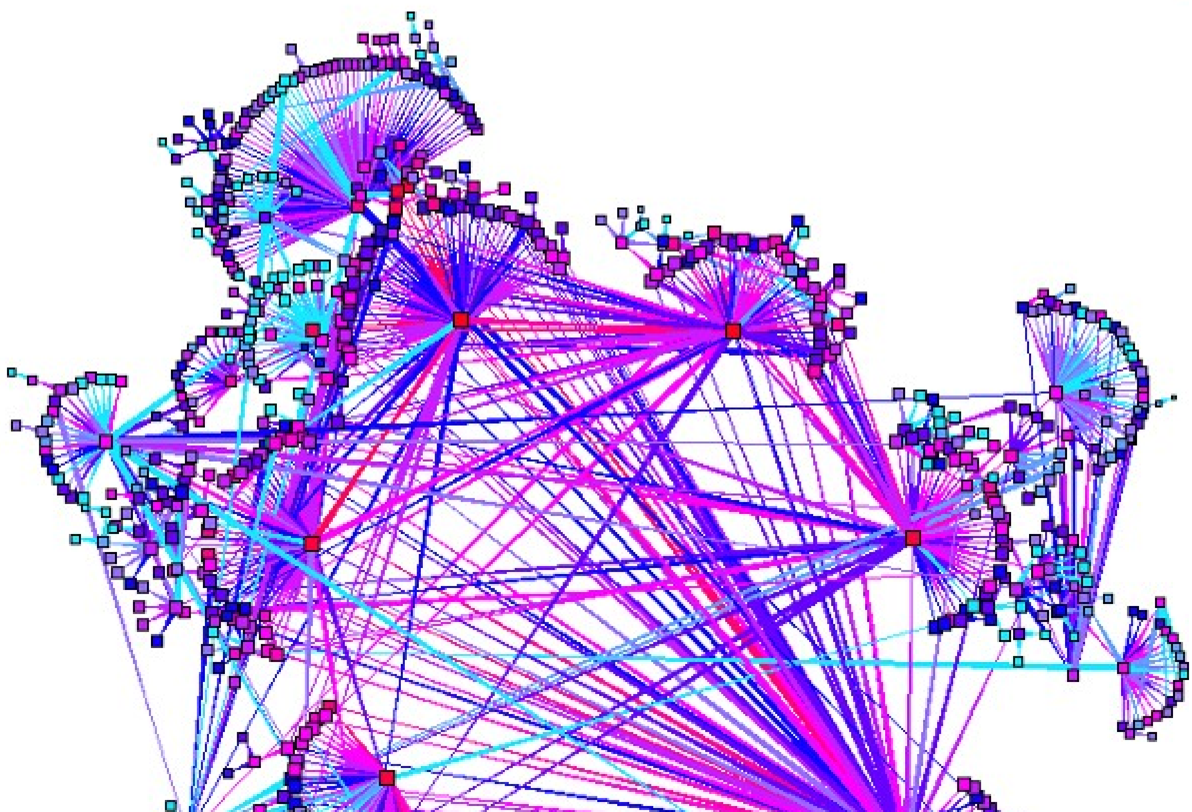
Fiches rétroprojecteur supplémentaires

World Wide Web en représentation tridimentionelle



Source : CAIDA, données de Skitter visualisées par le logiciel Walrus

World Wide Web en représentation hierarchisée par Plankton

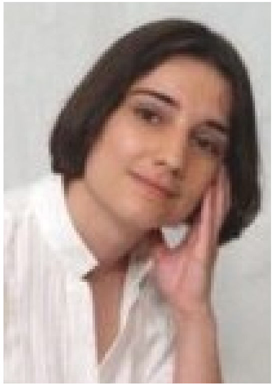


Source : CAIDA, données de Skitter visualisées par le logiciel Plankton

Architecture des réseaux « Scale-Free »

Albert-László Barabási
1967-
scientifique hongrois

J'ai conçu un algorithme de génération de réseaux aléatoires de type Scale-Free, d'après le modèle d'**attachement préférentiel**.



Réka Albert

Un nouveau noeud aura tendance à se lier aux nœuds de haut degrés (hubs).

