Chapitre 4 (suite) :
Le modèle LSI (Latent Semantic Indexing)
Indexation sémantique Latente

1

1. Introduction

La problématique dans les modèles de RI basés sur les termes, est que les documents **pertinents** n'ayant aucun terme de la requête n'ont aucune chance d'être sélectionnés. Cependant, ces documents peuvent contenir des termes synonymes à ceux de la requête (exemple : voiture/automobile, élève/étudiant, livre/ouvrage, …etc.)

Le modèle LSI, propose une solution pour cette problématique, en regroupant les termes sémantiquement reliés (co-occurrence), dans un même concept et faire une recherche par concept.

Ce modèle propose alors :

- ✓ Une indexation par concept selon la co-occurrence entre termes
- ✓ Un calcul de similarité par concept entre document/requête

2

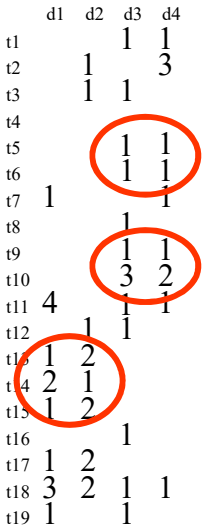
En résumé :

- LSI est une approche vectorielle
- Exploite les co-occurrences entre termes
- Réduit l'espace des termes, en regroupant les termes co-occurrents (similaires) dans les mêmes dimensions
- les documents et les requêtes sont alors représentés dans un espace plus réduit, composé de concepts de haut niveau

3

Exemple

Matrice Terme x Doc



4

2. L'idée générale du modèle LSI

Afin de trouver les concepts basés sur les co-occurrences entre termes, le modèle LSI utilise la technique de la décomposition d'une matrice en ***valeurs singulières*** ***(SVD*** *:* ***Singular Value Decomposition).***

Avant de voir le détail du modèle LSI, il faut bien comprendre la technique SVD …

La technique SVD se base sur la notion des valeurs propres et des vecteurs propres d'une matrice

5

2.1. Rappel sur les valeurs propres et les vecteurs propres d'une matrice

Pour une matrice S $m \times m$, il existe un vecteur propre $\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0}$ et une valeur propre $\lambda \in \mathbb{R}$, telque :

$$\mathbf{Sv} = \lambda \mathbf{v}$$

Exemple

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

2.1.1. Nombre de valeurs propres et de vecteurs propres possible

$$\mathrm{S}\mathrm{v} = \lambda \mathrm{v} \iff (\mathrm{S} - \lambda \mathrm{I})\,\mathrm{v} = 0$$ *Possède une solution unique si* $|\mathbf{S} - \lambda \mathbf{I}| = 0$

Un système à m équations en λ peut avoir au plus ***m solutions*** *distinctes. Alors pour une matrice m×m on peut avoir m valeurs propres et m vecteurs propres.*

2.2. Décomposition en valeurs singulières

Pour une matrice ***A*** *(*$M \times N$*)* ***il existe une décomposition***
(Singular Value Decomposition = ***SVD****) :*

$$A = U \Sigma V^T$$

$M \times M$ $\quad$ $M \times N$ $\quad$ V $N \times N$

- *Les colonnes de U sont les vecteurs propores de* $\boldsymbol{AA^T}$.
- *Les colonnes de V sont les vecteurs propores de* $\boldsymbol{A^TA}$.
- *les valeurs propres* $\lambda_1 \ldots \lambda_r$ *de* $\boldsymbol{AA^T}$ *sont également celles de* $\boldsymbol{A^TA}$.

$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag\left(\sigma_1 \ldots \sigma_r\right)$$ ← *Valeurs singulières*

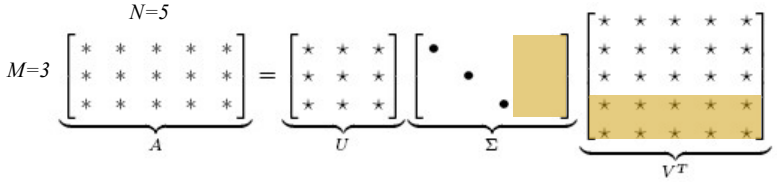
Remarque :

La "diagonale" de Σ contient les ***valeurs singulières*** de ***A***.

