Chapitre 4
Les modèles de base de RI

1

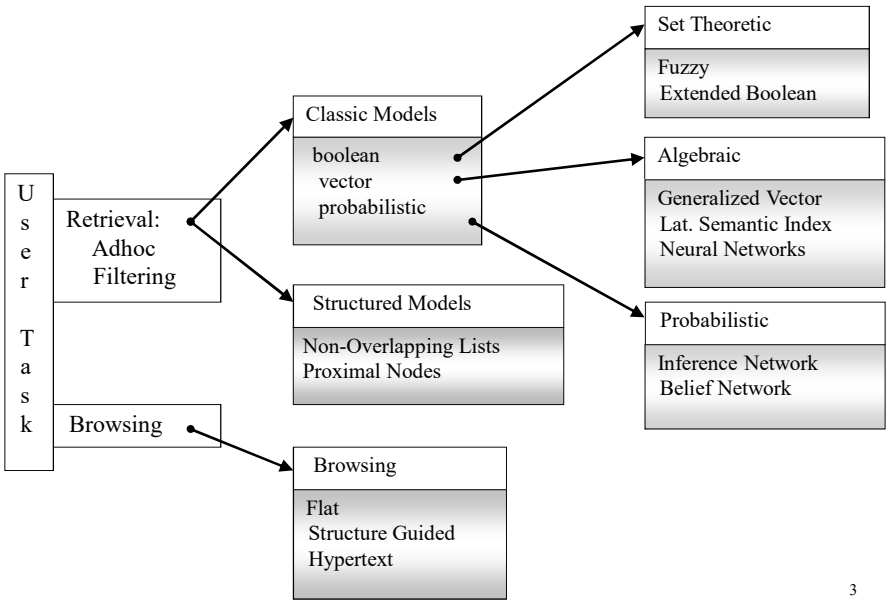
1. Introduction

Un modèle est une abstraction d'un processus. Un modèle de RI doit comporter au minimum les modules suivants :

- Un module de représentation des documents (indexation)
- Un module de représentation des requêtes
- Un module d'appariement entre un document et une requête (similarité)

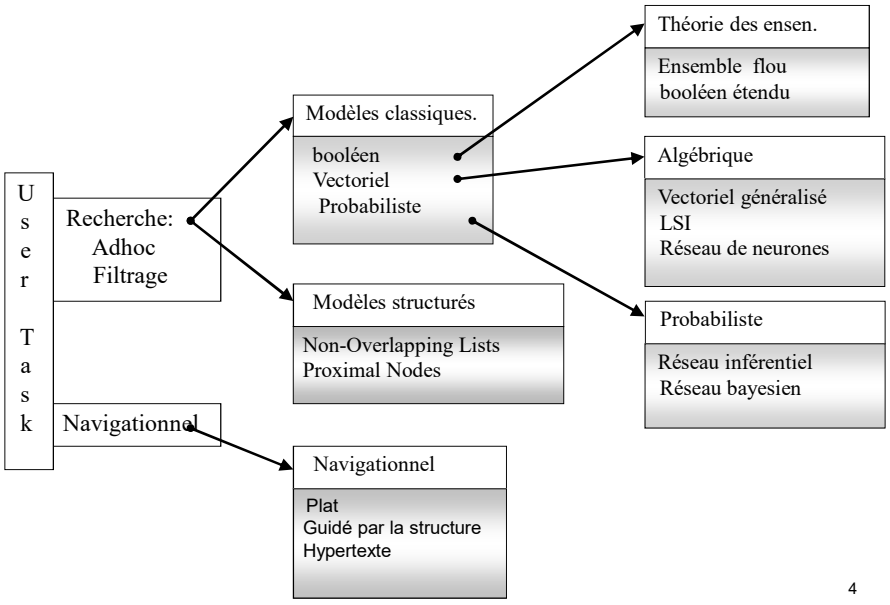
2

Modèles de RI (en anglais)



3

Modèles de RI (en français)



4

2. Modèles de base de RI à étudier

Dans ce chapitre nous allons étudier les modèles de base suivants :

- **Modèle booléen de base**
- **Modèle vectoriel**
- **Modèle booléen basé sur les ensembles flous**
- **Modèle booléen étendu**
- **Modèle P-norme**

5

2.1. Le Modèle Booléen

- ✓ Le premier modèle de RI
- ✓ Basé sur la théorie des ensembles

2.1.1. Module de représentation des documents dans ce modèle

Un document est représenté par un ensemble de termes, un terme a le poids 1 s'il existe dans le document, 0 sinon.

Donc dans ce modèle on ne calcul ni la fréquence de terme, ni son poids.

