RECHERCHE D'INFORMATION & TRAITEMENT AUTOMATIQUE DU LANGAGE

RI - Recherche web

2022-23

Benjamin Piwowarski / Laure Soulier

Introduction

Analyse de liens

- Popularisée par Google avec PageRank
- Actuellement une composante parmi beaucoup d'autres des moteurs de recherche
- De l'ordre de 400 caractéristiques prises en compte
- Cours : 2 algorithmes historiques
  - PageRank (Brin & Page 1998)
  - HITS (Kleinberg 1998)
  - Très nombreuses variantes

    E.g. trustrank

Les liens

- Le web est vu comme un graphe orienté
- Les liens sont porteurs d'information
  - Un lien entre pages indique une relation de pertinence
    Un lien est un indicateur de qualité
  - Le texte d'un lien résume la page cible
    L'indexation d'une page doit prendre en compte les liens vers cette page (contexte)

PageRank

PageRank in 1 slide

- Principe général
  - Popularized by google
  - Assign an authority score for each web page
  - Using only the structure of the web graph (query independent)
  - Now one of the many components used for computing page scores in Google S.E.
- Intuition
  - Assign higher scores to pages with many in-links from authoritative pages with few out-links
- Modèle
  - Random surfer model : Stationary distribution of a Markov Chain
  - Principal eigenvector of a linear system

Notations

- Graphe orienté $G = (V, E)$
- A matrice d'adjacence

$$a_{ij} = \begin{cases} 1 & \text{s'il y a un lien de i vers j} \\ 0 & \text{sinon} \end{cases}$$

- Le nombre de liens entrants pour une page i est

$$d_i = \sum_j a_{ij}$$

- P matrice de transition

$$P = \left(\frac{a_{ij}}{d_i} \right)_{ij}$$

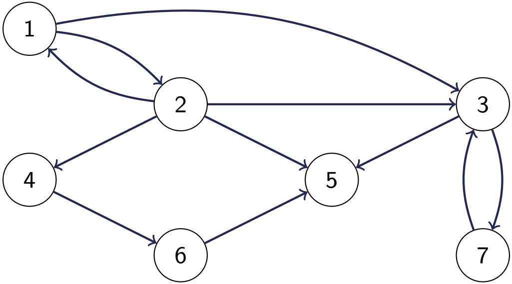
avec P_{ij} la probabilité de transition, i.e. d'aller de j à i. On a :

$$\sum_j p_{ij} = 1$$

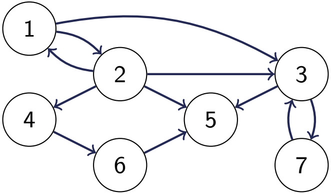
i.e.

$$p_{ij} \stackrel{\text{def}}{=} p(\text{aller sur } j | \text{point de départ } i)$$

Graphe

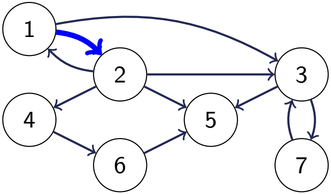


Représentation matricielle



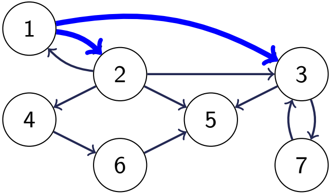
$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle



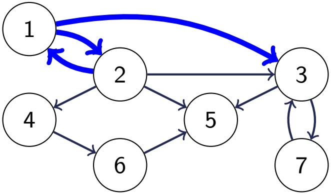
$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle



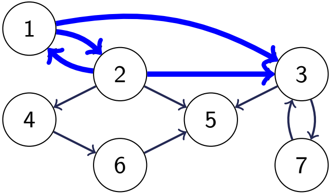
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle



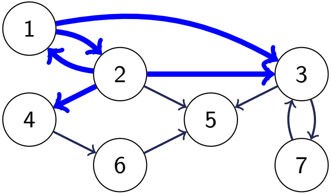
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle



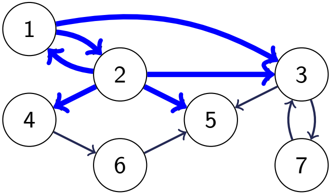
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle



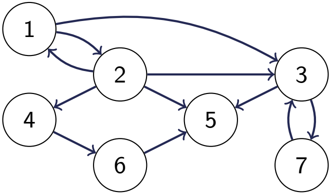
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle



