Indexation sémantique Latente
Latent Semantic Indexing

Latent Semantic Indexing

- Problématique
  - Dans les modèles de RI basés sur les termes, les documents **pertinents** n'ayant aucun terme de la requête n'ont aucune chance d'être sélectionnés
  - Ces documents contiennent plutôt des termes synonymes à ceux de la requête (exemple : voiture/automobile, …)

- Une réponse
  - Regrouper les termes sémantiquement reliés (co-occurrence, …), dans un même concept
  - Faire une recherche par concept

Latent Semantic Indexing

- LSI apporte une solution
  - 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 espace plus réduit, composé de concepts de haut niveau
- Comment réduire l'espace des termes:
  - Techniques d'analyse de données (**SVD**, AFC, ACP, …)

Exemple

Matrice Terme x Doc

	d1	d2	d3	d4
t1			1	1
t2		1		3
t3		1	1	
t4				
t5			1	1
t6			1	1
t7	1			1
t8			1	
t9		1	1	1
t10	1		3	2
t11	4		1	1
t12		1	1	
t13	1	2		
t14	2	1		
t15	1	2		
t16			1	
t17	1	2		
t18	3	2	1	1
t19	1		1	

Rappel : Vecteurs propres & Valeurs propres

- **Vecteurs propres** (pour une matrice S $m \times m$)

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

Vecteur propre $\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0}$

Valeurs propre $\lambda \in \mathbb{R}$

Exemple

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

- Combien de valeurs propres ?

$$\mathbf{S}\mathbf{v} = \lambda\mathbf{v} \iff (\mathbf{S} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0}$$

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*

Décomposition d'une matrice

- Soit $\mathbf{S} \in \mathbb{R}^{m \times m}$ une matrice carré **avec *m* vecteurs propres linéairement indépendant**
- **Théoreme**: Il existe une **décomposition**

diagonale

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

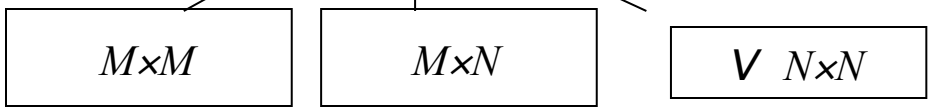
- Colonnes de ***U*** sont des **vecteurs propres de *S***
- Diagonale de $\mathbf{\Lambda}$ valeurs propres de $\mathbf{S}$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

Décomposition en valeurs singulières

Pour une matrice **A** (M × N) **il existe une factorisation** (Singular Value Decomposition = **SVD**) :

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



Les colonnes de U sont les vecteurs propores $\boldsymbol{AA^T}$.

Les colonnes de $\boldsymbol{V}$ *sont les vecteurs propores de* $\boldsymbol{A^TA}$.

les valeurs propres $\lambda_1 \ldots \lambda_r$ *de* $\boldsymbol{AA^T}$ *sont également celles de* $\boldsymbol{A^TA}$.

