UCLouvain

PageRank

Classement de réseaux socieaux et pages web

LINFO1114 - Mathématiques discrètes

Alexandre DEWILDE
Brieuc DUBOIS
Theo TECHNICGUY

Étudiants

Marco SAERENS

Professeur

Sylvain COURTAIN
Pierre LELEUX

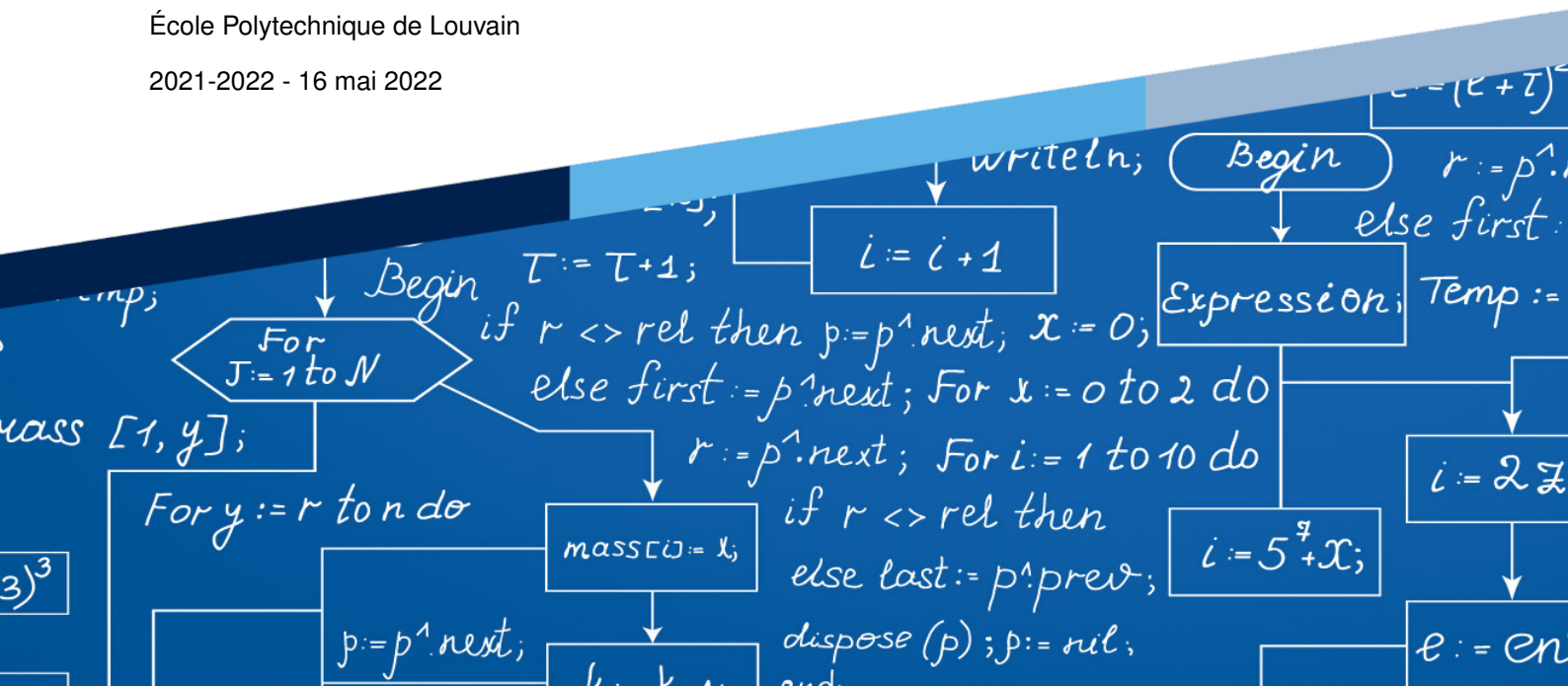
Assistants

SINF1BA · Groupe 42

Université Catholique de Louvain

École Polytechnique de Louvain

2021-2022 - 16 mai 2022



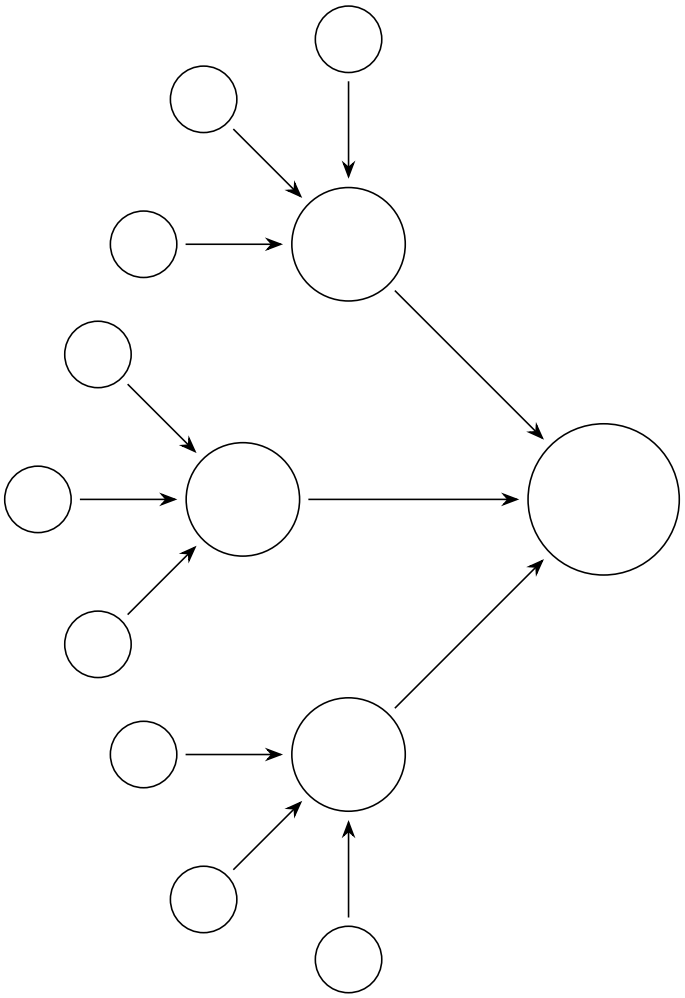
1 Introduction

Le cours *LINFO1114* est composé d'un projet : une implémentation de l'algorithme dit *de Google* ou *PageRank*. Il s'agit d'un programme permettant la classification et la quantification de l'*importance* des pages internet, crée par Larry PAGE. Cet algorithme est partiellement utilisé par le moteur de recherche *Google*[1].

2 Rappel Théorique

PageRank est basé sur l'hypothèse qu'une page importante serait référence plus souvent qu'une page moins importante. Si l'on se limite à uniquement cette définition, il serait facile d'artificiellement augmenter l'importance d'une page en créant des pages vide de véritable contenu. En ajoutant la contrainte de l'importance de la page référente, nous mitigeons ce facteur.

Une page plus importante est une page qui est référencé par beaucoup de pages importante.



Le graphe dirigé ci-dessus montre une potentielle relation entre sites. Plus le nœud est grand, plus il est important. Mathématiquement, cela est équivalent à dire :

1. http://www.google.com

$$x_i \propto \sum_{j=0}^{n} \frac{w_{ji} * x_j}{w_{j.}} \tag{1}$$

