Web Structure Mining: Analyse de Liens

HITS et PageRank: les algorithmes qui valaient des millions de dollars. . .

source *Modeling the Internet and the Web*
P. Baldi, P. Frasconi, P. Smyth

L. Ralaivola & C. Capponi

15 janvier 2008

Web Structure Mining: Analyse de Liens

HITS et PageRank: les algorithmes qui valaient des millions de dollars. . .

source *Modeling the Internet and the Web*
P. Baldi, P. Frasconi, P. Smyth

L. Ralaivola & C. Capponi

15 janvier 2008

Plan

Graphes possibles du Web

Graphe physique

Repose sur les couches physiques du réseau : e.g., routeurs, ordinateurs pour les noeuds et « câbles ». . .

- représentation limitée
- pas de capture des relations au niveau informatif

Graphes possibles du Web

Graphe physique

Repose sur les couches physiques du réseau : e.g., routeurs, ordinateurs pour les noeuds et « câbles ». . .

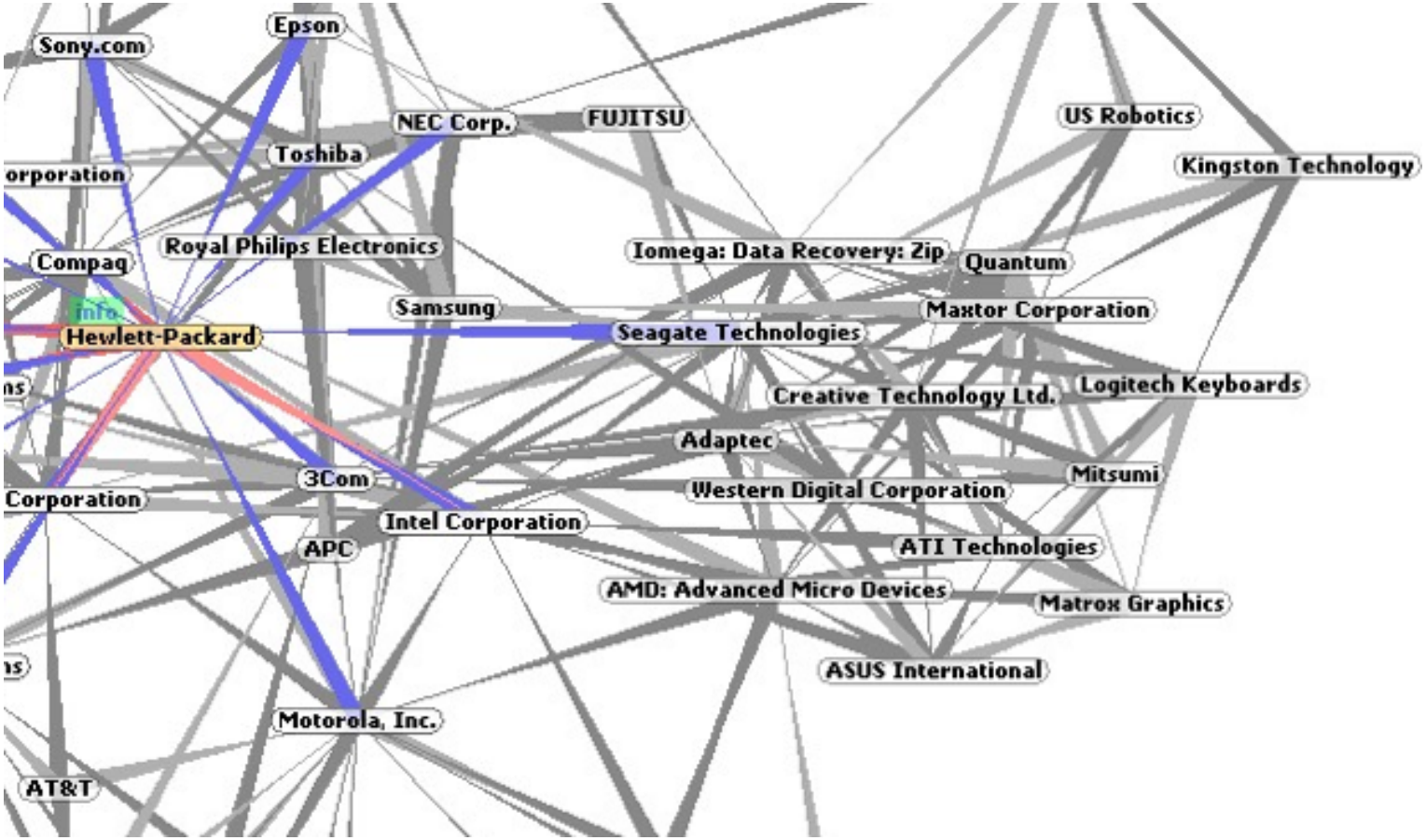
- représentation limitée
- pas de capture des relations au niveau informatif

Graphe logique

Graphe considéré par les algorithmes de WSM. Graphe orienté

- nœuds : pages Web
- arcs : hyperliens

Extrait du graphe logique du Web



http://www.touchgraph.com/TGGoogleBrowser.html

Considérations sur le graphe loqique

Structure

- graphe orienté, non orienté
- graphe dynamique
  - nœuds et arêtes (arcs) ajoutés/enlevés fréquemment
  - contenu des nœuds varie
- composante connexe importante, quelques composantes connexes plus petites

