Fouille du web

Quentin Fournier <quentin.fournier@polymtl.ca>

http://infolab.stanford.edu/~ullman/mmds/book.pdf

Les diapositives ont été créées par Daniel Aloise <daniel.aloise@polymtl.ca>

13 juin 2020

Fouille de graphes

- Il y a deux types principaux d'applications pour lesquelles la fouille de graphes est naturelle :
  1. Dans des applications telles que les données chimiques et biologiques, une base de données de nombreux petits graphes est disponible.
  2. Dans les applications telles que le Web et les réseaux sociaux, un seul grand graphe est disponible.
- Dans cette séance, on va traiter le deuxième type d'application.

PageRank

calculer l'importance des pages en s'intéressant à la structure du web plutôt qu'au contenu des pages.

page bcp référecé (ex: Wikipedia)

- Une manière intuitive de savoir si une page est populaire consiste à compter le nombre d'hyperliens entrants (c.-à-d. qui pointent vers cette page).
- Supposons l'existence de deux pages A et B, chacune avec trois hyperliens entrants.
  - A est pointé par les sites de Pierre, Paul, et Jack.
  - B est pointé par LeMonde, LaPresse, et LeDevoir.
- Doit-on considérer que A est aussi "populaire" que B ?

Non, car on doit prendre en compte l'importance des sites de référence

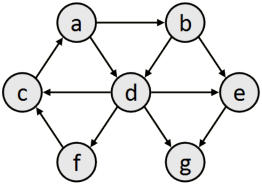
PageRank

les sommets: correspondent à un site internet
graphe orienté
les arcs entrants (c vers a) veut dire que le site c contient un lien qui pointe vers le site a

- PageRank était la sauce secrète originale derrière Google.
- Ignore le contenu textuel des pages Web, pour se concentrer uniquement sur la structure des hyperliens entre les pages.

Idée de base

```
while not converged:
 for each node v
   rang(v) = sum of the ranks of the
                  incoming hyperlinks
```



+ il y a des lien qui nous références, + on a d'importance

PageRank

Première amélioration

d'autres sites

- Si une page Y liée à la page X a des millions de liens sortants, cette connexion est moins importante que celle venant d'une page W liée à X avec seulement quelques liens sortants.

on divise de manière équitable

- Modification simple :

rang d'un site

$$rang(j) = \sum_{i \to j} \frac{rang(i)}{L(i)}$$

les sites i pointant le site j

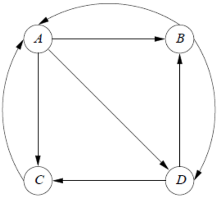
nb arcs sortants et se rendant à j

où $L(i)$ est le nombre d'arcs sortant de i.

Interprétation par algèbre linéaire

n: nb de noeuds dans notre graphe, ou le nb de sites webs

- Soit M la matrice ($n \times n$) de transition.
- M_{ij} est la probabilité de que la prochaine page visitée après la page j soit la page i.
- Alors $M_{ij} = 1/L(j)$ s'il existe un arc $j \to i$, $M_{ij} = 0$ sinon.
- Exemple :



Voir notes cahier

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

MATRICE SELON LES COLONNES!!

Interprétation par algèbre linéaire

n: nb de sites internet

- Le PageRank initiale de toutes les pages (PR_0) vaut $1/n$:
  - Un utilisateur peut commencer à surfer en principe à partir de n'importe quelle page du web.

- Le PageRank à la k-eme itération est alors donné par :

$$PR_k = M \cdot PR_{k-1}$$

Interprétation par algèbre linéaire

- Le PageRank initiale de toutes les pages (PR_0) vaut $1/n$:
  - Un utilisateur peut commencer à surfer en principe à partir de n'importe quelle page du web.
- Le PageRank à la k-eme itération est alors donné par :

$$PR_k = M \cdot PR_{k-1}$$

- L'algorithme (Markovian process) converge alors à condition que : faut respecter ces 2 conditions ci-dessous:
  - Le graphe soit fortement connexe : il existe un chemin entre chaque sommet.
  - M soit stochastique : $\sum_{i=1}^{n} M_{ij} = 1$.

    faut pas qu'on ait un cul de sac

!!! Pour toutes les lignes, chacune des colonnes doit sommer à 1 ! !!!!

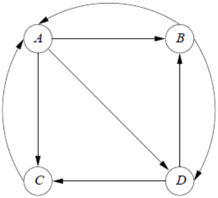
EXAMEN

enregistrement 15 min

Exemple

3 itérations, il nous donne le graphe, chacun des sites est équiprobable,

- Considérons le graphe ci-dessous et sa matrice M



$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- En partant de $PR_0 = [\frac{1}{4} \frac{1}{4} \frac{1}{4} \frac{1}{4}]^T$, on obtient à partir de l'itération de $PR_j = M \cdot PR_{j-1}$

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}, \begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}, \ldots, \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

PR0 PR1 PR2 PR3

Structure du web

- On pourrait penser que :
  - Le web est un ensemble de sites indépendants.
  - L'information peut se trouver n'importe où !

