M1 Info - Ranking et Reommandation

Cours 1 : Introduction

Franck.Quessette@uvsq.fr

UVSQ université PARIS-SACLAY
UFR des Sciences
CAMPUS DE VERSAILLES

Version du 13 février 2025

Organisation de l'UE

- Partie 0 : **Chaînes de Markov**, 2 CM+TD, commun avec Simulation ;
- Partie 1 : **Ranking**, 3 CM+TD, Franck Quessette ;
- Partie 2 : **Recommandation**, 3 CM+TD, Sandrine Vial ;
- CC : 1 Projet ;
- 1 Examen final.

Plan de la Partie 1 : Ranking

Introduction

Ranking

Problématique

- Donner un ordre aux pages du web pour répondre à une requête.
- Comment traiter en temps réel sachant la taille du Web ?
- Comment donner une pertinence à une page ?

Notions abordées

- L'approche de PageRank.
- Gestion de grands objets en mémoire (matrice creuse).
- Convergence de PageRank.
- Valeurs Propres, Vecteurs Propres.
- Les alternatives à PageRank (SALSA, HITS).
- Agir sur PageRank.

Recommandation

Problématique

- Connaissant vos achats dans le passé ainsi que tous les achats de tous les clients, comment vous proposer de nouveaux achats pour augmenter le chiffre d'affaires ?
- Plus sophistiqué : connaissant les notes de vos achats par le passé ainsi que toutes les notes données par les clients, comment vous proposer de nouveaux achats pour augmenter le chiffre d'affaires ?
- Même questions avec les contenus que vous regardez et que l'on vous propose.

Notions abordées

- Similarité de clients et similarité d'objets.
- Calcul d'une note virtuelle.
- Singular Value Decomposition.

Notions pas abordées

On ne fera pas

- La gestion des robots qui indexent le web.
- Les techniques pour que votre page soit bien traitée
- Les questions liées au langage naturel.
- La détection de communautés.
- L'algorithmique du texte.

Ranking

Hypothèses de départ

Données d'entrée

Le WEB est visité par des robots qui indexent :

- les textes avec leurs propriétés typographiques ;
- les URL (liens) ;
- les images ;
- les vidéos.

Modélisation avec deux objets :

- le graphe du WEB ;
- le dictionnaire des mots avec une pondération qui indique leur importance pour la page.

Graphe du WEB

Graphe du WEB :

- Les nœuds sont les pages.
- Les arcs sont les URL entre les pages. Comme il peut y avoir plusieurs URL entre deux pages, il peut y avoir plusieurs arcs de la page i vers la page j : c'est un multigraphe et non pas un graphe.

$\Longrightarrow$ **Multigraphe orienté**

Dictionnaire :

- Pour chaque page, on collecte les mots présents dans la page.
- On ajoute les synonymes (voiture = automobile).
- On ajoute une note qui prend en compte l'importance dans la page :
  - nombre d'occurrences dans la page ;
  - attributs typographiques (couleur, fonte, gras, souligné, italique) ;
  - place dans le texte (titre, en-tête de paragraphe).

  Le calcul de cette note est peu documenté, il y a des conseils de Google pour qu'une page soit bien notée.
- Pour chaque mot du dictionnaire, on a donc la liste des pages WEB avec la note de ce mot pour cette page.
- La structure de données stockant le dictionnaire doit être très efficace (Algorithmique du texte).

Aspects temporel

Mises à jour du graphe et du dictionnaire :

- L'indexation du web est faite en continu.
- Le graphe est construit périodiquement (tous les mois ?) pour en tirer les notes de pertinence.
- Le dictionnaire peut être mis à jour plus souvent. Le contenu des pages bouge plus que les liens entre les pages.
- Attention : la construction du dictionnaire permet de censurer très vite et de faire disparaitre une partie du Web.

Construction de la matrice associée au graphe du WEB

Graphe du WEB stocké sous forme d'une matrice

- Chaque indice de ligne/colonne correspond à un sommet du graphe donc à une page WEB.
- Chaque entrée de la matrice correspond aux nombre de liens d'une page vers l'autre.