Intérêt de la représentation en graphe

- théorie des graphes : domaine très étudié
- graphe du web similaire à d'autres graphes : sociaux, biologiques, etc.
- organisation naturelle de l'information
- algorithmes efficaces de navigation

Points importants d'analyse

Statistiques d'étude

- taille et connectivité du graphe
- nombre de composantes connexes
- distribution du nombre de pages par site
- distributions des liens entrants et sortants par site
- tailles moyenne et maximale des chemins entre 2 nœuds quelconques

Propriétés générales

Topologie

- graphe creux (faiblement connecté)
- graphe « petit monde »

Propriétés générales

Topologie

- graphe creux (faiblement connecté)
- graphe « petit monde »

Graphe « petit monde »

- Poétiquement
  - millions de nœuds mais faible longueur des chemins séparant n'importe quelle paire de nœuds
  - cf. expérience de Milgram
- Mathématiquement
  - diamètre du graphe faible $O(\log n)$, n nb de nœuds (rassurant pour les analyses futures)
  - intéressant car graphe faiblement connecté

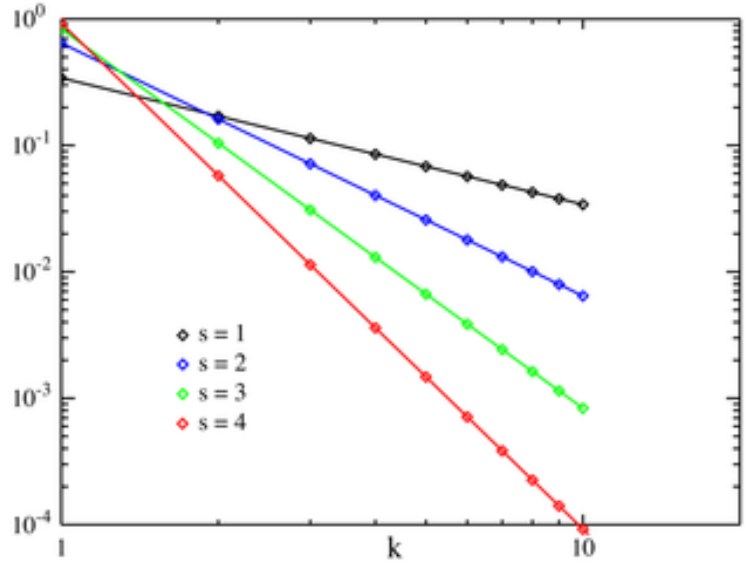
Propriétés générales

Rappel : lois exponentielles

K, variable aléatoire

- cas discret : $P(K = k) = Ck^{-\gamma}$
- cas réel : $p(x) = Cx^{-\gamma}$ (densité de probabilité)

Échelle log / log, loi de Zipf



Distributions de la connectivité, taille des sites

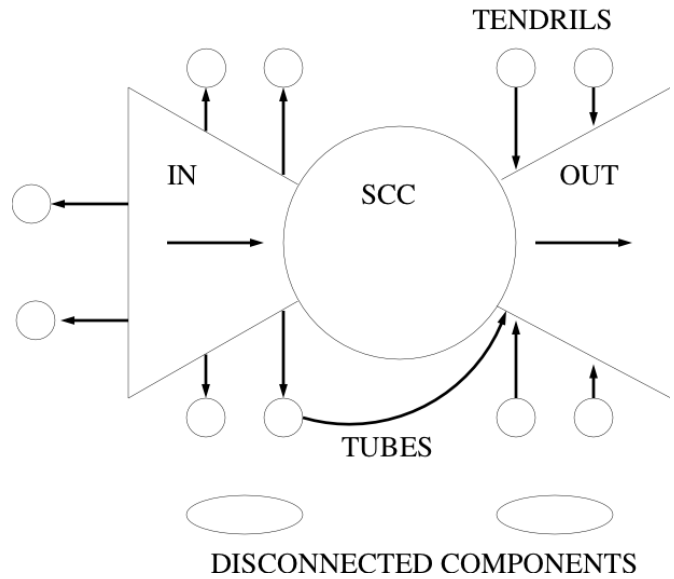
- connectivité suit une loi exponentielle ; étude université de Notre Dame
  - $\gamma = 2.45$ pour le degré sortant
  - $\gamma = 2.15$ pour le degré entrant
- taille des sites suit une distribution exponentielle ($\gamma \in [1.6; 1.9]$)

Réalité : modèle en nœud papillon

Presque un « petit monde »

Résultats des crawls d'Altavista entre mai et octobre 1999 Broder et al. [2000] (200 millions de nœuds et 1.5 milliard de liens)

- chaque composante fait la même taille (50M)... en 1999
- quelques composantes ne sont pas atteignables
- chemins parfois longs entre deux nœuds
- lois exponentielles s'appliquent pour la connectivité (in $\gamma = 2.1$, out $\gamma = 2.72$)
- SCC « petit monde »