Structure du web

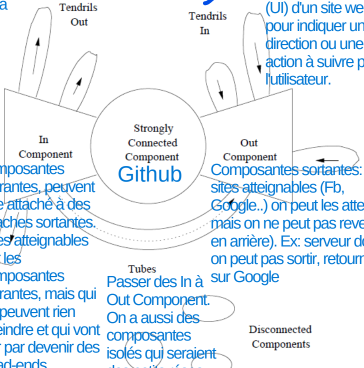
- On pourrait penser que :
  - Le web est un ensemble de sites indépendants.
  - L'information peut se trouver n'importe où !
- C'est faux...
  - Le web est un ensemble de sites interconnectés.
  - L'information se trouve sur les sites plus fortement connectés.

EXAMEN:
DONNER STRUCURE DU WEB ET EXPLIQUER SES COMPOSANTES EXAMEN!!!
Enregistrement environ 24 min et prendre + de notes

- Sites fortement connexes (SFC) : pages les plus intéressants. pages qui peuvent être atteintes des autres pages à l'intérieur des SFC. 1 site peut atteindre les autres sites
- Composantes entrantes (CE) : pages qui peuvent atteindre les SFC. mais pas atteignables par les SFC
- Composantes sortantes (CS) : pages qui peuvent être atteints par les SFC mais ne peuvent pas les atteindre.
- Attaches sortantes : pages atteints des CE qui ne peuvent pas atteindre les SFC.
- Attaches entrantes : pages qui atteignent les CS mais ne peuvent pas atteindre les SFC.
- Tubes : pages liant les CE et les CS qui sautent les SFC.
- Composantes isolées : ne peuvent pas atteindre ni être atteintes du reste du réseau.

liens qui pointent vers d'autres sites depuis un site web donné. Utilisés pour fournir des références ou des sources supplémentaires d'information sur un sujet ou pour aider à orienter les utilisateurs vers d'autres ressources en ligne.

utilisés dans l'interface utilisateur (UI) d'un site web pour indiquer une direction ou une action à suivre pour l'utilisateur.



Github

composantes entrantes, peuvent être attaché à des attaches sortantes. Sites atteignables par les composantes entrantes, mais qui ne peuvent rien atteindre et qui vont finir par devenir des dead-ends

Composantes sortantes: sites atteignables (Fb, Google..) on peut les atteindre mais on ne peut pas revenir en arrière). Ex: serveur dont on peut pas sortir, retourner sur Google

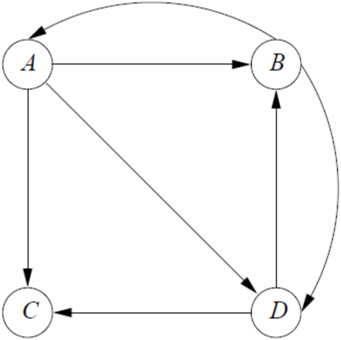
Passer des In à Out Component. On a aussi des composantes isolés qui seraient des petits réseaux internets pas accèssibles à l'extérieuret qui ne peut pas accèder à l'extérieur

Structure du web

Voir notes de cours

- Plusieurs de ces structures violent les hypothèses nécessaires pour que l'itération de PageRank converge vers une limite :
  - Les usagers ne peuvent pas sortir des CS (dead-end). site qu'on est bloqué et qu'on ne peut pas retourner en arrière
  - Les usagers vont certainement finir dans les CS ou les attaches sortantes.
- Conséquence : les PageRanks des SFC ou des CE seront nuls à la fin.

Dead-end



$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 3/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 5/48 \\ 7/48 \\ 7/48 \\ 7/48 \end{bmatrix}, \begin{bmatrix} 21/288 \\ 31/288 \\ 31/288 \\ 31/288 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

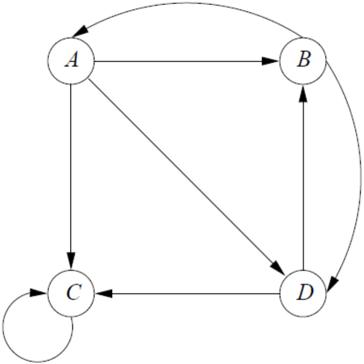
- La probabilité qu'un internaute finisse de surfer à une page quelconque du web passe à zéro au fil du temps.
- M n'est pas stochastique.

processus ou à un événement qui est aléatoire ou incertain. Cela signifie que le résultat d'un processus stochastique ne peut pas être prédit avec certitude et varie chaque fois qu'il est exécuté.