- Ce sont des nombres réels et non négatifs.
- La partie supérieure de la diagonale de Σ contient les valeurs singulières strictement positives.
  - ✓ leur nombre est égal à *r*, le rang de ***A***. Le rang d'une matrice est donc révélé par le nombre de valeurs singulières non nulles.
  - ✓ elles sont égales aux racines carrées positives des valeurs propres de $\boldsymbol{AA^T}$.
  - ✓ la partie inférieure de la diagonale contient les (*n* - *r*) valeurs singulières nulles.

8

Illustration de la SVD



Exemple d'une decomposition SVD

Soit $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$

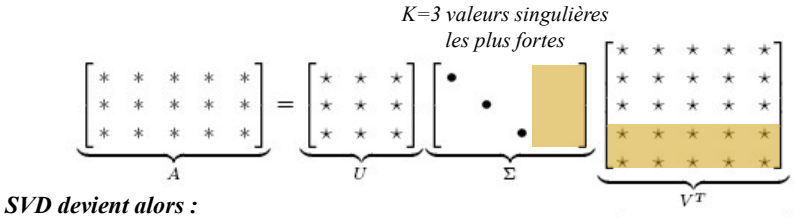
M=3, N=2. son SVD est

$$\begin{bmatrix} 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

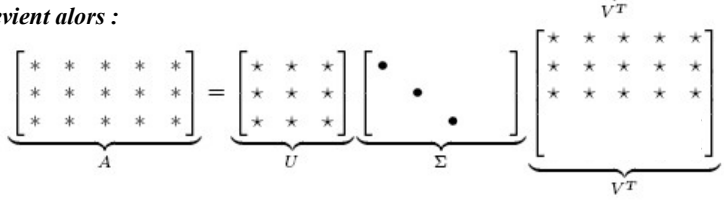
Les valeurs singulières sont rangées par ordre décroissant

2.3. SVD Réduite

- Si on retient les *k premières valeurs singulières (les plus fortes)* on peut réduire la matrice Σ
- Σ devient $k \times k$, U devient $(M \times k)$ et V^T devient $(k \times N)$,
- C'est la version réduire de de SVD



SVD devient alors :



Résumé de la SVD :

- Base mathématique de la LSI : décomposition par valeur singulière de la matrice terme-document
- SVD identifie un ensemble utile de vecteurs colonnes couvrant le même espace de vecteurs associés à la représentation des documents
- SVD décompose la matrice W en trois matrices
  - T : matrice de termes
  - D : matrice de documents
  - S : matrice de valeurs singulières

$$W_{txd} = T_{t\times r} \times S_{r\times r} \times D^T_{r\times d}$$

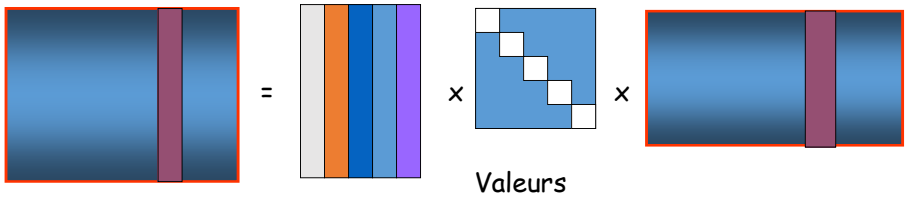
Remarque: nous pouvons faire cette décomposition SVD en python en exploitant le module « numpy » (qu'il faut installer) comme suit :

```
from numpy import *
import numpy as np
t, s, d = np.linalg.svd(W, full_matrices=False)
s = np.diag(s)
```

12

Résumé de la SVD (suite) :

W = T x S x D



Matrice termes

Valeurs singulières

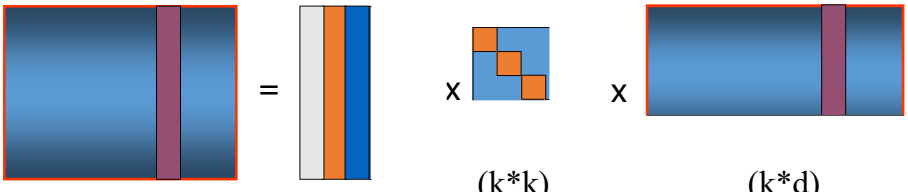
Matrice documents

13

Résumé de la SVD (suite) :

Sélectionner les k premières valeurs singulières de S

W = T x S x D



(t*k) (k*k) (k*d)

Les colonnes de la matrice D représentent les documents dans le nouvel espace vectoriel (espace des concepts)