Plan

Objectifs

Identification de pages de référence

- trouver des pages reliées à une page pertinente (recommender systems)
- catégorisation de pages web
- crawling du web
- Web communities (collection de pages web tq chaque nœud membre a plus de liens dans sa communauté qu'en dehors)
- Web usage mining

Portée de l'analyse et ses propriétés

- un seul nœud (caractéristiques = pertinence ou qualité)
- un sous-ensemble de nœuds (caractéristiques = distance, communauté)
- le web entier (caractéristiques = diamètre)

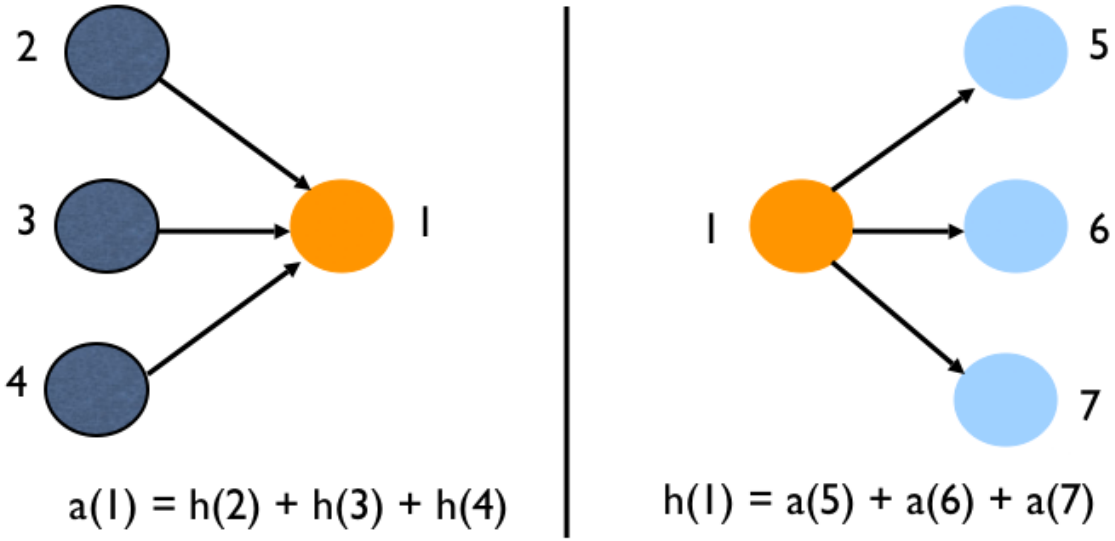
Plan

HITS – Hypertext Induced Topic Selection [Kleinberg, 1999]

Hubs et autorités

- Page d'autorité : vers laquelle se concentrent un bon nombre de liens (provenant de pivots)
- Page pivots (hubs) : qui pointe sur un bon nombre de pages d'autorité
- Dépend d'une requête q : graphe $G_q = (V_q, E_q)$

HITS – Hypertext Induced Topic Selection [Kleinberg, 1999]



Définition récursive

- Score d'autorité d'une page p proprotionnel à la somme des scores de pivot des pages qui pointent sur p
- Score de pivot d'une page p proportionnel aux scores d'autorité des pages que p pointe.

$$\mathbf{a} \propto E^T \mathbf{h} \text{ et } \mathbf{h} \propto E\mathbf{a}$$

Algorithme HITS

HubsAuthorities(G, ε), $G = (V, E)$

$\mathbf{1} \leftarrow [1 \cdots 1] \in \mathcal{R}^{|V|}$
$t \leftarrow 1$
repeat
  for all $v \in V$ **do**
    $a(v) \leftarrow \sum_{w \in pa[v]} h_{t-1}(w)$
    $h(v) \leftarrow \sum_{w \in ch[v]} a_{t-1}(w)$
  end for
  $\mathbf{a}_t \leftarrow \mathbf{a}/\|\mathbf{a}\|$
  $\mathbf{h}_t \leftarrow \mathbf{h}/\|\mathbf{h}\|$
  $t \leftarrow t + 1$
until $\|\mathbf{a}_t - \mathbf{a}_{t-1}\| + \|\mathbf{h}_t - \mathbf{h}_{t-1}\| < \varepsilon$
return $\mathbf{a}_t, \mathbf{h}_t$

Quelques détails

- G graphe de base obtenu par l'algorithme BaseSubgraph (cf. slide suivant)
- $pa[v]$ parents de v
- $ch[v]$ enfants de v
- ε défini par l'utilisateur

Algorithme HITS

HubsAuthorities(G, ε), $G = (V, E)$

$\mathbf{1} \leftarrow [1 \cdots 1] \in \mathcal{R}^{|V|}$
$t \leftarrow 1$
repeat
  for all $v \in V$ **do**
    $a(v) \leftarrow \sum_{w \in pa[v]} h_{t-1}(w)$
    $h(v) \leftarrow \sum_{w \in ch[v]} a_{t-1}(w)$
  end for
  $\mathbf{a}_t \leftarrow \mathbf{a}/\|\mathbf{a}\|$
  $\mathbf{h}_t \leftarrow \mathbf{h}/\|\mathbf{h}\|$
  $t \leftarrow t + 1$
until $\|\mathbf{a}_t - \mathbf{a}_{t-1}\| + \|\mathbf{h}_t - \mathbf{h}_{t-1}\| < \varepsilon$
return $\mathbf{a}_t, \mathbf{h}_t$

Quelques détails

- G graphe de base obtenu par l'algorithme BaseSubgraph (cf. slide suivant)
- $pa[v]$ parents de v
- $ch[v]$ enfants de v
- ε défini par l'utilisateur

Les joies du TD (ou des devoirs maison ?)