Notons A cette matrice et i, j des indices de cette matrice. On a :

- $A[i,j]$ nombre d'arcs de i vers j.
- A est un matrice positive : $\forall i,j,\ A[i,j] \geq 0$.

Attention en anglais "Positive matrix" signifie $\forall i,j,\ A[i,j] > 0$.
Idem pour "negative matrix".
Donc ici on dira en anglais "non negative matrix" pour dire ≥ 0.

On veut faire quoi 1/2

Question de base : Pour un mot donné lister les pages du WEB en ordre décroissant de notes

Difficulté : Taille du graphe (de la matrice) de l'ordre de 10^{10}.
Problème de stockage et de temps de calcul.
Sur Google environ 10^5 requêtes par secondes, 8,6 milliards par jour.

Il faut tenir compte de la « pertinence des pages » et aller plus vite qu'un algo naïf qui trie toutes les pages.

On veut faire quoi 2/2

Principe : Calculer deux notes :

- Une **note de contenu** qui repose sur la requête et le dictionnaire : rapide.
- Une **note de pertinence de page** qui repose sur la pertinence de la page sur le web et qui se calcule avec le graphe du web : lent.
- Combinaison des deux notes. A priori multiplicative pour éliminer les pages proches de zéro sur une des deux notes.

Note de contenu

- Pour chaque mot de la requête accéder au dictionnaire.
- Combiner les notes de chaque mot, a priori une somme.
- Utilisation de synonymes ou de correction orthographiques.
- Éventuellement requêtes en éliminant un mot.

Complexité mieux que linéaire dans la taille du dictionnaire.

Note de pertinence de page

Ce que l'on veut

- On voudrait une note par page liée à l'«expertise» de l'auteur.
- L'«expertise» est reconnue par les pairs (même idée qu'en bibliographie).
- Les mots de la requête ne sont pas pris en compte.
- Cette note n'est pas recalculée en temps réel.
- Cette note tient compte des avis de tous les auteurs à propos de toutes les pages Web (au moins quadratique sur la taille du web).

Complexité quadratique dans la taille du web mais c'est trop si on le fait en force brute.

Note de de pertinence de page

Ce qui est fait

2 axiomes :

- Une page avec un certain niveau de pertinence **pointe vers** des pages de niveau de pertinence équivalent.
- Une page avec un certain niveau de pertinence **est pointée par** des pages de niveau de pertinence équivalent.

3 autres idées sur les combinaisons de pertinence :

- Chaque page distribue de la pertinence et en reçoit (équilibre).
- Hypothèse 1 : La pertinence d'une page est divisée équitablement sur les pages vers lesquelles elle pointe.
- Hypothèse 2 : La pertinence d'une page est la somme des parts de pertinence qui pointe vers elle.

Note de pertinence page

Critique

- Toute citation est positive.
- Dire qu'une page est nulle ou qu'elle contient des erreurs tout en pointant vers elle, renforce la pertinence de cette page (pour PageRank).
- Il n'y a pas d'analyse sémantique du contenu.
- La pertinence est additive : beaucoup de soutiens de pages peu pertinentes peut être plus valorisant qu'un avis positif de l'expert du domaine. C'est un effet réseaux sociaux.
  C'est le principe qui permet le «Google bombing»
  https://fr.wikipedia.org/wiki/Bombardement_Google.

Note de pertinence de page

Implémentation

- La pertinence de la page i est un réel dans $[0;1]$, notée $\pi(i)$.
- On construit la matrice P de distribution de pertinence (P est une matrice de transition).

Construction de P avec l'hypothèse 1

Soit $d^+(i)$ le nombre de liens présents dans la page i.

$$\text{Si } d^+(i) > 0, P[i,j] = \frac{A[i,j]}{d^+(i)} \text{ sinon } P[i,j] = 0.$$

Formule de calcul de la pertinence avec l'hypothèse 2

$$\forall j,\ \pi(j) = \sum_i \pi(i) P[i,j]$$

Note de pertinence de page

Implémentation 1 En notation matricielle, on a

$$\pi = \pi P$$

Le problème se ramène à calculer π.