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

$$\Sigma = diag(\sigma_1 \ldots \sigma_r)$$

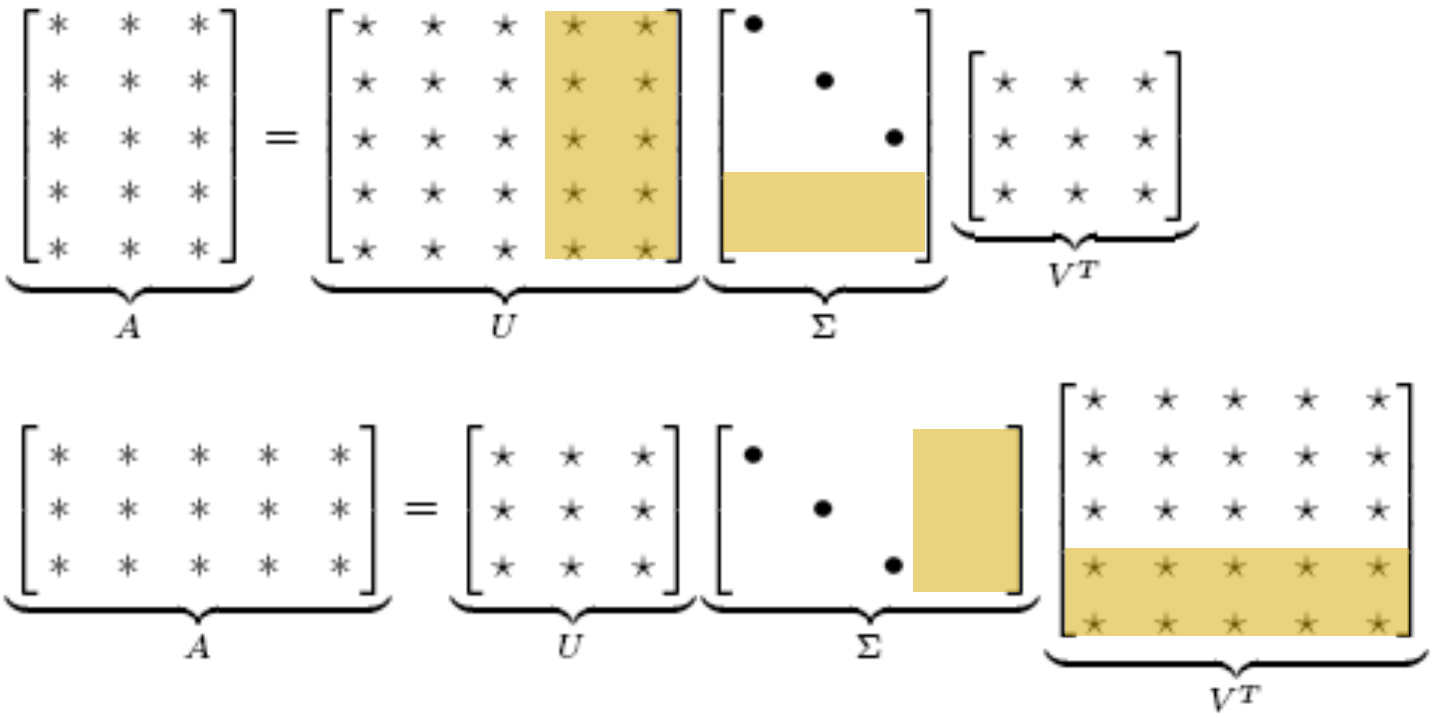
Valeurs singulières

Décomposition en valeurs singulières

- 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. → le rang d'une matrice est le nombre de colonnes linéairement indép.)
    - 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.

Décomposition en valeurs singulières

- Illustration de la SVD



SVD example

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

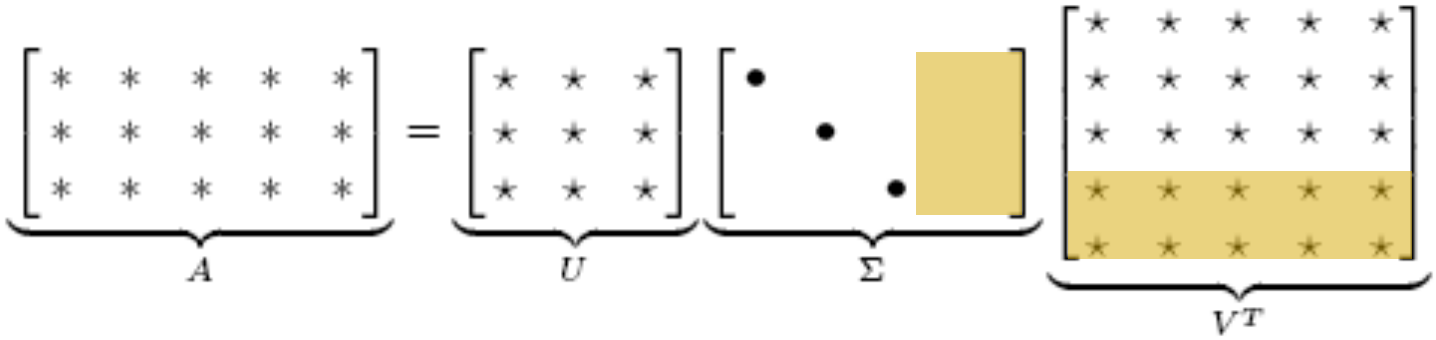
M=3, N=2. son SVD est

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

Les valeurs singulières sont rangées par ordre décroissant

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 ($M \times k$) et V^T (k×N),
- C'est la version réduire de de SVD