Un peu d'algèbre linéaire permet de montrer que sous des hypothèses assez faibles, il y a convergence de l'algorithme.

Algorithme BaseSubgraph(R, d)

BaseSubgraph(R, d)

$S \leftarrow R$
for all $v \in R$ **do**
  $S \leftarrow S \cup ch[v]$
  $P \leftarrow pa[v]$
  if $|P| > d$ **then**
    $P \leftarrow$ sous-ensemble arbitraire de P de taille d
    $S \leftarrow S \cup P$
  end if
end for
return S

Quelques détails

- R : ensemble de pages dont on sait qu'elles ont un rapport avec le sujet (ou la requête) étudié
- d : paramètre défini par l'utilisateur
- S (+ arcs) : graphe à fournir à HubsAuthorities

Algorithme BaseSubgraph(R, d)

BaseSubgraph(R, d)

$S \leftarrow R$
for all $v \in R$ **do**
  $S \leftarrow S \cup ch[v]$
  $P \leftarrow pa[v]$
  if $|P| > d$ **then**
    $P \leftarrow$ sous-ensemble arbitraire de P de taille d
    $S \leftarrow S \cup P$
  end if
end for
return S

Quelques détails

- R : ensemble de pages dont on sait qu'elles ont un rapport avec le sujet (ou la requête) étudié
- d : paramètre défini par l'utilisateur
- S (+ arcs) : graphe à fournir à HubsAuthorities

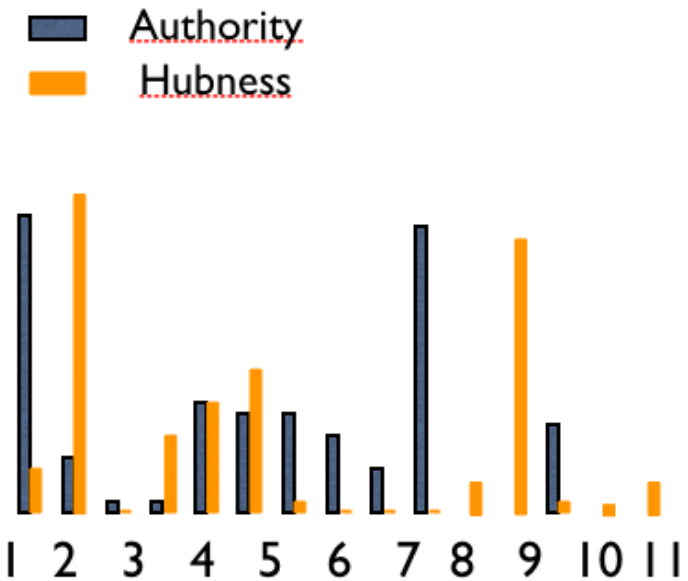
Les joies du TD (ou des devoirs maison ?)

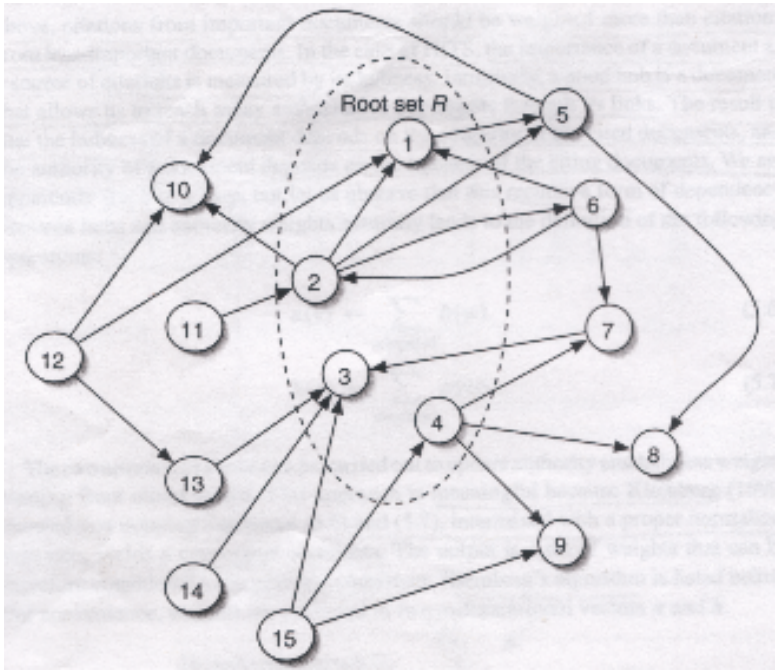
- Quid si d trop grand ? trop petit ?
- Supposons que d soit inférieur à $|P|$ (initialement). Combien faut-il tirer au hasard de sous-ensembles de P de taille d pour être sûr à 95% que l'un d'eux fait partie des 5% meilleurs sous-ensembles ?