Dans un premier temps, on va supposer que $\forall i, d^+(i) > 0$.

Problème classique d'algèbre linéaire. Mais la taille de P est très grande.

Algèbre Linéaire

Matrice

Rappels

- On ne considère que des matrices positives (en anglais «non negative matrices»).
- Une mattrice A de taille $N \times M$ possède N lignes et M colonnes.
- $A[i, j]$ est l'élément de la matrice A à l'intersection de la $i$ème ligne et de la $j$ème colonne.
- Une matrice est dite carrée si elle a autant de lignes que de colonnes. On dit que A est de taille $N \times N$ ou parfois A carrée de taille N ou encore A de taille N si avec le contexte on sait que A est carrée.

Somme matricielle

Soient A, B et C trois matrices de même tailles $N \times M$, $C = A + B$ est définie par :

$$\forall i \in 1..N,\ \forall j \in 1..M,\quad C[i,j] = A[i,j] + B[i,j]$$

- L'addition est commutative : $A + B = B + A$.
- L'addition est associative $(A + B) + C = A + (B + C)$.

Produit matriciel

Soient A une matrice de taille $N \times K$, B une matrice de taille $K \times M$ et C une matrice de taille $N \times M$ $C = AB = A \times B$ est définie par :

$$\forall i \in 1..N,\ \forall j \in 1..M,\quad C[i,j] = \sum_{k=1}^{K} A[i,k] \times B[k,j]$$

Attention à la compatibilité des tailles.

- La multiplication n'est en général pas commutative, à cause des tailles
- Même pour les matrices carrées, en général $AB \neq BA$.
- La multiplication (avec des tailles compatibles) est associative : $(AB)C = A(BC)$.

Vecteur ligne, vecteur colonne

Un vecteur est une matrice avec une de ses dimension égale à un :

- Une matrice u de dimension $1 \times N$ est appelé un **vecteur ligne** de taille N. Exemple

$$u = (1, 3, 0, 6)$$

- Une matrice v de dimension $N \times 1$ est appelé un **vecteur colonne** de taille N. Exemple

$$v = \begin{pmatrix} 1 \\ 0 \\ 8 \end{pmatrix}$$

- Quand on dit simplement vecteur, c'est que le contexte permet de savoir si c'est un vecteur ligne ou un vecteur colonne.

Éléments neutres

Matrice nulle

La matrice nulle ne contient que des 0, c'est l'élément neutre de l'addition matricielle.

Matrice Identité

La matrice identité I ou Id ou (Id_N pour préciser la taille) est définie pour les matrices carrées, c'est l'élément neutre pour la multiplication :

$$\begin{cases} \forall i \in 1..N, \quad Id_N[i,i] = 1 \\ \forall i \in 1..N, \ \forall j \in 1..N, \ j \neq i, \quad Id_N[i,j] = 0 \end{cases}$$

Id est l'élément neutre à gauche et à droite : pour toute matrice carrée A de taille $N \times N$ on a :

$$Id \times A = A \times Id = A$$

Transposée, diagonale, symétrique

Matrice transposée

Soit A une matrice de taille $N \times M$ la **tranposée de** A, notée A^t est une matrice de taille $M \times N$ (attention à l'inversion) telle que :

$$\forall i \in 1..M,\ \forall j \in 1..N,\quad A^t[i,j] = A[j,i]$$

- Pour A et B de tailles compatibles :

$$(AB)^t = B^t A^t$$

- Une matrice carrée égale à sa transposée est dite **matrice symétrique**.
- Pour une matrice carrée A, les éléments $A[i,i]$ sont appelés la diagonale de A.
- Une matrice carrée avec uniquement des valeurs non nulles sur sa diagonale est appelée **matrice diagonale**.
- Une matrice diagonale est égale à sa transposée.
- La matrice Id est diagonale.

