Les modèles de RI

Qu'est ce qu'un modèle de RI ?

- Un modèle est une abstraction d'un processus (ici recherche d'info)
- Les modèles mathématiques sont souvent utilisés pour
  - formaliser les propriétés d'un processus,
  - élaborer des conclusions, faire des prévisions, etc.
- Les Conclusions dérivées d'un modèle dépendent de la qualité du modèle
  - Question : est ce que le modèle est une bonne approximation du processus ?

Qu’est ce qu’un modèle de RI ?

- Les modèles de RI peuvent décrire
  - Le processus de mesure de pertinence : comment les documents sont sélectionnés et triés
  - L’utilisateur : besoin en information, interaction
  - L’information

- Les modèles de RI manipulent plusieurs variables : les besoins, les documents, les termes, les jugements de pertinence , les utilisateurs, …

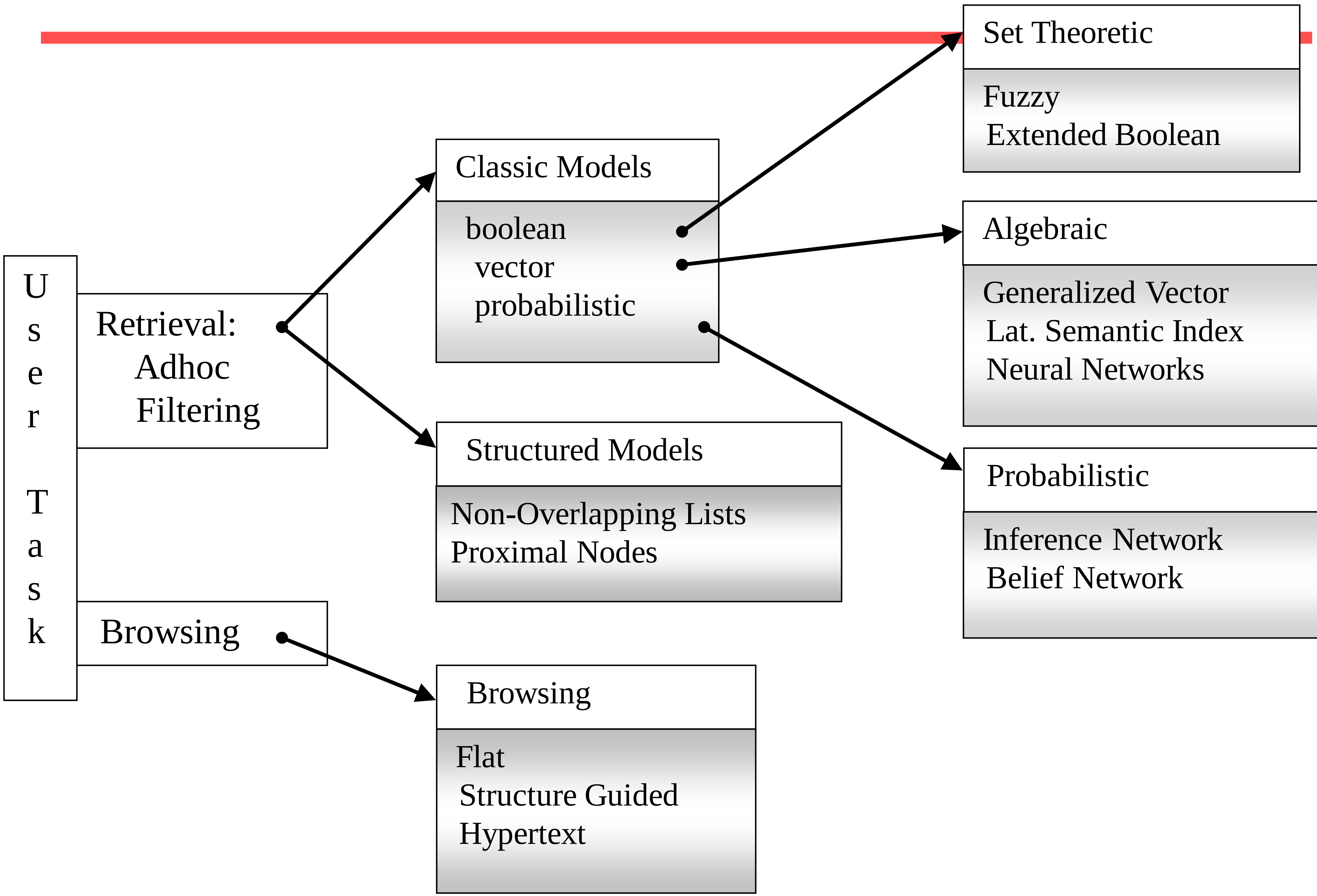
- Les modèles de RI se distinguent par le principe d’appariement (*matching*) : appariement exact /approché *(Exact /Best matching)*