Réseau initial de m_0 *nœuds* (m_0 >2) <u>non connectés entre eux initialement</u>

Chaque nœud ajouté se lie à un nœud existant de degré k avec la probabilité :

$$p = \frac{k}{\sum_j k_j}$$

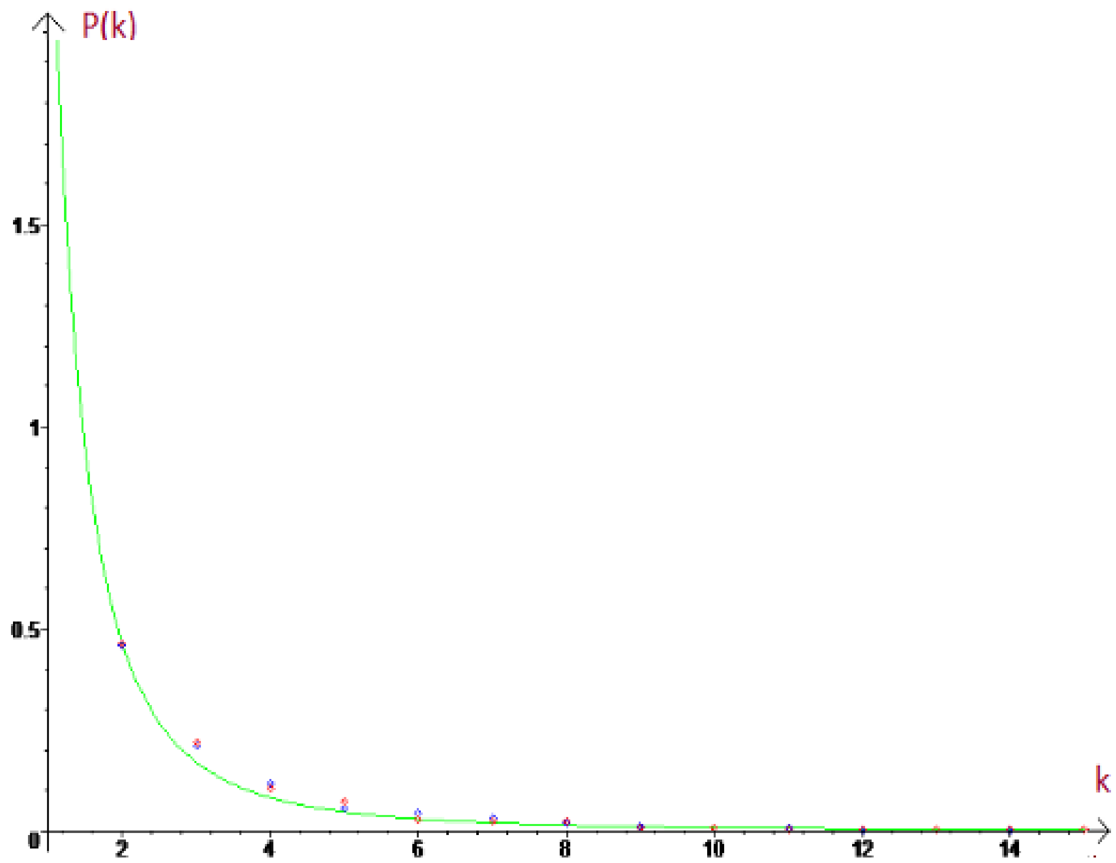
Exemple de réseau créé (1000 noeuds)

$$entrant := k \rightarrow ea\, k^{eb}$$

$$sortant := k \rightarrow sa\, k^{sb}$$

$$\{ ea = 2.527053043 \,,\, eb = -2.448371849 \}$$

$$\{ sb = -2.518864797 \,,\, sa = 2.642133663 \}$$



<u>Degré maximal</u> : entre 40 et 80 suivant les exemples.

Ils vérifient de plus la **théorie des 6 degrés**

Une telle modélisation peut être utile pour utiliser divers phénomènes sur les réseaux sociaux (virus sur internet, épidémies humaines)...

Convergence du Pagerank

L'algorithme itératif considère un internaute fictif parcourant le web.

$$P_a = [\,1 - c\,] + c\left(\frac{P_1}{\text{degré}(1)} + \frac{P_2}{\text{degré}(2)} + \;\cdots\; + \frac{P_q}{\text{degré}(q)}\right)$$

Matriciellement : $\mathrm{T}(x) = c\,\varepsilon + (1 - c)\,A\,x$

- ***x*** : vecteur dont chaque coordonnée représente la probabilité de présence sur une page (*son pagerank*).