Exemple HITS

Résultat

- Nœuds de départ : $\{1, 2, 3, 4\}$
- Sous graphe pour HubsAuthorities : 15 nœuds, 20 arcs



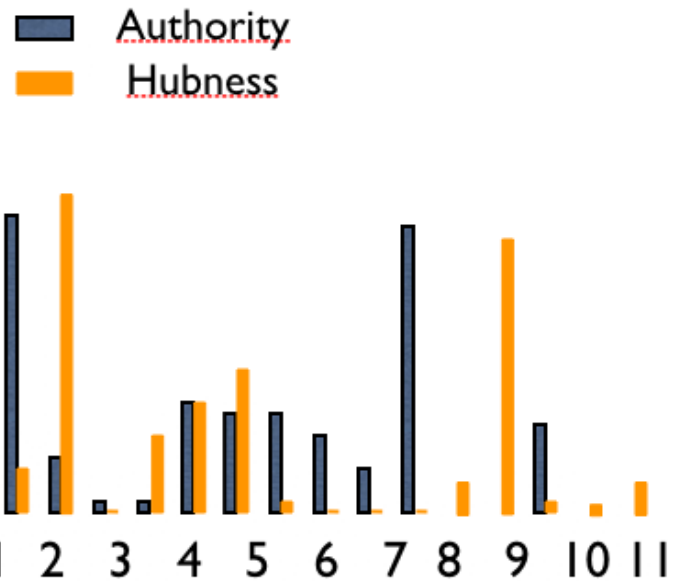


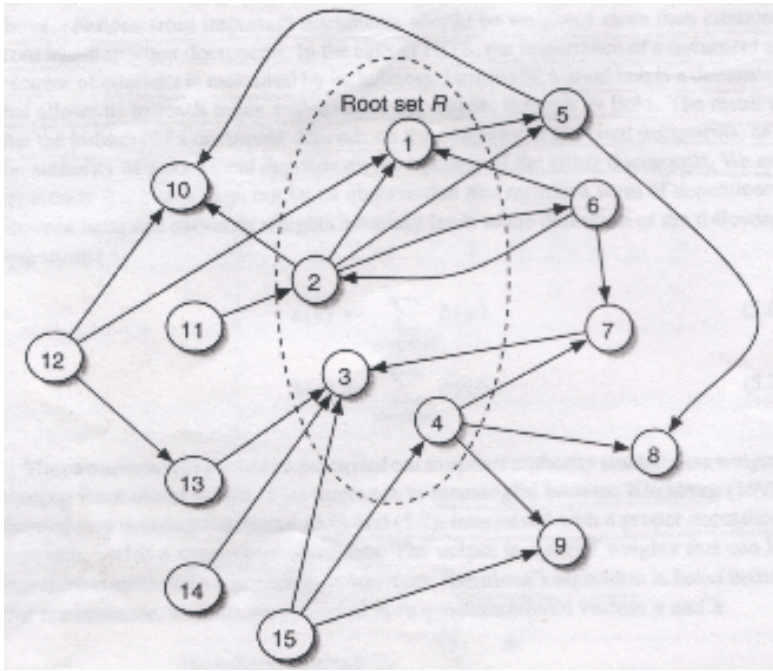
Résultat surprenant au niveau du nœud 3 ! ?

Exemple HITS

Résultat

- Nœuds de départ : $\{1, 2, 3, 4\}$
- Sous graphe pour HubsAuthorities : 15 nœuds, 20 arcs





Résultat surprenant au niveau du nœud 3 ! ?

Les joies du TP

Programmer les algorithmes décrits. Les tester sur le graphe fourni et d'autres graphes (aléatoires).

Améliorations de HITS

Problèmes HITS

- Renforcement mutuel artificiel lorsqu'un document contient plusieurs liens vers un autre même document
- Liens artificiels non informatifs avec des pages créées artificiellement (e.g. newsgroups)
- Topic drift : problème lié à un mauvais sous-graphe de départ

Bharat and Henzinger [1998]

- Corriger les poids des liens multiples en les ajustant proportionnellement à l'inverse de leur multuplicité
- Elagage de nœuds et liens par rapport à une mesure de pertinence définie par ailleurs

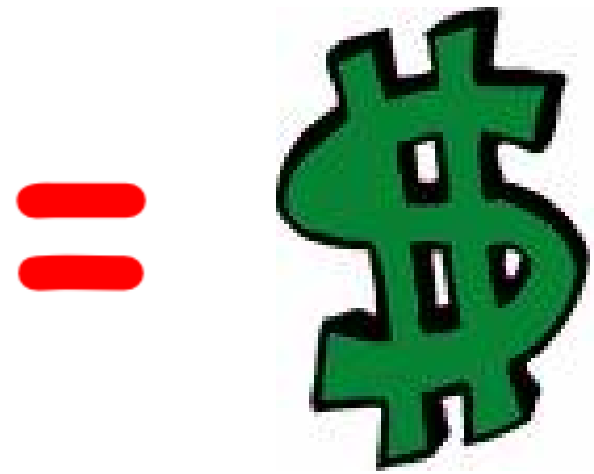
Plan

Les mathématiques de la fortune

Rev. T. Bayes

Les mathématiques de la fortune

Rev. T. Bayes



Idée de PageRank, [Page et al., 1998]