6

Exemple : soit la collection de 3 documents suivants :

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 }

Stoplist = { de est très pour le sur un en }

Dans le modèle booléen cette collection sera représentée comme suit :

Documents / Termes	poids des termes dans le modèle booléen		
	D1	D2	D3
langage	1	1	1
programmation	1		1
python	1		
Utilisé	1		1
traitement	1		
Texte	1		
JAVA		1	
Basé		1	
C++		1	
Traduire			1
algorithme			1
programme			1

2.1.2. Module de représentation d'une requête

Une requête est un ensemble de mots avec des opérateurs booléens : AND ($\wedge$), OR($\vee$), NOT ($\neg$)

Exemple : q = $t1 \wedge (t2 \vee \neg t3)$

2.1.3. Module d'appariement

Dans ce modèle la similarité entre un document est une requête est calculée par une valeur exacte basée sur la présence ou l'absence des termes de la requête dans les documents, qui est soit 1 soit 0.

On note Appariement (q,d) par RSV(q,d) qui signifie « Retrieval Status Value »

9

RSV(q,d) = 1 si en remplaçant les termes dans la requête par leurs poids dans le document (0 ou 1), puis en évaluant cette requête comme une expression logique, elle donnera 1

RSV(q,d) = 0 sinon

10

Exemple :

Soit l'ensemble des termes d'indexation = (document, web, information, recherche, image, contenu).

Soit le document d1=(document web document web document)
Soit les 3 requêtes :

q1 = document ET web OU image
q2= (document OU web) ET image
q3= (web OU image) ET document

La similarité entre d1 et chaque requête, par le modèle booléen

RSV(q1,d1)= 1 ET 1 OU 0 = 1
RSV(q2,d1)= (1 OU 1) ET 0 = 0
RSV(q3,d1)= (1 ou 0) ET 1 = 1

Donc, d1 est pertinent pour q1 et q3, mais il n'est pas pertinent pour q2

11

2.1.4. Inconvénient du Modèle Booléen

- La sélection d'un document est basée sur une décision binaire
- Pas d'ordre pour les documents sélectionnés
- Formulation de la requête difficile pas toujours évidente pour beaucoup d'utilisateurs
- Problème de collections volumineuses : le nombre de documents retournés peut être considérable

12

2.2. Modèle Vectoriel (Vector Space Model) (VSM)

Ce modèle est basé sur les formules de similarité entre vecteurs. Il est proposé par Salton dans le système SMART (Salton, G. 1970)

2.2.1. Module de représentation des documents dans ce modèle

Les documents sont représentés sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents. Chaque terme est pondéré selon une formule de pondération comme TF*IDF. Donc une collection est représentée par un fichier inverse avec pondération (vu dans le chapitre précédent).

13

2.2.2. Module de représentation des requêtes dans ce modèle

Les requêtes sont aussi représentées sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents. Chaque terme est pondéré par 1 s'il existe dans la requête, 0 sinon.

On note w_{iq} le poids du terme i dans la requête q

Donc, une requête q= (w_{1q}, w_{2q}, …, w_{Mq}) avec M le nombre de termes dans la collection.

2.2.3. Module d'appariement dans ce modèle

La pertinence d'un document pour une requête dans ce modèle est traduite en une similarité vectorielle. Un document est d'autant plus pertinent à une requête que le vecteur associé est similaire à celui de la requête.

14

La similarité un document est une requête est calculée selon l'une des forentre mules suivantes :

Produit interne $RSV(d_j, q) = \sum w_{ij} * w_{iq}$

Coef. de Dice $RSV(d_j, q) = \frac{2 * \sum w_{ij} * w_{iq}}{\sum {w_{ij}}^2 + \sum {w_{iq}}^2}$

Mesure du cosinus $RSV(d_j, q) = \frac{\sum w_{ij} * w_{iq}}{\sqrt{\sum {w_{ij}}^2 * \sum {w_{iq}}^2}}$

Mesure du Jaccard $RSV(d_j, q) = \frac{\sum w_{ij} * w_{iq}}{\sum {w_{ij}}^2 + \sum {w_{iq}}^2 - \sum w_{ij} * w_{iq}}$

15

Exercice 1

Soit la collection de 3 documents suivants :

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 }

La requête q : { langage python java }

1. Donner le fichier inverse de la collection avec la formule : poids(ti, dj)=(freq(ti,dj)/max(freq(dj))*log(N/ni +1)
2. Calculer la similarité entre chaque document et la requête q par les quatre formules du modèle vectoriel.