- ***c*** : probabilité de changer de page.

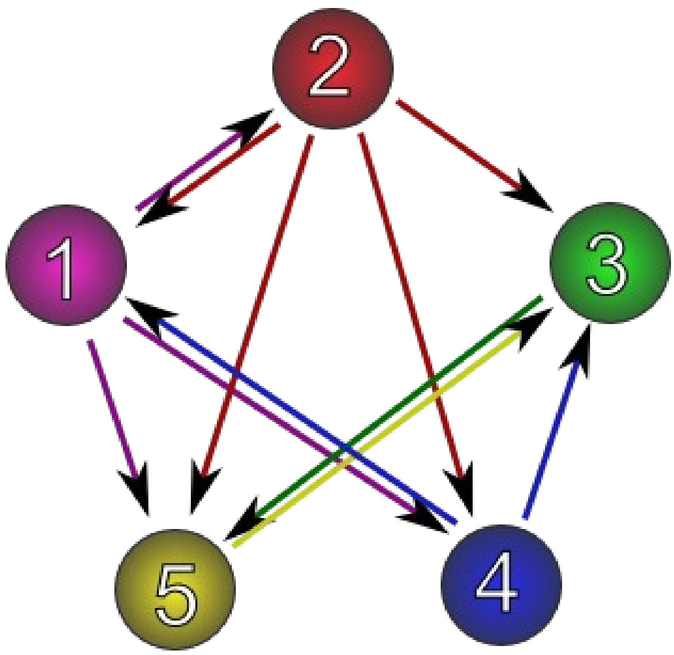
- ***A*** : matrice contenant les liens entre les sites web

  si le site *j* présente un lien vers le site *i* : $A_{i,j} = \dfrac{1}{\text{degré}(i)}$

  sinon : $A_{i,j} = 0$

Exemple :

$$\begin{array}{rc}
\text{de} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{r} \text{vers } 1 \\ \text{vers } 2 \\ \text{vers } 3 \\ \text{vers } 4 \\ \text{vers } 5 \end{array} &
\left[\begin{array}{ccccc}
0 & \frac{1}{4} & 0 & \frac{1}{2} & 0 \\
\frac{1}{3} & 0 & 0 & 0 & 0 \\
0 & \frac{1}{4} & 0 & \frac{1}{2} & 1 \\
\frac{1}{3} & \frac{1}{4} & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{4} & 1 & 0 & 0
\end{array}\right]
\end{array}$$



Il s'agit de montrer que l'application $\mathrm{T}(x) = c\,\varepsilon + (1 - c)\,A\,x$ *de* R^n *dans* R^n *est une application contractante, avec* **$c \approx 0{,}25$**

$$z = \mathrm{T}(x) - \mathrm{T}(y) = (1 - c)\,A\,(x - y)$$

Il aura pour coordonnées :

$$z_i = \sum_{k=0}^{n} (1 - c)\,A_{i,k}\,(x_k - y_k)$$

D'où les inégalités :

$$\| z \|_1 = \sum_{i=0}^{n} |z_i| \leq \sum_{i=0}^{n} \left(\sum_{k=0}^{n} (1 - c)\,|A_{i,k}|\,|x_k - y_k| \right)$$

Or $\sum_{i=0}^{n} |A_{i,k}| = 1$ d'où $\| z \|_1 \leq (1 - c) \sum_{k=0}^{n} |x_k - y_k|$

Ainsi $$\| \mathrm{T}(x) - \mathrm{T}(y) \|_1 \leq (1 - c)\, \| x - y \|_1$$

avec **$1 - c < 1$**.

L'application T est donc **contractante**, d'où la convergence du processus itératif.