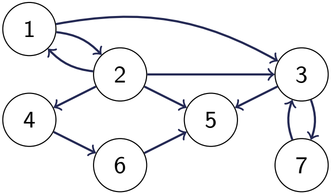
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle



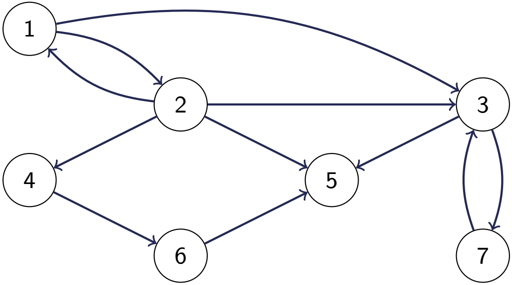
$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Représentation matricielle

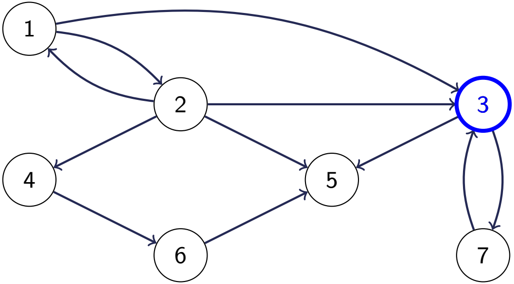


$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

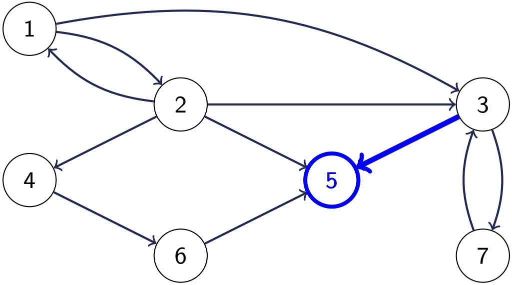
Surfer stochastique



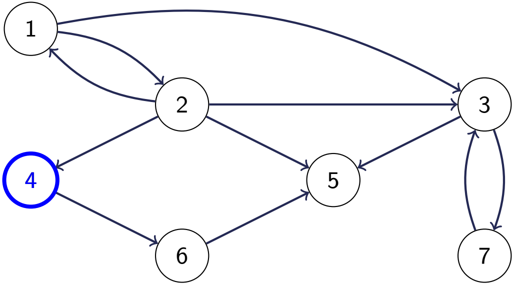
Surfer stochastique



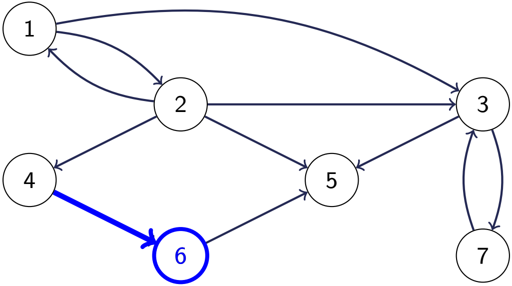
Surfer stochastique



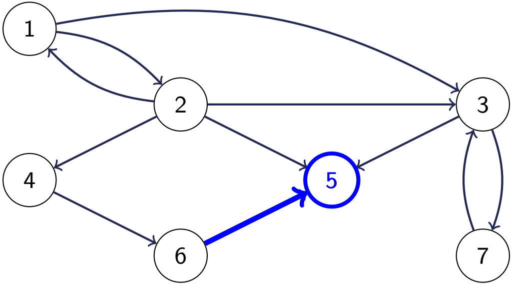
Surfer stochastique



Surfer stochastique



Surfer stochastique



PageRank

Distribution stationnaire $\mathbf{s}_j$

$$\mathbf{s}_j = \underbrace{\sum_i p_{ij}\mathbf{s}_i}_{\text{propagation de l'importance}} \tag{1}$$

- $\mathbf{s}_j$ correspond à la probabilité que la page i soit importante $\sum_i \mathbf{s}_i = 1$

PageRank

Distribution stationnaire $\mathbf{s}_j$

$$\mathbf{s}_j = \underbrace{d \sum_i p_{ij} \mathbf{s}_i}_{\substack{\text{propagation} \\ \text{de l'importance}}} + \underbrace{(1-d)\, a_j}_{\text{a priori}} \tag{1}$$

- d est le facteur d'amortissement (*damping factor*). Les valeurs typiques de d sont autour de 0.8
- a_i correspond à la probabilité que la page i soit importante *a priori* $\sum_i a_i = 1$

