Modèles et algorithmes des réseaux

PageRank

et les chaînes de Markov

Ana Busic

Inria Paris - DI ENS

http://www.di.ens.fr/~busic/
ana.busic@inria.fr

Paris, Septembre 2018

Recherche sur Internet en utilisant les liens

Un peu d'histoire

- web : 1989
- browser : 1990
- web portal : 1994

Explosion de pages web : comment retrouver les pages pertinentes ?

Recherche sur Internet en utilisant les liens

Un peu d'histoire

- web : 1989
- browser : 1990
- web portal : 1994

Explosion de pages web : comment retrouver les pages pertinentes ?

Brin & Page : algorithme PageRank

S. Brin, L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, 1998.

Recherche sur Internet en utilisant les liens

Un peu d'histoire

- web : 1989
- browser : 1990
- web portal : 1994

Explosion de pages web : comment retrouver les pages pertinentes ?

Brin & Page : algorithme PageRank

S. Brin, L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, 1998.

Si beaucoup de pages
pointent vers une page alors c'est une page pertinente.

Recherche sur Internet en utilisant les liens

Un peu d'histoire

- web : 1989
- browser : 1990
- web portal : 1994

Explosion de pages web : comment retrouver les pages pertinentes ?

Brin & Page : algorithme PageRank

S. Brin, L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, 1998.

Si beaucoup de pages pertinentes
pointent vers une page alors c'est une page pertinente.

Recherche sur Internet en utilisant les liens

Un peu d'histoire

- web : 1989
- browser : 1990
- web portal : 1994

Explosion de pages web : comment retrouver les pages pertinentes ?

Brin & Page : algorithme PageRank

S. Brin, L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, 1998.

Si beaucoup de pages pertinentes
pointent vers une page alors c'est une page pertinente.

Principe de vote basé sur les liens entrants
avec le principe des améliorations répétées.

Recherche sur Internet en utilisant les liens

Un peu d'histoire

- web : 1989
- browser : 1990
- web portal : 1994

Explosion de pages web : comment retrouver les pages pertinentes ?

Brin & Page : algorithme PageRank

S. Brin, L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, 1998.

Si beaucoup de pages pertinentes
pointent vers une page alors c'est une page pertinente.

Principe de vote basé sur les liens entrants
avec le principe des améliorations répétées.

Pages web, articles scientifiques, documents juridiques, ...

Algorithme PageRank

Graphe dirigé (V, A), avec $n = |V|$ pages ;

Notation :

- $i \to j \Leftrightarrow (i, j) \in A$
- d_i degré sortant de i

$$H_{ij} = \begin{cases} 1/d_i, & i \to j \\ 0, & \text{sinon} \end{cases}$$

Idée de base :

$$\begin{aligned} x(0) &= (1/n, \ldots, 1/n) \\ x_i(k+1) &= \sum_{j \to i} \frac{x_j(k)}{d_j} \end{aligned}$$

Problème : et les pages sans liens sortants ?

Algorithme PageRank

Pour résoudre le problème de nœuds absorbants :

(notation en vecteurs colonnes)

$$\hat{H} = H + w\frac{1}{n}\mathbf{1}^t.$$

où $w_i = \begin{cases} 1, & d_i = 0 \\ 0, & \text{sinon} \end{cases}$

Algorithme PageRank

Pour résoudre le problème de nœuds absorbants :
(notation en vecteurs colonnes)

$$\hat{H} = H + w\frac{1}{n}\mathbf{1}^t.$$

où $w_i = \begin{cases} 1, & d_i = 0 \\ 0, & \text{sinon} \end{cases}$

Variante plus générale :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$ (vecteur stochastique).

Algorithme PageRank

Pour résoudre le problème de nœuds absorbants :
(notation en vecteurs colonnes)

$$\hat{H} = H + w\frac{1}{n}\mathbf{1}^t.$$

où $w_i = \begin{cases} 1, & d_i = 0 \\ 0, & \text{sinon} \end{cases}$

Variante plus générale :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$ (vecteur stochastique).

Questions :

- Sous quelles hypothèses $x(k)$ converge ?
- Comment calculer la limite ?