Doogle Bot

Classes d'objets

page

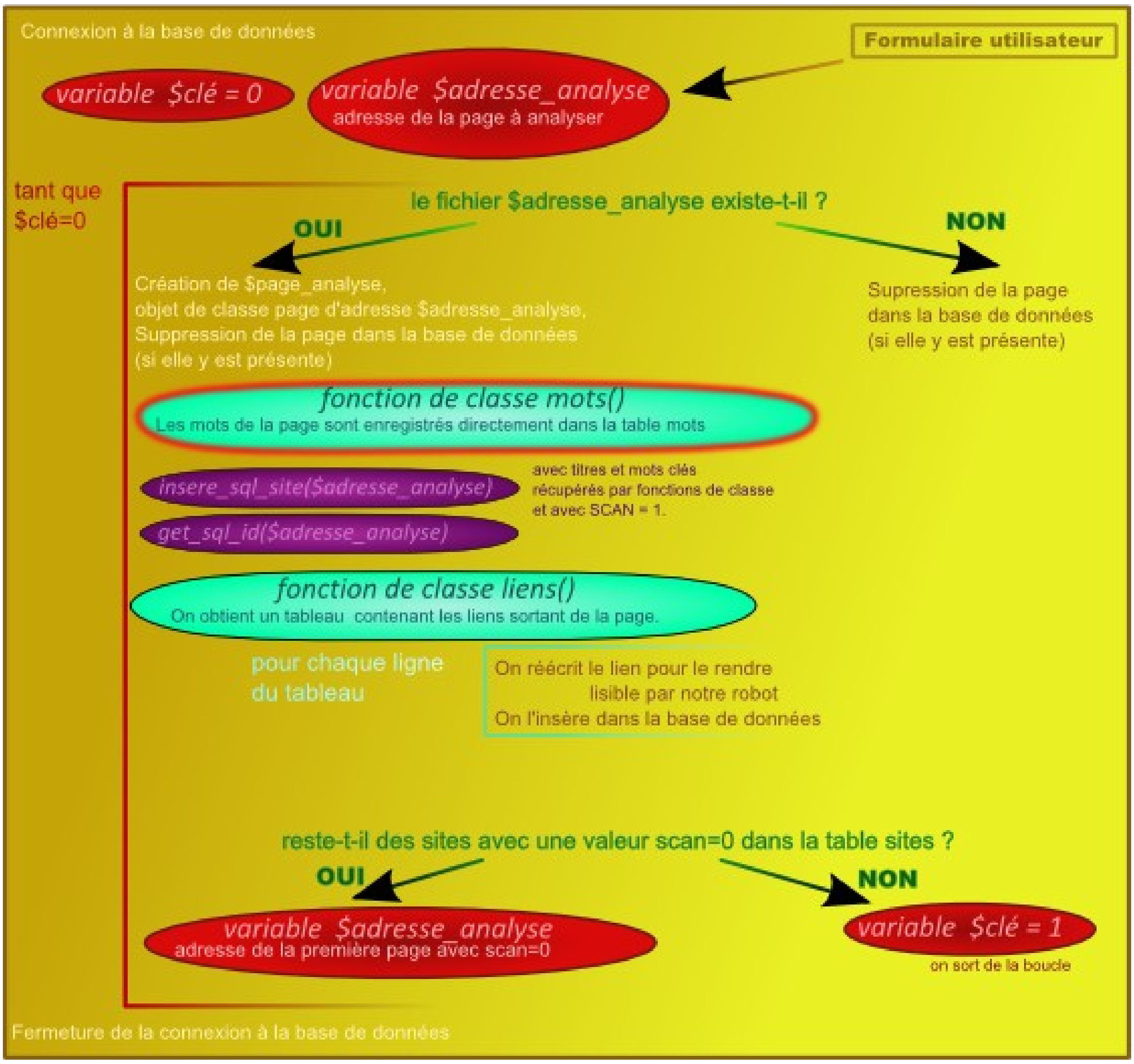
variable $adresse

- *fonction titre()* — Récupère le titre de la page.
- *fonction keywords()* — Récupère les mots clefs de la page.
- *fonction liens()* — Récupère les liens sortant de la page.
- *fonction mots()* — Parcours la page, recense les mots de plus de 5 lettres, et les stocke dans la base de données

Fonctions

- *fonction insere_sql_site()* — Ajoute un site dans la base de données
- *fonction insere_sql_liens()* — Ajoute un lien dans la base de données
- *fonction get_id_sql()* — Récupère l'identifiant d'un site dans la base de données.

