Un texte, une modélisation

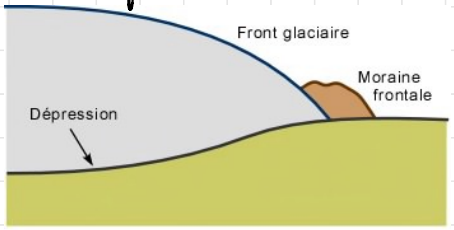
Laurent Dumas

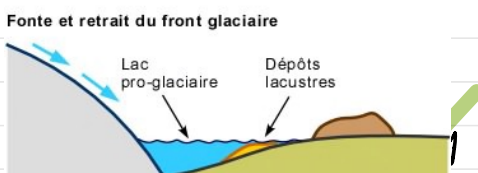
Texte 4 : une requête bibliographique

* Objectif : classer la pertinence de documents par rapport à une requête par mots clés.
* Exemple étudié : classer la pertinence de 12 documents par rapport à une requête polysémique.
* Outils mathématiques :
  Algèbre linéaire (décomposition en valeurs singulières).

dépression

dépression

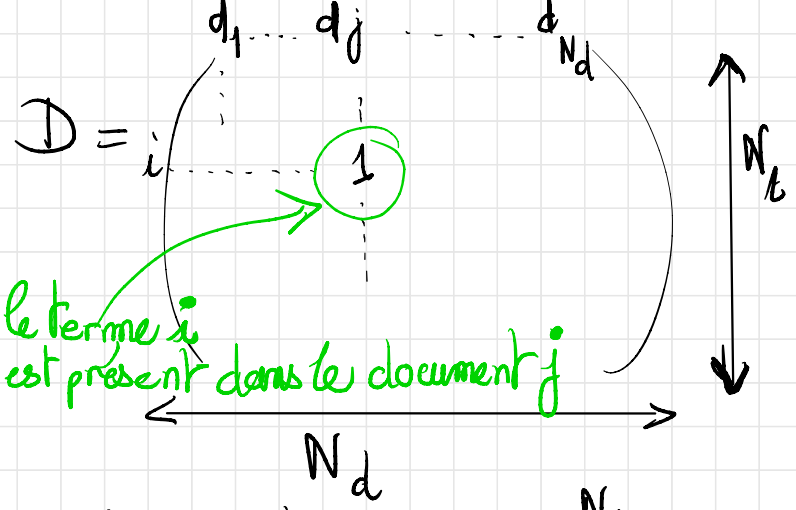




Etape 1 : modélisation d'une requête bibliographique

* On dispose de :
  → N_d documents
  contenant :
  → N_t termes
et on cherche la pertinence d'une requête choisie parmi ces termes.
* On construit :

→ $D \in \mathcal{M}_{N_t, N_d}(\mathbb{R})$ la matrice termes-documents :



→ la requête $q \in \mathbb{R}^{N_t}$:

$q =$ (1 en position i) le terme i est présent dans la requête q

→ la réponse sous la forme d'un vecteur de score : $s \in \mathbb{R}^{N_d}$:

$$s_j = \frac{\langle q, d_j \rangle}{\|q\| . \|d_j\|}$$

Etape 2 : décomposition d'une matrice en valeurs singulières

* La méthode précedente comporte plusieurs difficultés :
  - → coût de calcul
  - → gestion des aspects de redondance, synonymie, polysemie.

* On peut corriger ces difficultés tout en simplifiant le problème !

Théorème 1 : soit $A \in \mathcal{M}_{n,m}(\mathbb{R})$ de rang $r \in \min(n, m)$.
Il existe une unique famille $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ et un couple de matrices (U, V) tel que $U \in O_n(\mathbb{R})$, $V \in O_m(\mathbb{R})$ et

$$A = U \begin{pmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ 0 & & & 0 \end{pmatrix} {}^t V$$

* Les réels σ_i s'appellent les valeurs singulières.

Etape 3 : réduction de rang et calcul d'un score approché

* Un aspect important de la décomposition SVD est le suivant :

Théorème 2 : la meilleure approximation de A parmi les matrices de rang k est donnée par :

$$A_k = U \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_k & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} {}^t V$$

(au sens de la norme $\|\|_2$)

* A partir de la décomposition SVD de la matrice $D \in \mathcal{M}_{N_l, N_d}(\mathbb{R})$ on calcule une matrice approchée :

$$D_k = \underbrace{U_k}_{N_l \times k} \underbrace{\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}}_{\Sigma_k} \underbrace{{}^t V_k}_{k \times N_d}$$

puis un score approché $\tilde{s}$:

$$\tilde{s}_j = \frac{\langle q, D_k e_j \rangle}{\|q\| \cdot \|D_k e_j\|}$$

$$= \frac{\langle {}^t U_k q, \Sigma_k {}^t V_k e_j \rangle}{\|{}^t U_k q\|_2 \, \|\Sigma_k {}^t V_k e_j\|_2}$$

Etape 4 : simulation avec le logiciel Scilab

Documents économiques	Documents géologiques
Doc. 1 : chômage, impôts, dépression, commerce	**Doc. 8 :** bassin, faille, dérive
Doc. 2 : économie, marché, commerce	**Doc. 9 :** dépression, faille
Doc. 3 : commerce, marché, production	**Doc. 10 :** bassin, drainage, vallée
Doc. 4 : dépression, liquidation, emploi, redressement	**Doc. 11 :** dépression, drainage, érosion
Doc. 5 : indemnisation, chômage, liquidation	**Doc. 12 :** bassin, drainage, volcan
Doc. 6 : commerce, marché, prix, emploi	
Doc. 7 : bénéfices, indemnisation, chômage	

On prend $N_d = 12$, $N_t = 20$

* On teste les requêtes :

↗ requête 1 : [depression, commerce]

↘ requête 2 : [emploi]

Pour aller plus loin :

→ aspects algorithmiques de la SVD

→ lien avec la recherche de valeurs propres.

Références :

www.agreg.org

(texte de modélisation)