La fonction qui permet le passage de l'espace des termes à l'espace des concepts est $M = T.S^{-1}$

14

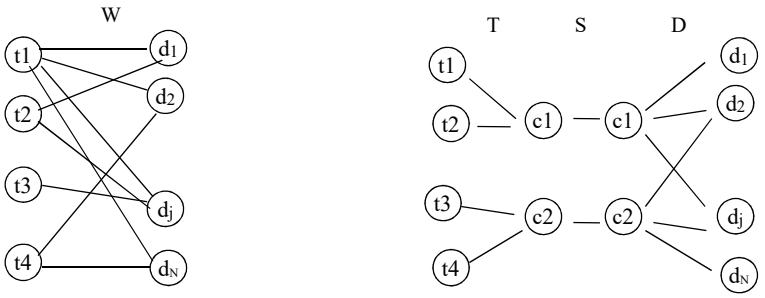
SVD : algorithme

- Calculer la SVD de la matrice terme document
- Sélectionner les k premières valeurs singulières de la matrice S
- Garder les colonnes correspondantes dans les matrices T et D
- La matrice D représente les vecteurs documents dans le nouvel espace M.
- La fonction de changement de repère est donc formée par la matrice $M=T[t,k].S^{-1}[k,k]$

15

Où sont les concepts ?

$$W_{t\times d} = T_{t\times r} \times S_{r\times r} \times D_{r\times d}$$



✓ S est une matrice diagonale composée de valeurs singulières.
✓ Une bonne approximation de la matrice W , est obtenue en utilisant la matrice S(k,k) constituée des k plus fortes valeurs singulières

16

3. Représentation de la requête par LSI :

Pour évaluer une requête Q, représenter la requête dans l'espace M, par la formule :

$$Q_{new} = Q^T.M$$

4. Appariement documents/requête par LSI :

Le calcul de la similarité entre chaque document et la requête, tous représentés dans le nouvel espace vectoriel M, se fait par la formule : $sim = Q_{new} * S^2 * D.$

Le résultat **"sim"** est un vecteur de M cases. Chaque sim[j] représente RSV(Dj, Q)

17

Récapitulatif

- Très bonnes performances pour K= 250, 300, selon l'éxpérimentations reportées par S. Dumais sur les collections TREC disque 1/2/3 (plusieurs centaines de milliers de documents).
- LSI a plusieurs autres applications
  - ✓ classification de termes,
  - ✓ classification de documents,
  - ✓ croisement de langues, …
- Ce modèle est très coûteux en calcul

18

Exemple

Soit la collection suivante :

D1: " t1 t2 t3 t4 t5 t6 t7 "
D2: " t8 t2 t9 t10 t5 t6 t11 t9 "
D3: " t1 t2 t3 t10 t5 t6 t11 "

Soit le requête suivante :

Q : " *t3 t9 t11* "

La matrice Termes*Documents est comme suit :

Questions :

1. Décomposer la matrice par SVD
2. Réduire la matrice S (prendre les k valeurs fortes >=2)
3. Mesurer la similarité entre les documents et la requête Q

Documents / Termes	D1	D2	D3
t6	1	1	1
t10	0	1	1
t4	1	0	0
t8	0	1	0
t7	1	0	0
t3	1	0	1
t5	1	1	1
t2	1	1	1
t1	1	0	1
t9	0	2	0
t11	0	1	1

19

Décomposition de la matrice par SVD et Réduire la matrice S (prendre les k valeurs fortes supérieures ou égales à 2)

T

-0,4201	0,0748
-0,2995	-0,2001
-0,1206	0,2749
-0,1576	-0,3046
-0,1206	0,2749
-0,2626	0,3794
-0,4201	0,0748
-0,4201	0,0748
-0,2626	0,3794
-0,3151	-0,6093
-0,2995	-0,2001

S

4,0909	0
0	2,3616

D

-0,4945	-0,6458	-0,5817
0,6492	-0,7194	-0,2469

20

Calcul de similarités entre les documents et la requête

Q : " t3 t9 t11 "

Q
0
0
0
0
0
1
0
0
0
1
1

$Q_{new} = Q^T.T.S^{-1} =$ | -0.214 | -0.182 |

- **La similarité entre les vecteurs documents et la requête =**

$RSV = Q_{new} * S^2 * D =$ | 1.05 | 3.1043 | 2.3503 |

RSV(d1,q) RSV(d2,q) RSV(d3, q)

21