16

2.2.4. Avantages du modèle vectoriel :

- ✓ La pondération améliore les résultats de recherche
- ✓ La mesure de similarité permet d'ordonner les documents selon leur pertinence vis à vis de la requête
- ✓ Simple à programmer

2.2.5. Inconvénients du modèle vectoriel :

- ✓ La représentation vectorielle suppose l'indépendance entre termes
- ✓ Le sens des termes n'est pas pris en compte

17

Extension du modèle Booléen

18

3. Modèle booléen basé sur les ensembles flous

Ce modèle est une extension du modèle booléen de base. Un document est un ensemble de termes. Chaque terme à un poids qui mesure à quel point le terme caractérise le document. La similarité entre un document et une requête (booléenne) de termes est calculée selon les formules suivantes :

RSV(dj, ti) = Wij (poids TF*IDF du terme i dans le document j)

RSV(dj, t1$\wedge$ t2) = min(RSV(dj,t1), RSV(dj ,t2))

RSV(dj,t1$\vee$ t2) = max(RSV(dj,t1), RSV(dj ,t2))

RSV(dj, not ti) = 1- RSV(dj, ti)

19

Exemple

	théorie des ensemble					ensemble. flous			
t1	1	1	0	0		0.5	0.5	0	0
t2	1	0	1	0		0.7	0	0.6	0
t1$\cap$ *t2*	1	0	0	0		0.5	0	0	0
t1$\cup$ *t2*	1	1	1	0		0.7	0.5	0.6	0

20

4. Modèle booléen étendu

Ce modèle est une autre extension du modèle booléen. Un document est un ensemble de termes. Chaque terme à un poids qui mesure à quel point le terme caractérise le document. La similarité entre un document et une requête (booléenne) de termes est calculée selon les formules suivantes :

$RSV(dj, ti) = wij$

$$RSV\ (d_j, t_1 \vee t_2) = \frac{\sqrt{\left(w_{1j}^2 + w_{2j}^2\right)}}{\sqrt{2}}$$

$$RSV\,(d_j, t_1 \wedge t_2) = 1 - \frac{\sqrt{\left((1-w_{1j})^2 + (1-w_{2j})^2\right)}}{\sqrt{2}}$$

$$RSV\ (dj, qnot\ ) = 1 - RSV\ (dj, q)$$

21

5. Modèle pnorm (requête avec termes non pondérés)

Ce modèle est une généralisation du modèle précédent pour m termes de la requête. Ce modèle introduit aussi un paramètre P qui peut être associé par exemple aux opérateurs logiques.

La similarité entre un document et une requête (booléenne) de termes est calculée selon les formules suivantes :

$$RSV\ (dj, qor\ ) = \left(\frac{w_{1j}^p + w_{2j}^p + \ldots + w_{mj}^p}{m}\right)^{1/p}$$

$$RSV\,(d_j, qand) = 1 - \left(\frac{(1-w_{1j})^p + (1-w_{2j})^p + \ldots + (1-w_{mj})^p}{m}\right)^{1/p}$$

$$RSV\ (dj, qnot\ ) = 1 - RSV\ (dj, q)$$

22

5.1. Quelques valeurs du paramètre P

- Si p = 1 alors (on retrouve le modèle vectoriel)
  - RSV(dj,qor) = RSV(dj,qand)
- Si p = ∞ alors (modèle booléen)
  - RSV(d_j,qor) = max (wxj)
  - RSV(d_j,qand) = min (wxj)
- p=2 correpond à la distance euclidienne, semble être le meilleur choix

23

6. Modèle pnorm (requête avec termes pondérés)

Ce modèle ajoute la possibilité de donner des poids aux termes de la requête, soit qi le poids d'un terme i dans la requête, donné par l'utilisateur.

Exemple : q = *(t1,0.6) ∧ ((t2,0.3) ∨ ¬ (t3,0.7))*

La similarité entre un document et une requête (booléenne) de termes est calculée selon les formules suivantes

$$RSV\ (dj, qor) = \left(\frac{\sum q_i^p * w_{ij}^p}{\sum q_i^p}\right)^{1/p}$$

$$RSV\ (dj, qand) = 1 - \left(\frac{\sum q_i^p * (1 - w_{ij})^p}{\sum q_i^p}\right)^{1/p}$$

$$RSV\ (dj, qnot) = 1 - RSV\ (dj, q)$$

24

Exercice 2

Soit l'ensemble des termes d'indexation = (document, web, information, recherche, image, contenu).

Soit d1=(document 1, web 0,5)

Soient q1 = (document OU web)

q2= (web ET document)

q3= ((web OU document) ET image)

Questions :

Calculer la similarité entre d1 et chaque requête par :

1. le modèle booléen basé sur les ensembles flous
2. le modèle booléen étendu
3. le modèle p_norme avec p=2.

25