Vision matricielle

- Version initiale

$$\mathbf{s}_j = \underbrace{d \sum_i \mathbf{p}_{ij}\mathbf{s}_i}_{\text{propagation de l'importance}} + \underbrace{(1-d)\,\mathbf{a}_j}_{\text{a priori}}$$

Vision matricielle

- Version initiale

$$\mathbf{s}_j = \underbrace{d \sum_i \mathbf{p}_{ij} \mathbf{s}_i}_{\text{propagation de l'importance}} + \underbrace{(1-d)\,\mathbf{a}_j}_{\text{a priori}}$$

- Version matricielle

$$\mathbf{s} = d\mathbf{s}\mathbf{P} + (1-d)\,\mathbf{a}$$

Un peu d'algèbre linéaire...

$$\mathbf{s} = d\mathbf{s}\mathbf{P} + (1-d)\,\mathbf{a} = \mathbf{s}\,(d\mathbf{P} + (1-d)\,\mathbf{E}) = s\mathbf{P}'$$

Un peu d'algèbre linéaire...

$$\mathbf{s} = d\mathbf{s}\mathbf{P} + (1-d)\,\mathbf{a} = \mathbf{s}\,(d\mathbf{P} + (1-d)\,\mathbf{E}) = s\mathbf{P}'$$

- Que vaut s ?

Un peu d'algèbre linéaire...

$$\mathbf{s} = d\mathbf{s}\mathbf{P} + (1-d)\,\mathbf{a} = \mathbf{s}\,(d\mathbf{P} + (1-d)\,\mathbf{E}) = s\mathbf{P}'$$

- Que vaut s ?
- Rappel sur les valeurs propres : $AX = \lambda X$ (λ : valeur propre et X : vecteur propre

Un peu d'algèbre linéaire...

$$\mathbf{s} = d\mathbf{s}\mathbf{P} + (1-d)\,\mathbf{a} = \mathbf{s}\,(d\mathbf{P} + (1-d)\,\mathbf{E}) = s\mathbf{P}'$$

- Que vaut s ?
- Rappel sur les valeurs propres : $AX = \lambda X$ (λ : valeur propre et X : vecteur propre
- s est le vecteur propre associé à la valeur propre 1

Un peu d'algèbre linéaire...

$$\mathbf{s} = d\mathbf{s}\mathbf{P} + (1-d)\,\mathbf{a} = \mathbf{s}\,(d\mathbf{P} + (1-d)\,\mathbf{E}) = s\mathbf{P}'$$

- Que vaut s ?
- Rappel sur les valeurs propres : $AX = \lambda X$ (λ : valeur propre et X : vecteur propre
- s est le vecteur propre associé à la valeur propre 1
- Il existe une solution unique

Explication théorique

A est une matrice irréductible : correspond à un graphe fortement connecté

- A square matrix A_{nxn} is ***non negative*** if $a_{ij} \geq 0$
  - Notation $A \geq 0$
  - Example: graph incidence matrix
- A_{nxn} is ***positive*** if $a_{ij} > 0$
  - Notation $A > 0$
- A_{nxn} is ***irreducible*** if
  - $\forall\, i, j, \exists\, t \in N \,/ (A^t)_{ij} > 0$
  - If A is a graph incidence matrix, this means that G is strongly connected
    - There is a path between any pair of vertices
- A_{nxn} is ***primitive*** if $\exists\, t \in N \,/ A^t > 0$
  - A primitive matrix is irreducible
  - Converse is false

Explication théorique

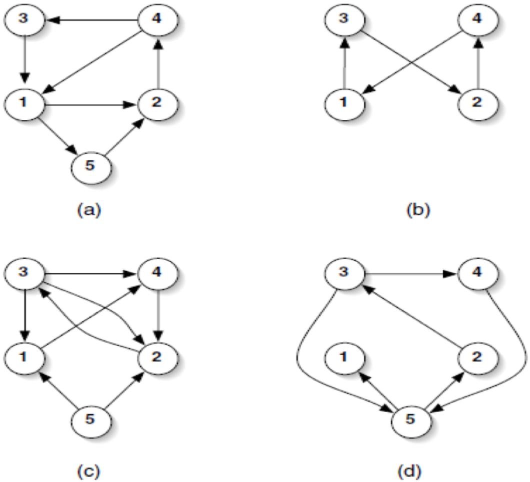
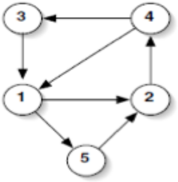


Figure 5.1 Graphs with different types of incidence matrices. (a) is primitive, (b) is irreducible (with period 4) but not primitive, (c) and (d) are reducible.