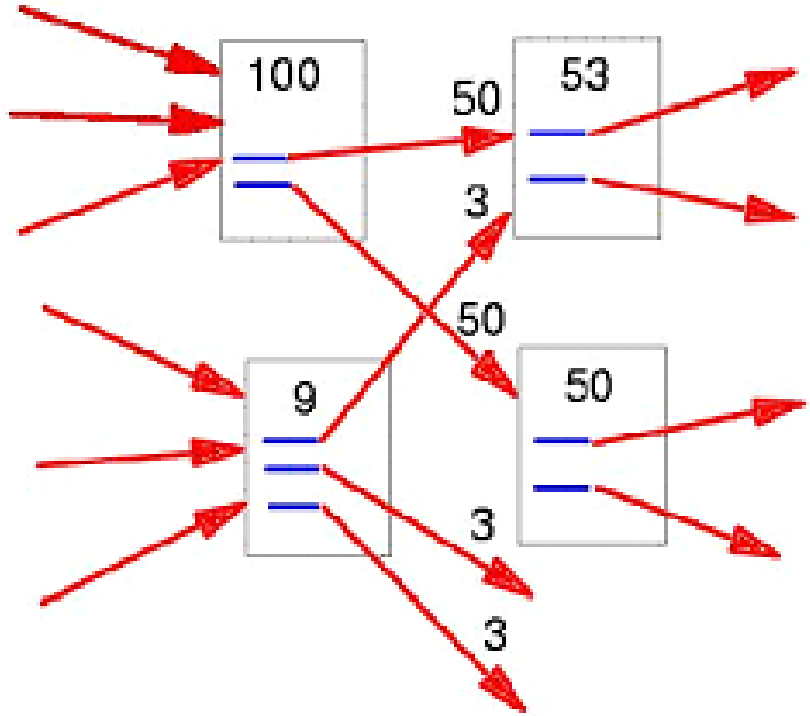
Par Google Technology

PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives ; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

Première approximation

Modèle : $r(v)$, PageRank de v

$$r(v) = \alpha \sum_{w \in pa[v]} \frac{r(w)}{|ch[w]|}$$



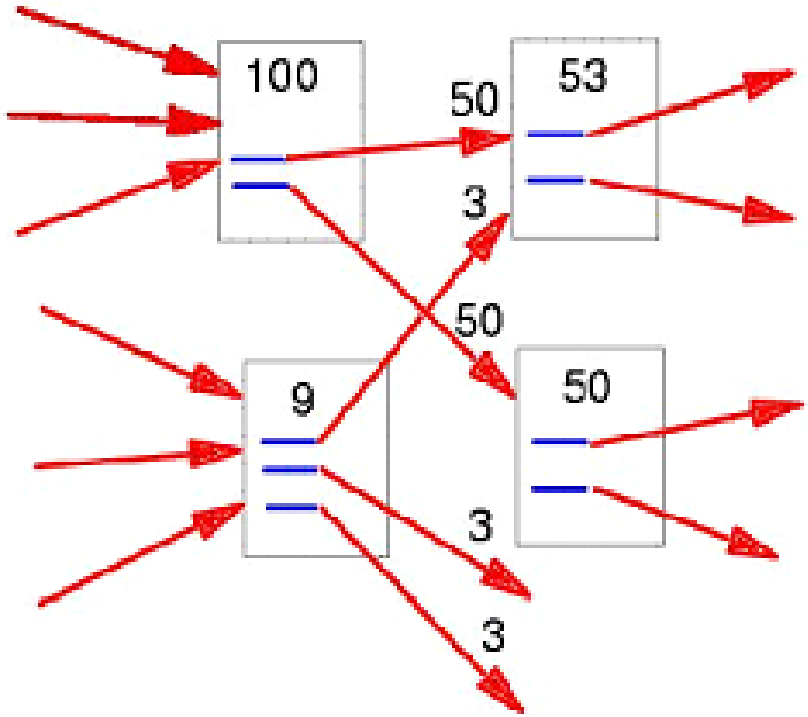
Différences avec HITS

- 1 seul type de poids
- r d'un nœud proportionnel au r de ses parents, modulé par leurs degrés sortants