où

— x_i est la valeur PageRank de la page i

— n le nombre de pages référant vers la page i

— w_{ji} l'élément en position ji de la matrice W

— x_j la valeur PageRank de la page référente j

$$w_{j.} = \sum_{i=0}^{n} w_{ji} \tag{2}$$

Notons que cette définition est récursive. Il faut une page importante pour en créer une autre. La question se pose alors *comment déterminer une page importante sans avoir de référence initiale ?*. Introduisons un nouveau concept : le marcheur aléatoire.

Le marcheur aléatoire est un programme qui visite les pages sur la toile et suit un des liens inclus dans la page. Pendant ce procédé, il tient compte du nombre de fois qu'il a visité cette page. Plus il a visité la page, plus il est probable qu'elle est importante ; c'est la définition initiale.

Étant donné que pas toutes les pages de la toile sont reliés, le marcheur aléatoire pourrait ne visiter qu'une facette d'Internet, sans visiter d'autres sites. Pour remédier à ce problème, un paramètre de téléportation est introduit dans le calcule (section 3).

$$P(page(k+i) = i | page(k) = j) = \frac{w_{ji}}{w_{j.}} \tag{3}$$

L'évolution de cette équation, en fonction du temps $x_i(k)$ se réécris comme

$$x_i(k+i) = \sum_{j=1}^{n} p_{ji} * x_j(k) \tag{4}$$

Avant de déterminer le classement des sites, il est possible de biaiser les résultats en introduisant un vecteur de personnalisation permettant de valoriser certains sites par rapport aux autres.