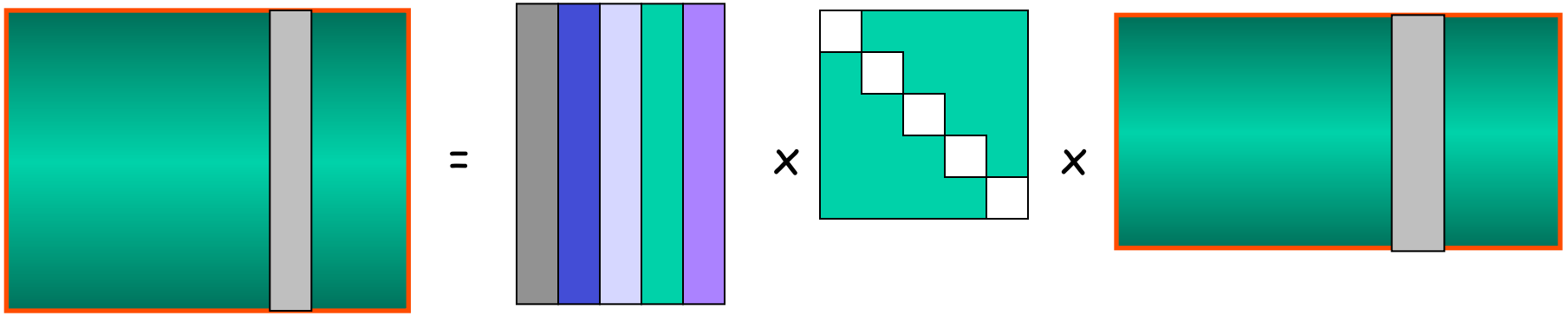
Décomposition en Valeur Singulière (SVD)

- 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 terme
  - D matrice document
  - S matrice de valeurs singulières

$$\underset{tXd}{W} = \underset{t\times r}{T} \times \underset{r\times r}{S} \times \underset{r\times d}{D^T}$$

Décomposition en Valeur Singulière (SVD)

W = T x S x D



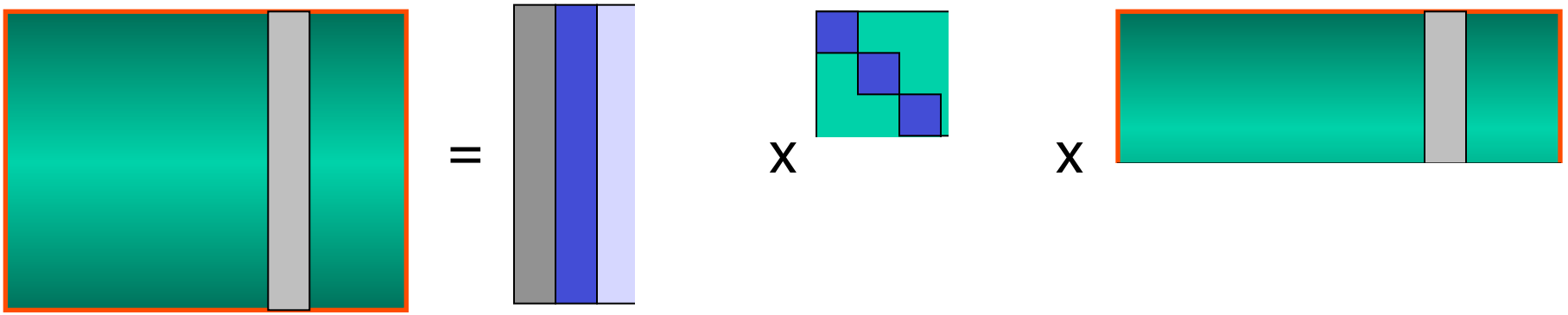
Valeurs singulières

Matrice documents

Décomposition en Valeur Singulière (SVD)

Sélectionner les k premières valeurs singulières de S

W = T x S x 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[t,k].S-1[k,k]

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 $M=T[t,k].S^{-1}[k,k]$

SVD : algorithme (suite)

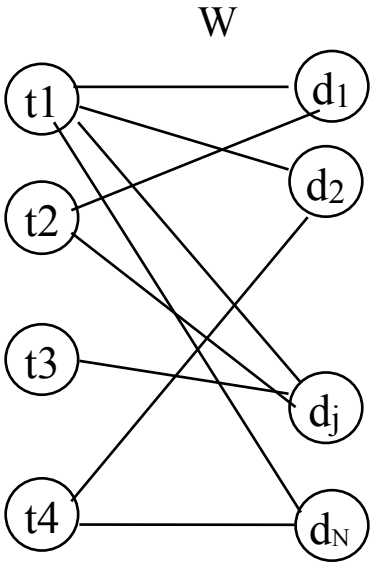
- Pour évaluer une requête Q
  - Représenter la requête dans l'espace M