Première approximation

Modèle : $r(v)$, PageRank de v

$$r(v) = \alpha \sum_{w \in pa[v]} \frac{r(w)}{|ch[w]|}$$

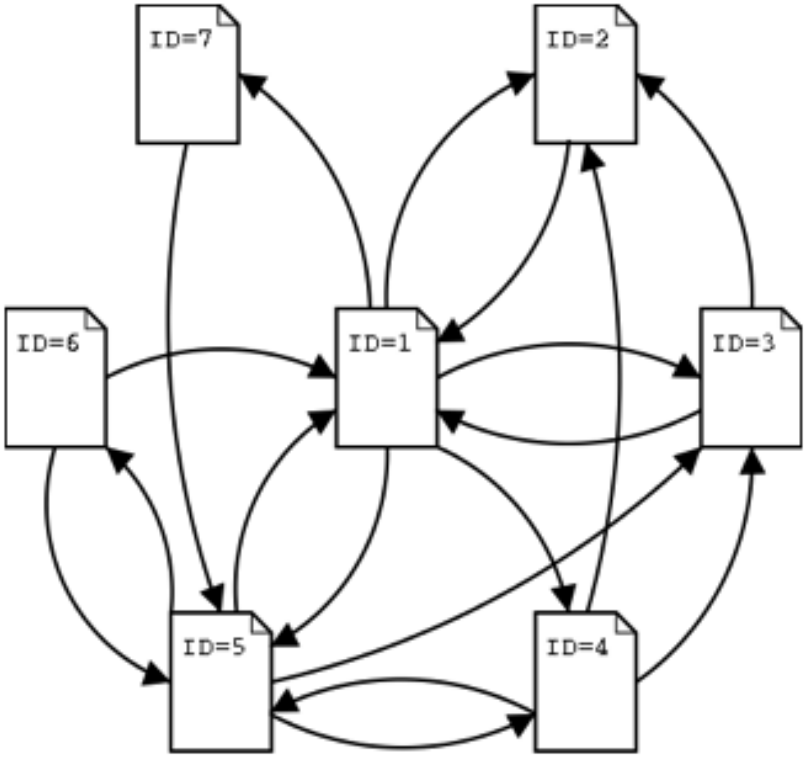


Sous forme matricielle

$$\mathbf{r} = \alpha B \mathbf{r} = M\mathbf{r}, \quad \text{avec } b_{uv} = \begin{cases} \frac{a_{uv}}{\sum_w a_{uw}}, & \text{si } ch(u) \neq \emptyset, \\ a_{uv} = 0 & \text{sinon} \end{cases}$$

avec A matrice d'adjacence du graphe

Exemple



Page ID	OutLinks
1	2,3,4,5,7
2	1
3	1,2
4	2,3,5
5	1,3,4,6
6	1,5
7	5

Matrice d'adjacence

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

http://www.kusatro.kyoto-u.com

Exemple

PageRank: recherche du vecteur propre associé à la plus grande valeur propre de B

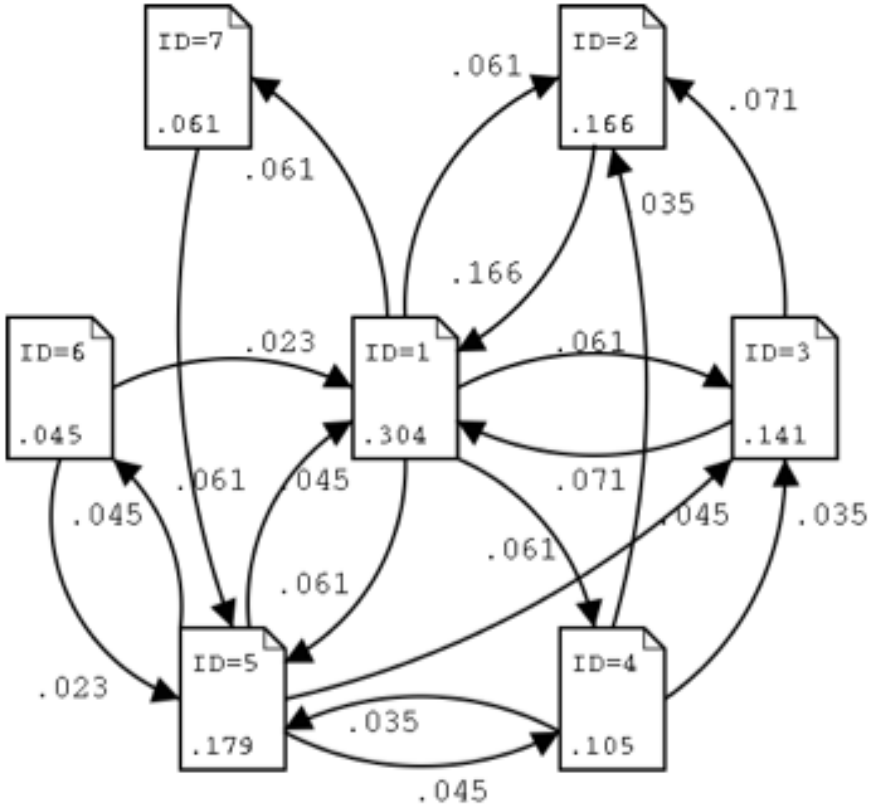
$$B = R\,D\,R^T = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} \lambda_1 & & O \\ & \lambda_2 & \\ O & & \ddots \\ & & \lambda_n \end{pmatrix}$$

$$R = \begin{pmatrix} r_1 & r_2 & \cdots & r_n \end{pmatrix}$$

$$\text{PageRank } r_1 = \begin{pmatrix} 0.69946 \\ 0.38286 \\ 0.32396 \\ 0.24297 \\ 0.41231 \\ 0.10308 \\ 0.13989 \end{pmatrix} \xrightarrow{\text{normalisation}} \begin{pmatrix} 0.303514 \\ 0.166134 \\ 0.140575 \\ 0.105431 \\ 0.178914 \\ 0.044728 \\ 0.060703 \end{pmatrix}$$

Exemple



PR	ID	OutLink	InLink
0.304	1	2,3,4,5,7	2,3,5,6
0.179	5	1,3,4,6	1,4,6,7
0.166	2	1	1,3,4
0.141	3	1,2	1,4,5
0.105	4	2,3,5	1,5
0.061	7	5	1
0.045	6	1,5	5

Question

Que faire pour avoir une haute valeur de PR ? (des sociétés sont spécialisées dans ce domaine)

Et Bayes sur son surf, dans tout ça ?