Explication théorique

Examples (Baldi et al. 2003)



(a)

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}^7 = \begin{pmatrix} 4 & 2 & 3 & 3 & 1 \\ 3 & 4 & 1 & 1 & 3 \\ 1 & 3 & 1 & 3 & 1 \\ 2 & 6 & 1 & 4 & 3 \\ 3 & 1 & 2 & 1 & 1 \end{pmatrix}$$

Matrice Primitive
« Il existe une puissance pour laquelle tous les éléments sont strictement positifs »

En termes de graphe : On peut naviguer entre tous les noeuds

Explication théorique

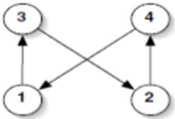
Examples (Baldi et al. 2003)

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

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}^2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\equiv \qquad \equiv$$

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}^3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$



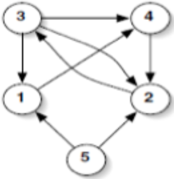
Irréductible :
« Il existe un exposant qui permet pour tout élément d'avoir une valeur de 1 »

En termes de graphe : On peut naviguer entre tous les noeuds

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}^4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Explication théorique

Examples (Baldi et al. 2003)



(c)

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}^{10} = \begin{pmatrix} 4 & 8 & 5 & 6 & 0 \\ 8 & 17 & 11 & 13 & 0 \\ 11 & 24 & 17 & 19 & 0 \\ 5 & 11 & 8 & 9 & 0 \\ 7 & 16 & 12 & 13 & 0 \end{pmatrix}$$

Réductible :
« On a toujours un bloc de 0, quelque que soit la puissance »

En termes de graphe : On ne peut naviguer entre tous les noeuds

Perron-Frobenius

Définitions (Matrice irréductible)

- A est irréductible si $\exists n\, \mathbf{A}^n > 0$
- Correspond à un graphe fortement connecté

- Si $\mathbf{A}_{n \times n}$ est une matrice non-négative irréductible et apériodique
  - A a une valeur propre réelle λ_1 et positive supérieure unique telle que $\forall j,\ |\lambda_1| > |\lambda_j|$
  - Le vecteur propre s associé à λ_1 est strictement positif
- Dans notre cas, la valeur propre maximum est ?

Perron-Frobenius

Définitions (Matrice irréductible)

- A est irréductible si $\exists n\, \mathbf{A}^n > 0$
- Correspond à un graphe fortement connecté

- Si $\mathbf{A}_{n\times n}$ est une matrice non-négative irréductible et apériodique
  - A a une valeur propre réelle λ_1 et positive supérieure unique telle que $\forall j,\ |\lambda_1| > |\lambda_j|$
  - Le vecteur propre s associé à λ_1 est strictement positif
- Dans notre cas, la valeur propre maximum est 1

Preuve de convergence

- Méthodes des puissances
  $A = U\Sigma U^t \rightarrow A^n = U\Sigma^n U^t$

$$\Sigma^n \xrightarrow{\infty} \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

  Revient à sélectionner le premier vecteur propre

En pratique...

- Version itérative :
  - Initialiser s aléatoirement
  - Répéter pour chaque noeud

$$\mathbf{s}_j = d \sum_i \mathbf{p}_{ij} \mathbf{s}_i + (1 - d)\, a_j$$

- Dans les deux cas, jusqu'à ce que $\left| s^{(t+1)} - s^{(t)} \right| < \epsilon$
- Rapidité de la convergence : géométrique avec un ratio λ_1 / λ_2

Démonstration

$$\mathbf{A} = \begin{pmatrix} . & . & 1.0 & 1.0 & 1.0 & . & 1.0 & . & 1.0 & . \\ . & . & . & 1.0 & 1.0 & 1.0 & . & . & . & . \\ 1.0 & . & . & 1.0 & 1.0 & 1.0 & . & . & . & . \\ 1.0 & . & 1.0 & . & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ . & . & 1.0 & 1.0 & . & 1.0 & . & 1.0 & 1.0 & . \\ 1.0 & 1.0 & 1.0 & . & 1.0 & . & 1.0 & . & 1.0 & 1.0 \\ . & 1.0 & . & . & 1.0 & 1.0 & . & 1.0 & . & . \\ 1.0 & 1.0 & . & 1.0 & 1.0 & 1.0 & . & . & 1.0 & . \\ . & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & . & . & . & . \\ 1.0 & 1.0 & 1.0 & 1.0 & . & 1.0 & 1.0 & . & . & . \end{pmatrix}$$