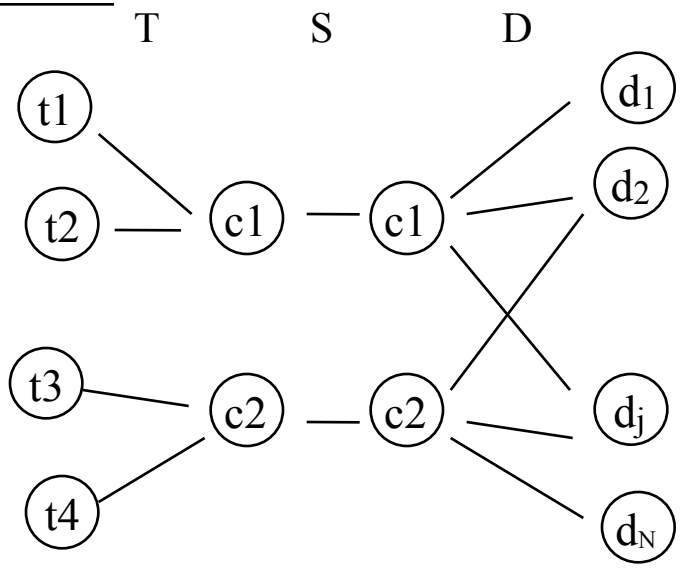
$$Q_{new}=Q^{T}.M$$

Puis calculer la similarité entre chaque document et la requête, tous représentés dans le nouvel espace vectoriel M

Ou sont les concepts ?

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





S 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

Récapitulatif

- Expérimentations reportées par S. Dumais sur les collections TREC disque 1/2/3 (plusieurs centaines de milliers de documents, voir chapitre évaluation)
  - Très bonnes performances pour K= 250, 300,
- LSI a plusieurs autres applications
  - classification de termes,
  - classification de documents,
  - croisement de langues, …
- Très coûteux en calcul

Exemple

Soit la collection suivante :
Q : “or argent cargo” :
D1: “cargaison d’or endommagée dans un incendie”
D2: “Envoi d’argent arrivé dans un cargo argent”
D3: “cargaison d’or arrivé dans un cargo.”

Exemple

terme	D1	D2	D3
un	1	1	1
arrivé	0	1	1
endommagé	1	0	0
envoi	0	1	0
incendie	1	0	0
or	1	0	1
dans	1	1	1
d	1	1	1
cargaison	1	0	1
argent	0	2	0
cargo	0	1	1

Exemple

- Décomposer la matrice par SVD (résultat sur feuille)
- Réduire la matrice S (prendre les k valeurs fortes ?)

Exemple

- Soit la requête suivante
  - Q(or, argent, cargo)

- Mesurer la similarité entre les vecteurs documents et requête

Exemple

T

-0,4201	0,0748	-0,046
-0,2995	-0,2001	0,4078
-0,1206	0,2749	-0,4538
-0,1576	-0,3046	-0,2006
-0,1206	0,2749	-0,4538
-0,2626	0,3794	0,1547
-0,4201	0,0748	-0,046
-0,4201	0,0748	-0,046
-0,2626	0,3794	0,1547
-0,3151	-0,6093	-0,4013
-0,2995	-0,2001	0,4078

S

4,0909	0	0
0	2,3616	0
0	0	1,2737

D

-0,4945	-0,6458	-0,5817
0,6492	-0,7194	-0,2469
-0,578	-0,255	0,775

S-1

0,2440	0
0	0,4234

qn (-0,8772, -0,43)

Fin

Exemple

Term-Doc	d1	d2	d3	d4	d5	d6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1

U:

0.44 -0.30 -0.57 0.58 -0.25
0.13 -0.33 0.59 0.00 -0.73
0.48 -0.51 0.37 0.00 0.61
0.70 0.35 -0.15 -0.58 -0.16
0.26 0.65 0.41 0.58 0.09

S:

2.16 0.00 0.00 0.00 0.00
0.00 1.59 0.00 0.00 0.00
0.00 0.00 1.27 0.00 0.00
0.00 0.00 0.00 1.00 0.00
0.00 0.00 0.00 0.00 0.39

DT:

0.75 0.28 0.20 0.45 0.33 0.12
-0.29 -0.53 -0.19 0.63 0.22 0.41
-0.28 0.75 -0.45 0.20 -0.12 0.33
0.00 0.00 0.58 0.00 -0.58 0.58
0.53 -0.29 -0.63 -0.19 -0.41 0.22

S-1

0.46 0.00
0.00 0.63