Produit vecteur-matrice, matrice-vecteur

Si u est un vecteur ligne de taille N et v est un vecteur colonne de taille N

- uv est un scalaire (réel) (chercher la définition de scalaire).
- vu est une matrice carrée de taille N :

$$\forall i \in 1..N,\ \forall j \in 1..N,\quad (vu)[i,j] = v[i]u[j]$$

Pour toute matrice A de taille $N \times M$ et tout scalaire $\lambda \in \mathbb{R}$, λA est la matrice de taille $N \times M$ définie par

$$\forall i \in 1..N,\ \forall j \in 1..N,\quad (\lambda A)[i,j] = \lambda \times A[i,j]$$

Inverse d'une matrice

On se restreint ici aux matrices carrées.

Une matrice A n'est pas toujours inversible. Quand elle l'est on note A^{-1} son inverse.

- Pour toute matrice A, son inverse A^{-1} existe, s'il satisfait :

$$AA^{-1} = A^{-1}A = Id$$

- Une matrice A est **non singulière** si son inverse A^{-1} existe.
- Si A et B sont non singulières alors

$$(AB)^{-1} = B^{-1}A^{-1}$$

- Le calcul de l'inverse d'une matrice de taille N est en $\mathcal{O}(N^3)$ sans algo spécifique.

Valeurs propres – Vecteurs propres

Soit P une matrice carrée, $\lambda \in \mathbb{R}$ et x vecteur ligne sont respectivement **valeur propre** et **vecteur propre à gauche** si

$$xP = \lambda x$$

Soit P une matrice carrée, $\lambda \in \mathbb{R}$ et y vecteur colonne sont respectivement **valeur propre** et **vecteur propre à droite** si

$$Py = \lambda y$$

- Si P est de taille N, il y a N valeurs propres et N vecteurs propres.
- Les valeurs propres sont les mêmes à gauche et à droite.
- Les vecteurs propres à gauche et à droite sont en général différents.

Calcul des Valeurs propres

Soit A une matrice carrée, les valeurs propres sont les racines du polynome en λ, poly(λ) defini par :

$$\text{poly}(\lambda) = \det(A - \lambda Id)$$

où $\det(X)$ est le déterminant de la matrice X.

- Le calcul des valeurs propres est généralement difficile.
- Il existe des algorithmes itératifs.
- L'ensemble (multi-ensemble) des valeurs propres est appelé le **spectre**, commande spec de Scilab.
- Les valeurs propres peuvent être approchées par les cercles de Geshgorin.
- Les valeurs propres peuvent être des nombres complexes même si la matrice ne contient que des nombres réels.

Matrice stochastique ou matrice de transition

Définition

Une matrice carrée P est une matrice stochastique si :

$$\begin{cases} \forall i,j, \quad P[i,j] \geq 0 \\ \forall i, \quad \sum_j P[i,j] = 1 \end{cases}$$

Propriétés

- $\forall i,j, \quad P[i,j] \leq 1$
- Si P et R sont stochastiques et $\alpha \in \mathbb{R}$ avec $0 \leq \alpha \leq 1$ alors $\alpha P + (1-\alpha)R$ est stochastique.
- Si P et R sont stochastiques alors $P \times R$ est stochastique.

Matrice irreductible, matrice primitive

Définition

Une matrice stochastique P est **irréductible** si le graphe associé est fortement connexe.

Définition

Une matrice stochastique P est **primitive** si il existe une puissance n de P telle que P^n est strictement positive :

$$\exists n,\ \forall i,j,\quad P^n[i,j] > 0$$

Remarque : Si P^n est strictement positive alors pour tout $m \geq n$ P^m est strictement positive.

Comment calculer π ?

Étant donné une matrice stochastique P, on veut calculer π tel que

$$\begin{cases} \pi P = \pi \\ \pi e = 1 \end{cases}$$

Méthode des puissances

- Choisir un vecteur $\pi^{(0)}$
- Itérer $\pi^{(k+1)} = \pi^{(k)} P$
- Jusqu'à $||\pi^{(k+1)} - \pi^{(k)}|| < \varepsilon$

Comment calculer π ?

Problèmes

- Comment stocker de très grandes matrices très creuses ?
- Comment faire la multiplication vecteur matrice avec une structure de données creuse ?

$\Longrightarrow$ Réponse en TD.

Pagerank