Appariement exact /Appariement approché

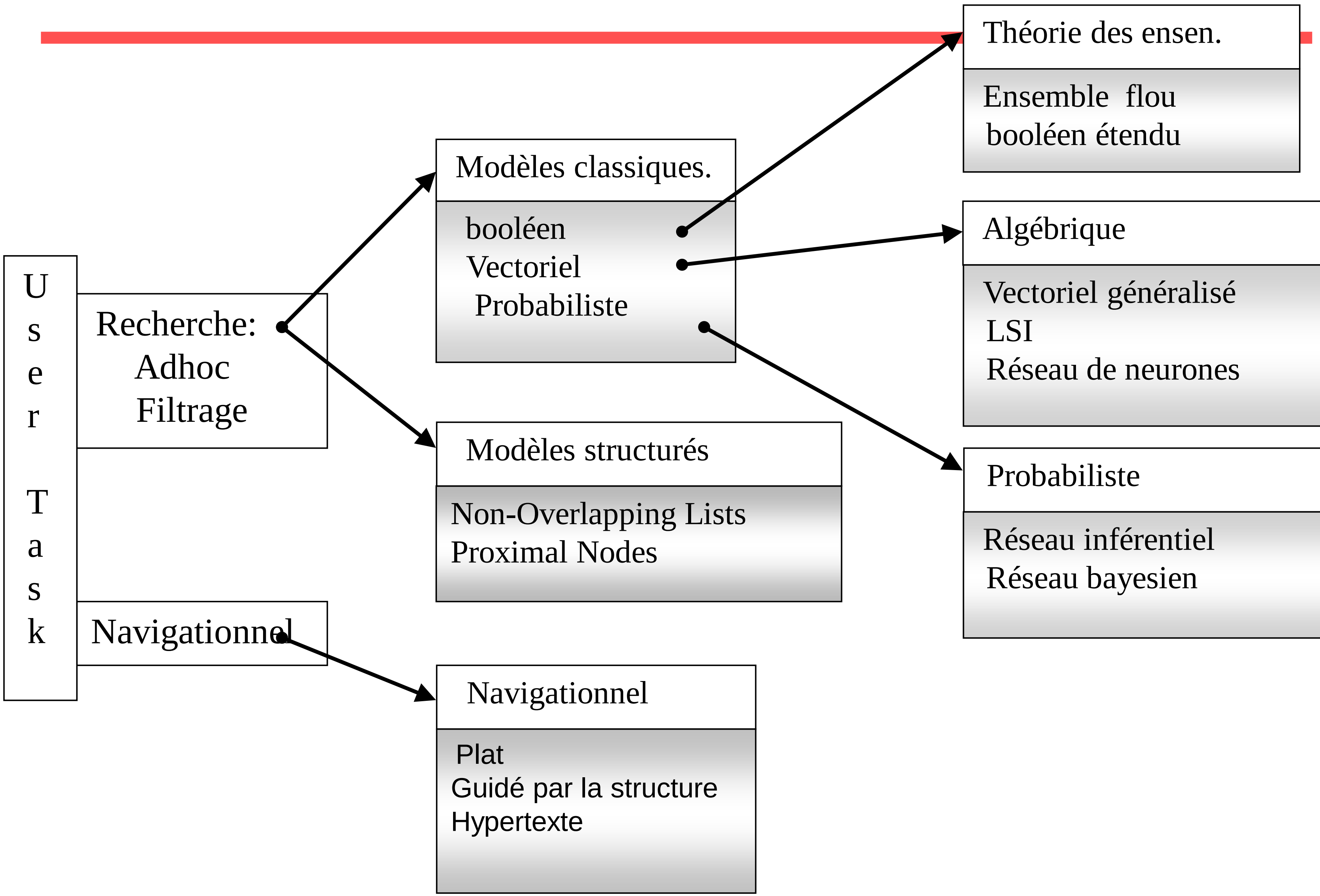
- Appariement exact
  - Requête spécifie de manière précise les critères recherchés
  - L’ensemble des documents respectant exactement la requête sont sélectionnés, mais pas ordonné

- Appariement approché
  - Requête décrit les critères recherchés dans un document
  - Les documents sont sélectionnés selon un degré de pertinence (similarité/ probabilité) vis-à-vis de la requête et sont ordonnés

Modèles de RI



Modèles de RI



Modèles de RI

- Panoplie de modèles
  - Modèle booléen (±1950)
  - Modèle vectoriel (±1970)
  - Modèle LSI (± 1994)
  - Modèle probabiliste (±1976)
  - Modèle inférentiel (±1992)
  - Modèle connexionniste (±1989)
  - Modèle de langage (±1998)