Finalement, il faut, pour la matrice *de Google*, il faut résoudre

$$G = \alpha P + (1 - \alpha) e v^T \tag{5}$$

avec

— α le paramètre de téléportation

— e la matrice identité

— v le vecteur de personnalisation

3 Paramètre de téléportation

Le paramètre de téléportation, noté α, est un paramètre qui décide de l'avancée du marcheur aléatoire. Plus le paramètre, limité entre 0 et 1, tends vers 1, plus le marcheur aléatoire a de chance d'effectivement rejoindre la destination du lien indiqué. α prend comme valeur par défaut 0.9, soi-disant que dans 90% des cas, le marcheur arrive à destination et dans 10% des cas, il se téléporte vers une autre page.

Si l'on diminue ce paramètre, la probabilité que le marcheur se téléporte augmente. Il aura donc moins de chance de parcourir les sites interconnectés par des liens, et donc d'*achever* la visitation de tous les sites d'une partie du réseau.

4 Vecteur de personnalisation

Le vecteur de personnalisation est un vecteur de même taille que la matrice d'adjacence. Il permet d'ajuster légèrement les scores obtenus pour favoriser une page web par rapport à une autre. Plus les valeurs, limitées entre 0 et 1, tendent vers 1, plus la page va être favorisé dans les résultats finaux. Il faut aussi noter que la somme du vecteur doit valoir 1.

Si l'on diminue les valeurs du vecteur, la page perd en valeur et sa position dans le classement final chute.

Dans le cadre du projet, ce vecteur nous a été assigné affin de différencier les valeurs de réponse.

5 Matrice de Probabilité

6 Système linéaire

Comme rappeler dans les sections précédentes, le système à résoudre est de cette forme :

$$(I - \alpha P)^T x = (1 - \alpha)v \tag{6}$$

La matrice I étant la matrice identité, P la matrice de probabilité, V le vecteur de personnalisation et α le paramètre de téléportation.

Pour résoudre le système, on commence par calculer le membre de gauche :

$$
\begin{aligned}
A &= (I - \alpha P)^T \\
&= \begin{pmatrix}
1.0 & -0.27 & 0.0 & -0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
-0.5625 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & -0.18 & 1.0 & 0.0 & -0.6429 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
-0.3375 & 0.0 & -0.5143 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.27 & 0.0 \\
0.0 & -0.09 & 0.0 & 0.0 & 1.0 & 0.0 & -0.45 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & -0.3857 & -0.6 & 0.0 & 1.0 & 0.0 & -0.18 & -0.45 & 0.0 \\
0.0 & -0.36 & 0.0 & 0.0 & -0.2571 & 0.0 & 1.0 & 0.0 & 0.0 & -0.5 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.9 & -0.45 & 1.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & -0.4 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.72 & -0.18 & 1.0
\end{pmatrix}
\end{aligned}
$$

Ensuite, on calcule le membre de droite :

$$
\begin{aligned}
B &= (1 - \alpha)v \\
&= 0.1 \begin{pmatrix} 0.1019 & 0.043 & 0.1877 & 0.1828 & 0.0439 & 0.1123 & 0.0596 & 0.1941 & 0.0476 & 0.0271 \end{pmatrix}^T \\
&= \begin{pmatrix} 0.0102 & 0.0043 & 0.0188 & 0.0183 & 0.0044 & 0.0112 & 0.006 & 0.0194 & 0.0048 & 0.0027 \end{pmatrix}^T
\end{aligned}
$$

Ce qui nous donne le systeme suivant à résoudre :

$$
\begin{pmatrix}
1.0 & -0.27 & 0.0 & -0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
-0.5625 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & -0.18 & 1.0 & 0.0 & -0.6429 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
-0.3375 & 0.0 & -0.5143 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.27 & 0.0 \\
0.0 & -0.09 & 0.0 & 0.0 & 1.0 & 0.0 & -0.45 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & -0.3857 & -0.6 & 0.0 & 1.0 & 0.0 & -0.18 & -0.45 & 0.0 \\
0.0 & -0.36 & 0.0 & 0.0 & -0.2571 & 0.0 & 1.0 & 0.0 & 0.0 & -0.5 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.9 & -0.45 & 1.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & -0.4 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & -0.72 & -0.18 & 1.0
\end{pmatrix}
x =
\begin{pmatrix}
0.0102 \\ 0.0043 \\ 0.0188 \\ 0.0183 \\ 0.0044 \\ 0.0112 \\ 0.006 \\ 0.0194 \\ 0.0048 \\ 0.0027
\end{pmatrix}
$$

La résolution de ce systeme donne :

$$
x = \begin{pmatrix} 0.0431 & 0.0286 & 0.0618 & 0.0841 & 0.0589 & 0.156 & 0.1155 & 0.2118 & 0.072 & 0.1682 \end{pmatrix}^T
$$

La matrice P donnée en entrée étant déjà normalisée, cette sortie l'est également.