Enregistrement 38min

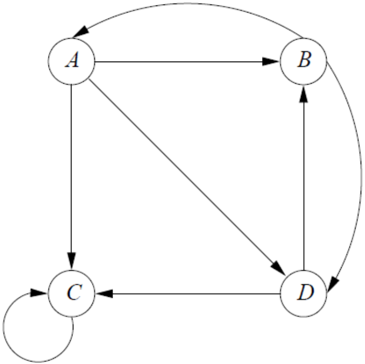
Spider Traps

- L'exemple précédent montre le souci d'un cul-de-sac pour l'algorithme PageRank.
- Et si quelqu'un essayait de capturer notre internaute à la page C ?



POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Spider Traps



$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 3/24 \\ 5/24 \\ 11/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 5/48 \\ 7/48 \\ 29/48 \\ 7/48 \end{bmatrix}, \begin{bmatrix} 21/288 \\ 31/288 \\ 205/288 \\ 31/288 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

- La probabilité qu'un internaute finisse de surfer à la page C est de 100% au fil du temps !
- M est stocastique par contre

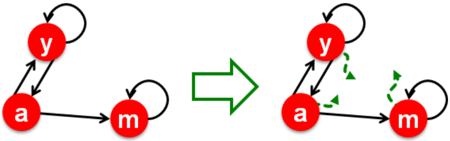
$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Voir notes de cours

Solution : téléportation

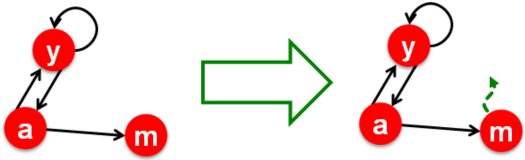
- La solution de Google pour les *spider traps* : à chaque pas de temps, l'internaute a deux options :
  - avec prob. β, il suit un lien au hasard
  - avec prob. $1-\beta$, il saute à une page quelconque du web
- Les internautes sortent ainsi des *spider traps* dans quelques itérations
- On utilise souvent $\beta \in [0.8, 0.9]$



Source : J. Leskovec, A. Rajaraman, J. Ullman : Mining of Massive Datasets

Solution : téléportation

- Pour le culs-de-sac on saute à une page quelconque avec prob. 1.
- Nous rendons M stochastique.



	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

	y	a	m
y	½	½	⅓
a	½	0	⅓
m	0	½	⅓

source : J. Leskovec, A. Rajaraman, J. Ullman : Mining of Massive Datasets

PageRank

- Équation PageRank [Brin & Page, 1998]

i: sites permettant d'atteindre le site j

$$rang(j) = \sum_{i \to j} \beta \frac{rang(i)}{L(i)} + (1 - \beta)\frac{1}{n}$$

n sites sur le web

L(i): nb arcs sortant de i

proba que je me téléporte que je me déplace dans un site web

Beta: proba de suivre un lien entre 0.8 et 0.9

- Formule algébrique :

$$PR_k = \beta M \cdot PR_{k-1} + (1 - \beta)\mathrm{e}/n$$

e = [1,1,1....1]^T

où e est le vecteur unitaire

EXAMEN FINAL

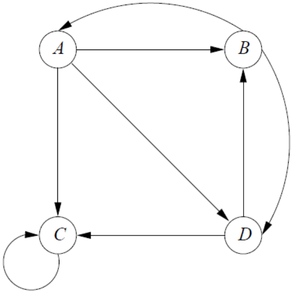
Exemple

Prof a dit une question sur PageRank au final
celle là était dans un ancien final

VOIR NOTES PRISES, C'EST SUR QUE C'EST À L'EXAM

REFAIRE POUR SE PRATIQUER
3 itérations, donc jusqu'à PR3 et COMMENTER LE RÉSULTAT
FEUILLE DE NOTES AUSSI

- Considérons le graphe :



avec $\beta = 0.8$

Exemple

- L'itération de PageRank est donnée par :

$$\mathbf{v}' = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \mathbf{v} + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

- Les premières itérations sont :

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix}, \begin{bmatrix} 41/300 \\ 53/300 \\ 153/300 \\ 53/300 \end{bmatrix}, \begin{bmatrix} 543/4500 \\ 707/4500 \\ 2543/4500 \\ 707/4500 \end{bmatrix}, \dots, \begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix}$$

- Remarquez que l'effet du *spider trap* a été limité.

PageRank

- Comment Google marche (version résumé) ?
- Google ne divulgue plus sa sauce secrète depuis 2016
- Comment Google est devenu un empire ?

Exercice typique examen
enregistrement 1h 03 environ

Exercice

- Obtenez la plus grande valeur de $rang(v)$ pour le graphe ci-dessous après trois itérations de *PageRank*
- Initialisez $rang(v) = 1/|n|$ pour tous les sommets
- Utilisez $\beta = 0.6$

Cet exo est sur moodle avec le code numpy

Très important, à faire!