Marche (surf) aléatoire

- Le PageRank d'une page v est la probabilité qu'un surfeur qui se déplace de nœud en nœud en suivant les liens du graphe, chacun ayant une probabilité bien définie d'être emprunté, se trouve sur cette page v
- Simulation d'une marche aléatoire

$$r_t(v) = P(S_t = v) = \sum_w P(S_t = v | S_{t-1} = w) P(S_t = w)$$
$$= \sum_w m_{wv} r_{t-1}(w)$$

- Soit, $\mathbf{r}_t = M^T \mathbf{r}_{t-1}$, avec M qui *doit* être stochastique
- Indépendant de toute requête
- Question : conditions pour qu'il y ait une convergence de $\mathbf{r}_t$?
- Problème lorsqu'on arrive à des pages qui n'ont pas de lien sortant (puits) (cf. nœud papillon

Et Bayes sur son surf, dans tout ça ?

M stochastique

Matrice de transition à considérer (n, nombre de nœuds considérés)

$$M = (B + E), \quad \text{avec } e_{uv} = \begin{cases} 0 & \text{si } ch(v) \neq \emptyset, \\ \frac{1}{n} & \text{sinon} \end{cases}$$

PageRank(M, n, ε)

$\mathbf{1} \leftarrow [1, \cdots, 1] \in \mathcal{R}^n$
$\mathbf{z} \leftarrow \frac{1}{n}\mathbf{1}$
$\mathbf{r}_0 \leftarrow \mathbf{z}$
$t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  $\mathbf{r}_t = M^T \mathbf{r}_{t-1}$
  $d_t \leftarrow \|\mathbf{r}_{t-1}\|_1 - \|\mathbf{r}_t\|_1$
  $\mathbf{r}_t \leftarrow \mathbf{r}_t + d_t \mathbf{z}$
  $\delta \leftarrow \|\mathbf{r}_{t-1} - \mathbf{r}_t\|_1$
until $\delta < \varepsilon$
return $\mathbf{r}_t$

Matrice primitive

Si M est primitive, i.e., il existe un t tel que $M^t > 0$ (chaque entrée strictement positive) alors PageRank converge (M est stochastique) et renvoie le premier vecteur propre de M.

Et Bayes sur son surf, dans tout ça ?

Rendre M primitive, éviter la périodicité

- il suffit de perturber le vecteur $\mathbf{r}$ en fonction de α

$$\mathbf{r} = \alpha\mathbf{e} + (1-\alpha)\mathbf{r}$$

  avec $\mathbf{e} = (1/n)\mathbf{1}$

- formule résultante

$$\mathbf{r}_t = [\alpha H + (1-\alpha)M]^T \mathbf{r}_{t-1}$$

  avec H d'ordre n et $h_{uv} = 1/n$

- matrice de transition stochastique *et* primitive : convergence
- en pratique ($\alpha \in [0.1; 0.2]$)

En pratique

- On peut calculer directement le premier vecteur propre de la matrice de transition. . . mais n trop grand
- PageRank implémente une stratégie de *puissance itérée*
- Google (1998) : 52 itérations pour convergence (322 millions de liens)

Pour finir

Utilisation de PageRank

- guidage du crawling
- combinaison avec pertinence de documents par rapport à une requête
- facteurs d'impact de revues scientifiques
- ...

Question/limitations PageRank/HITS

- Stabilité de l'algorithme
- Pas de prise en compte du contenu des pages
- *Link farms* et autres stratégies de *link buying*

Plan

What to take home ?

Web et graphes

- Représentation du Web comme un graphe
- Extraction d'information par rapport à l'architecture du graphe
- Algorithme HITS : hubs et authorities
- Algorithme PageRank : marche aléatoire sur le graphe

A retenir

Importance des idées plus que de la technologie/technicité des solutions

Plan

Pour s'amuser

Co-citation et matrice bibliographique

Soit A la matrice d'incidence du graphe G. Soit $C = A^T A$ et $B = AA^T$. Montrer que c_{uv} est le nombre de documents qui citent les documents u et v et b_{uv} est le nombre de pages qui sont citées par u et v

Convergence

Soit N une matrice carrée d'ordre n dont les lignes sont normalisées. Supposons que N n'est pas stochastique et R est une matrice telle que $N + R$ est stochastique. Montrer que PageRank(N, n, ε) et PageRank$(N + R, n, \varepsilon)$ convergent vers la même solution.

Références

Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, AU, 1998. URL `citeseer.ist.psu.edu/bharat98improved.html`.

A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web : experiments and models. In *9th WWW Conf.*, pages 309–320, 2000. URL `http://www9.org/w9cdrom/160/160.html`.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5) :604–632, 1999. URL `citeseer.ist.psu.edu/kleinberg99authoritative.html`.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. URL `citeseer.ist.psu.edu/page98pagerank.html`.