Algorithme PageRank

Pour résoudre le problème de nœuds absorbants :
(notation en vecteurs colonnes)

$$\hat{H} = H + w\frac{1}{n}\mathbf{1}^t.$$

où $w_i = \begin{cases} 1, & d_i = 0 \\ 0, & \text{sinon} \end{cases}$

Variante plus générale :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$ (vecteur stochastique).

Questions :

- Sous quelles hypothèses $x(k)$ converge ?
- Comment calculer la limite ?

Lien avec les chaînes de Markov (vision marche aléatoire)

Algorithme PageRank

Pour résoudre le problème de nœuds absorbants :
(notation en vecteurs colonnes)

$$\hat{H} = H + w\frac{1}{n}\mathbf{1}^t.$$

où $w_i = \begin{cases} 1, & d_i = 0 \\ 0, & \text{sinon} \end{cases}$

Variante plus générale :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$ (vecteur stochastique).

Questions :

- Sous quelles hypothèses $x(k)$ converge ?
- Comment calculer la limite ?

Lien avec les chaînes de Markov (vision marche aléatoire)
un marcheur dans un nœud i suit un lien sortant choisi selon la loi uniforme sur les liens sortants ; si pas de liens sortants alors il choisit le prochain nœud selon la loi v

Algorithme PageRank

Algorithme :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Algorithme PageRank

Algorithme :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Question : irréductible ? apériodique ?

Algorithme PageRank

Algorithme :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Question : irréductible ? apériodique ?

Pour résoudre ces problèmes :

$$G = \alpha\hat{H} + (1-\alpha)\frac{1}{n}\mathbf{1}\mathbf{1}^t.$$

avec $\hat{H} = H + wv^t$, pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Algorithme PageRank

Algorithme :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Question : irréductible ? apériodique ?

Pour résoudre ces problèmes :

$$G = \alpha\hat{H} + (1-\alpha)\frac{1}{n}\mathbf{1}\mathbf{1}^t.$$

avec $\hat{H} = H + wv^t$, pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Variante plus générale :

$$G = \alpha\hat{H} + (1-\alpha)\frac{1}{n}\mathbf{1}\hat{v}^t.$$

pour un vecteur $\hat{v} \geq 0,\ \hat{v}^t\mathbf{1} = \mathbf{1}^t$.

Algorithme PageRank

Algorithme :

$$\hat{H} = H + wv^t.$$

pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Question : irréductible ? apériodique ?

Pour résoudre ces problèmes :

$$G = \alpha\hat{H} + (1-\alpha)\frac{1}{n}\mathbf{1}\mathbf{1}^t.$$

avec $\hat{H} = H + wv^t$, pour un vecteur $v \geq 0,\ v^t\mathbf{1} = \mathbf{1}^t$.

Variante plus générale :

$$G = \alpha\hat{H} + (1-\alpha)\frac{1}{n}\mathbf{1}\hat{v}^t.$$

pour un vecteur $\hat{v} \geq 0,\ \hat{v}^t\mathbf{1} = \mathbf{1}^t$.

Généralisation : inclure pondérations sur les arcs.

Algorithme PageRank

PageRank : $G = \alpha \hat{H} + (1 - \alpha) \frac{1}{n} \mathbf{1} \hat{v}^t$.

Vision marche aléatoire

Algorithme PageRank

PageRank : $G = \alpha \hat{H} + (1-\alpha)\frac{1}{n}\mathbf{1}\hat{v}^t$.

Vision marche aléatoire

- avec probabilité α un marcheur dans un nœud i suit un lien sortant choisi selon la loi uniforme (ou une autre loi si pondérations) sur les liens sortants ; si pas de liens sortants alors il choisit le prochain nœud selon la loi v
- sinon il choisit un nouveau nœud selon la loi $\hat{v}$

Algorithme PageRank

PageRank : $G = \alpha \hat{H} + (1-\alpha)\frac{1}{n}\mathbf{1}\hat{v}^t$.

Vision marche aléatoire

- avec probabilité α un marcheur dans un nœud i suit un lien sortant choisi selon la loi uniforme (ou une autre loi si pondérations) sur les liens sortants ; si pas de liens sortants alors il choisit le prochain nœud selon la loi v
- sinon il choisit un nouveau nœud selon la loi $\hat{v}$

La limite (si elle existe) est la loi stationnaire de cette marche aléatoire.

Question : Pourquoi existe-t-elle ?