Chapitre 3
Pondération des termes

1. Introduction

- ✓ La pondération des termes est une étape importante dans le processus d'indexation dans la RI
- ✓ Elle permet d'attribuer une valeur à chaque terme pour mesurer son importance dans le document
- ✓ La 1ère valeur qui peut être utilisée c'est la fréquence du terme dans le document
- ✓ Un terme est important dans le document s'il est fréquent dans le document et moins fréquent dans la collection
- ✓ Une autre mesure qui peut être ajoutée à la fréquence du terme pour bien mesurer son poids c'est sa fréquence dans la collection. Cette valeur s'appelle la valeur de discrimination

2. Pondération basée sur la valeur de discrimination

L'approche TF*IDF est une approche très connue dans le domaine de la RI, qui se base sur la pondération par valeur de discrimination.

TF : signifie « term frequency » (Fréqeunce d'un terme dans le document)

IDF : signifie «inverse document frequency » (Fréquence inverse du document)

Donc, TF mesure la fréquence du terme dans le document et IDF mesure la valeur de discrimination du poids du terme.

3

3. Calcul de TF

Il existe plusieurs formules pour calculer TF, les plus connues sont :

$$tf_{ij} = fréq_{ij} \qquad tf_{ij} = 1 + \log(fréq_{ij}) \qquad tf_{ij} = \frac{fréq_{ij}}{\max(fréq_j)}$$

$$tf_{ij} = \frac{freq_{ij}}{(K + freq_{ij})}$$

K introduit pour tenir compte de la longueur des documents

$$tf_{ij} = \frac{fréq.}{(fréq_{ij}. + 0.5 + 1.5 * \frac{longueur\_doc}{longueur\_moy\_doc})}$$

4

3. Calcul de l'IDF

Il y a deux formules pour calculer l'IDF d'un terme t_i :

$$IDF_i = \log\left(\frac{N}{n_i}\right) \qquad IDF_i = \log\left(\frac{N}{n_i} + 1\right)$$

Avec :
N : le nombre de documents de la collection,
n_i le nombre de documents contenant le terme t_i

Remarque :
La 1ère formule retourne zéro pour un terme qui existe dans tous les documents de la collection. Pour éviter ce problème nous utilisons la 2ème formule dans la suite de nos cours.

5

5. Calcul du poids d'un terme par TF*IDF

Pour calculer le poids d'un terme par TF*IDF, il suffit de choisir une formule de TF est al multiplier par une formule de l'IDF.

Exercice : Soit la collection suivante de 3 documents :

D1 : { langage de programmation python est très utilisé pour le traitement de texte }

D2 : { le langage JAVA est basé sur le langage C++ }

D3 : { un langage de programmation est un langage utilisé pour traduire un algorithme en un programme }

Motsvides: { de, est, très, pour, le, un, en }

Donner le fichier inverse de la collection avec la formule :
poids(ti, dj)=(freq(ti,dj)/max(freq(dj))*log(N/ni +1)

6

Solution :

Terme \ doc	D1	D2	D3
langage	0.301	0.301	0.301
programmation	0.397		0.198
python	0.602		
utilisé	0.397		0.198
traitement	0.602		
texte	0.602		
java		0.301	
basé		0.301	
c++		0.301	
traduire			0.301
algorithme			0.301
programme			0.301

7