7 Power method

Comme mentionné dans le rappel théorique, le calcul du score *PageRank* via la power method est un procédé itératif, où le résultat est obtenu par convergence.

En partant de la matrice d'adjance A suivante :

$$A = \begin{pmatrix} 0 & 5 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 2 & 0 & 1 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 3 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 3 & 0 & 5 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 5 & 0 & 4 & 0 \end{pmatrix}$$

On peut calculer la matrice de probabilité de transition P

$P = \frac{A}{A_j}$ Où A_j contient la somme des lignes de A

$$= \begin{pmatrix} 0.0 & 0.625 & 0.0 & 0.375 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.0 & 0.2 & 0.0 & 0.1 & 0.0 & 0.4 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.5714 & 0.0 & 0.4286 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.3333 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6667 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.7143 & 0.0 & 0.0 & 0.0 & 0.2857 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.0 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.0 & 0.0 & 0.3 & 0.0 & 0.5 & 0.0 & 0.0 & 0.0 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.5556 & 0.0 & 0.4444 & 0.0 \end{pmatrix}$$

On calcule ensuite la matrice Google G

$$G = \alpha P + (1 - \alpha)v$$

$$= \begin{pmatrix} 0.0102 & 0.5668 & 0.0188 & 0.3558 & 0.0044 & 0.0112 & 0.006 & 0.0194 & 0.0048 & 0.0027 \\ 0.2802 & 0.0043 & 0.1988 & 0.0183 & 0.0944 & 0.0112 & 0.366 & 0.0194 & 0.0048 & 0.0027 \\ 0.0102 & 0.0043 & 0.0188 & 0.5326 & 0.0044 & 0.3969 & 0.006 & 0.0194 & 0.0048 & 0.0027 \\ 0.3102 & 0.0043 & 0.0188 & 0.0183 & 0.0044 & 0.6112 & 0.006 & 0.0194 & 0.0048 & 0.0027 \\ 0.0102 & 0.0043 & 0.6616 & 0.0183 & 0.0044 & 0.0112 & 0.2631 & 0.0194 & 0.0048 & 0.0027 \\ 0.0102 & 0.0043 & 0.0188 & 0.0183 & 0.0044 & 0.0112 & 0.006 & 0.9194 & 0.0048 & 0.0027 \\ 0.0102 & 0.0043 & 0.0188 & 0.0183 & 0.4544 & 0.0112 & 0.006 & 0.4694 & 0.0048 & 0.0027 \\ 0.0102 & 0.0043 & 0.0188 & 0.0183 & 0.0044 & 0.1912 & 0.006 & 0.0194 & 0.0048 & 0.7227 \\ 0.0102 & 0.0043 & 0.0188 & 0.2883 & 0.0044 & 0.4612 & 0.006 & 0.0194 & 0.0048 & 0.1827 \\ 0.0102 & 0.0043 & 0.0188 & 0.0183 & 0.0044 & 0.0112 & 0.506 & 0.0194 & 0.4048 & 0.0027 \end{pmatrix}$$

Nous pouvons utiliser ce résultat pour calculer le score *PageRank* à l'aide de la power method. Pour cela, nous initialisons chaque valeur du vecteur de scores X à $\frac{1}{|X|}$, de telle sorte que le vecteur X soit normalisé. Nous pouvons ensuite calculer X par récurrence.

$$
\begin{aligned}
X_0 &= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}^T \\
X_1 = G^T \cdot X_0 &= \begin{pmatrix} 0.0672 & 0.0606 & 0.1011 & 0.1305 & 0.0584 & 0.1728 & 0.1177 & 0.1544 & 0.0448 & 0.0927 \end{pmatrix}^T \\
X_2 = G^T \cdot X_1 &= \begin{pmatrix} 0.0657 & 0.0421 & 0.0672 & 0.105 & 0.0628 & 0.1764 & 0.0891 & 0.2279 & 0.0418 & 0.1219 \end{pmatrix}^T \\
X_3 = G^T \cdot X_2 &= \begin{pmatrix} 0.0531 & 0.0412 & 0.0667 & 0.0863 & 0.0483 & 0.16 & 0.0982 & 0.2183 & 0.0535 & 0.1743 \end{pmatrix}^T \\
\vdots \qquad & \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
X = G^T \cdot X_{-1}^T &= \begin{pmatrix} 0.0431 & 0.0286 & 0.0618 & 0.0841 & 0.0589 & 0.156 & 0.1155 & 0.2118 & 0.072 & 0.1682 \end{pmatrix}^T
\end{aligned}
$$

Nous pouvons remarquer que nous obtenons exactement les même résultats qu'avec le système d'équations linéaires, ce qui est en tout point cohérent.