Programme



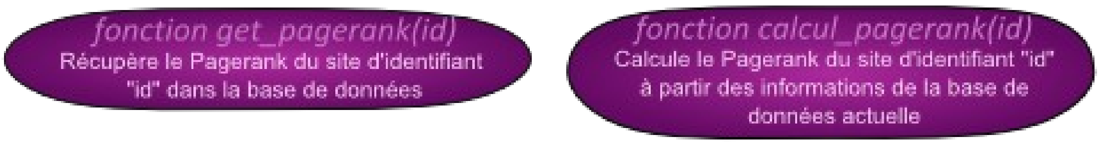
Doogle Search

Programme

Doogle Rank

Ancienne version

Fonctions



Programme

Doogle Rank

Nouvelle version

Variables globales

global $objectif
id de la page en cours d'analyse, necessaire pour le filtrage

global $liste_sites
Tableau contenant la liste des sites de la forme analogue à la table de la base de données

global $liste_liens
Tableau contenant la liste des liens de la forme analogue à la table de la base de données

Fonctions de filtrage

fonction sous_filtre(variable)
Retourne vrai si variable est égale à la variable globale $objectif

fonction filtre(variable)
Utilise la fonction sous_filtre et renvoie vrai si le tableau envoyé en variable contient au moins une occurence de la variable globale $objectif.

Fonctions

fonction calcul_pagerank(id)
Calcule le Pagerank du site d'identifiant "id" à partir des informations des tableaux et de la fonction filtre qui détermine la liste des sites ayant des liens vers "id"

fonction get_pagerank(id)
Récupère le Pagerank du site d'identifiant "id" dans le tableau de sites

Programme

Déclenchement par l'utilisateur

Connexion à la base de données

Une requête SQL permet d'initialiser les tableaux :

Pour chaque site — Insertion dans $liste_sites

Pour chaque liens — Insertion dans $liste_liens

variable $op = 0 — ancien Pagerank du premier site

variable $np = 1 — ancien Pagerank du premier site

Tant que $op ≠ $np

- $op ← $np
- $np ← calcul_pagerank (premier site)
- Pour chaque site: $liste_sites est mis à jour par calcul_pagerank (site)

On met à jour la base de données avec le tableau $liste_sites

Fermeture de la connexion à la base de données

TIPE

Fiches polycopiées

Algorithme itératif :

→ ajoute à chaque tour un nœud de degré ***d*** fixe par attachement préférentiel.

Au bout de ***t*** tours, on aura le nombre total de liens dans le réseau :

$$\sum_j k_j = k_t = 2 \cdot d \cdot t$$ car chaque lien a deux extrémités

Considérant les variations de degré comme continues :

$$\frac{\delta k_i}{\delta t} = d \cdot p(k_i) = d \cdot \frac{k_i}{\sum_j k_j} = d \cdot \frac{k_i}{2 \cdot d \cdot t} = \frac{k_i}{2 \cdot t}$$

$$\frac{\delta k_i}{k_i} = \frac{\delta t}{2 \cdot t}$$

$$\ln(k_i) = \frac{1}{2} \cdot \ln(t) + cste$$

Ce nœud a été ajouté à un temps t_i avec un degré ***d*** : $$k_i(t) = d\sqrt{\frac{t}{t_i}}$$

$$k < k_i(t) \iff k < d\sqrt{\frac{t}{t_i}}$$

$$\iff \frac{k^2}{d^2} < \frac{t}{t_i} \iff t_i > \frac{d^2}{k^2} \cdot t$$

En considérant que tous les noeuds ont été ajoutés à intervalles de temps égaux :

$$p_i(t_i) = \frac{1}{m_0 + t}$$

D'où :

$$p(k < k_i(t)) = p\left(t_i > \frac{d^2}{k^2} \cdot t\right)$$

$$= 1 - p\left(t_i \leqslant \frac{d^2}{k^2} \cdot t\right) = 1 - \frac{d^2}{k^2} \cdot t \cdot p_i(t_i)$$

$$= 1 - \frac{d^2}{k^2} \cdot t \cdot \frac{1}{m_0 + t}$$

et donc :

$$\mathrm{P}(k) = \frac{\delta p(k < k_i(t))}{\delta k} = \frac{2 \cdot d^2 \cdot t}{m_0 + t} \cdot \frac{1}{k^3}$$

La distribution de degrés d'un réseau créé par cet algorithme est bien une loi de puissance : c'est un réseau « scale-free ».
Exemples de réseaux générés par l'algorithme (non orientés/orientés) :