Démonstration

$$\mathbf{P} = \begin{pmatrix}
. & . & 0.2 & 0.2 & 0.2 & . & 0.2 & . & 0.2 & . \\
. & . & . & 0.33 & 0.33 & 0.33 & . & . & . & . \\
0.25 & . & . & 0.25 & 0.25 & 0.25 & . & . & . & . \\
0.12 & . & 0.12 & . & 0.12 & 0.12 & 0.12 & 0.12 & 0.12 & 0.12 \\
. & . & 0.2 & 0.2 & . & 0.2 & . & 0.2 & 0.2 & . \\
0.14 & 0.14 & 0.14 & . & 0.14 & . & 0.14 & . & 0.14 & 0.14 \\
. & 0.25 & . & . & 0.25 & 0.25 & . & 0.25 & . & . \\
0.17 & 0.17 & . & 0.17 & 0.17 & 0.17 & . & . & 0.17 & . \\
. & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & . & . & . & . \\
0.17 & 0.17 & 0.17 & 0.17 & . & 0.17 & 0.17 & . & . & .
\end{pmatrix}$$

Démonstration

Iteration 1, 2, 5, 10

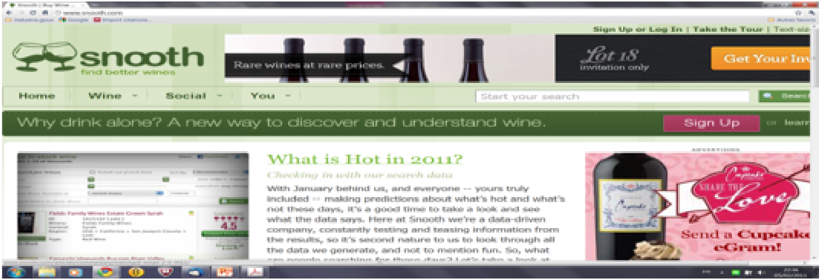
$$\mathbf{x}_1^\top = \begin{pmatrix} 0.07 \\ 0.078 \\ 0.13 \\ 0.17 \\ 0.14 \\ 0.15 \\ 0.07 \\ 0.062 \\ 0.12 \\ 0.015 \end{pmatrix}, \mathbf{x}_2^\top = \begin{pmatrix} 0.088 \\ 0.075 \\ 0.11 \\ 0.14 \\ 0.17 \\ 0.16 \\ 0.06 \\ 0.066 \\ 0.095 \\ 0.043 \end{pmatrix}, \mathbf{x}_5^\top = \begin{pmatrix} 0.086 \\ 0.075 \\ 0.12 \\ 0.14 \\ 0.16 \\ 0.16 \\ 0.064 \\ 0.065 \\ 0.1 \\ 0.04 \end{pmatrix}, \mathbf{x}_{10}^\top = \begin{pmatrix} 0.086 \\ 0.076 \\ 0.12 \\ 0.14 \\ 0.16 \\ 0.16 \\ 0.064 \\ 0.065 \\ 0.1 \\ 0.04 \end{pmatrix}$$

HITS

Hubs : page de pointeurs

- Points good authority pages

Authorities : pages de références thématiques



- Pointed by good hub pages
- Référence importante pour un thème

Hubs et Authorities

SCIENCES SORBONNE UNIVERSITÉ

- Pour une page i
  - Score hub = somme des score des *authorities* des pages pointées par i
  - Score authority = somme des scores des hubs des pages qui pointent vers i
- Formellement,

$$
\begin{aligned}
h_i &= \sum_{i \to j} a_j \\
a_i &= \sum_{j \to i} h_j
\end{aligned}
$$

Algorithme

- Itératif
  - On calcule a et h à partir des valeurs estimées à l'itération précédente
  - il faut normaliser a et h à chaque étape pour que cela fonctionne (norme L2 égale à 1)
- Version algèbre linéaire

$$
\begin{aligned}
\lambda_h \mathbf{h} &= \mathbf{h}\mathbf{P}\mathbf{P}^\top \\
\lambda_a \mathbf{a} &= \mathbf{a}\mathbf{P}^\top\mathbf{P}
\end{aligned}
$$