Modèles de RI

- Dans ce module nous allons étudier les modèles suivant
  - **Modèle booléen de base**
  - **Modèle booléen basé sur les ensembles flous**
  - **Modèle vectoriel de base**
  - **Modèle P-norme**
  - **Modèle LSI**
  - **Modèle probabiliste**
  - **Modèle de langage**

Le Modèle booléen
Boolean Model

Le Modèle Booléen

- Le premier modèle de RI
- Basé sur la théorie des ensembles
- Un document est représenté un ensemble de termes
  - Ex : d1(t1,t2,t5); d2(t1,t3,t5,t6); d3(t1,t2,t3,t4,t5)
- Une requête est un ensemble de mots avec des opérateurs booléens : AND ($\wedge$), OR($\vee$), NOT ($\neg$)
  - Ex: $q = t1 \wedge (t2 \vee \neg t3)$
- Appariement Exact basé sur la présence ou l'absence des termes de la requête dans les documents
  - Appariement (q,d) = RSV(q,d)=1 ou 0

Le Modèle Booléen

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

- d1(t1,t2,t5); d2(t1,t3,t5,t6); d3(t1,t2,t3,t4,t5)

Rsv(q,d1)=

Rsv(q,d2)=

Rsv(q,d3)=

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

Modèle Vectoriel
Vector Space Model (VSM)

Modèle Vectoriel *(Vector Space Model) (VSM)*

- Proposé par Salton dans le système SMART (Salton, G. 1970)
- Idée de base :
  - Représenter les documents et les requêtes sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents :

    $T\langle t_1, t_2, \ldots, t_M \rangle$ (un terme = une dimension)

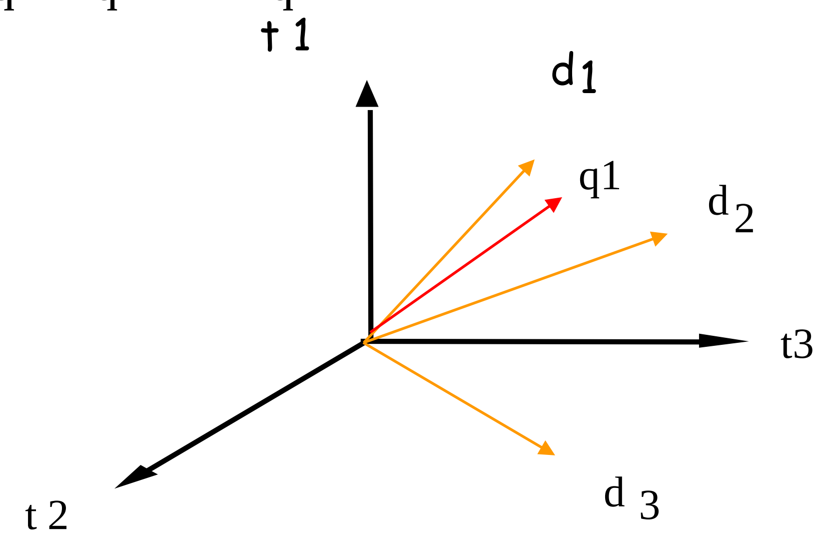
  - Document : $dj = (w_{1j}, w_{2j}, \ldots, w_{Mj})$
  - Requête : $q = (w_{1q}, w_{2q}, \ldots, w_{Mq})$

Modèle Vectoriel
The Vector Model. (VSM)

- Soit $T<t_1,t_2, \ldots, t_M>$: ensemble des M termes de la collection

$dj= (w_{1j}, w_{2j}, \ldots, w_{Mj})$

$q= (w_{1q}, w_{2q}, \ldots, w_{Mq})$



Le Modèle Vectoriel

- Une collection de *n* documents et *M* termes distincts peut être représentée sous forme de matrice

$$\begin{array}{cc} & \begin{array}{cccc} T_1 & T_2 & \cdots & T_M \end{array} \\ \begin{array}{c} D_1 \\ D_2 \\ \vdots \\ \vdots \\ D_n \end{array} & \begin{pmatrix} w_{11} & w_{21} & \cdots & w_{M1} \\ w_{12} & w_{22} & \cdots & w_{M2} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{Mn} \end{pmatrix} \end{array}$$

- La requête est également représentée par un vecteur.

Modèle Vectoriel
The Vector Model. (VSM)

- Exemple :
  - T=(document, web, information, recherche, image, contenu) : ensemble des termes d'indexation
  - d1=(document 2,web 1)
  - d2=(information 1, document 3, contenu 2)
  - q1 = (image web); q2(recherche, documentaire)

  - Représentation vectorielle
    - d1 (2,1,0,0,0,0)
    - d2 (3,0,1,0,0,2)
    - q1 (0,1,0,0,1,0)
    - q2(0,0,0,1,0,0)

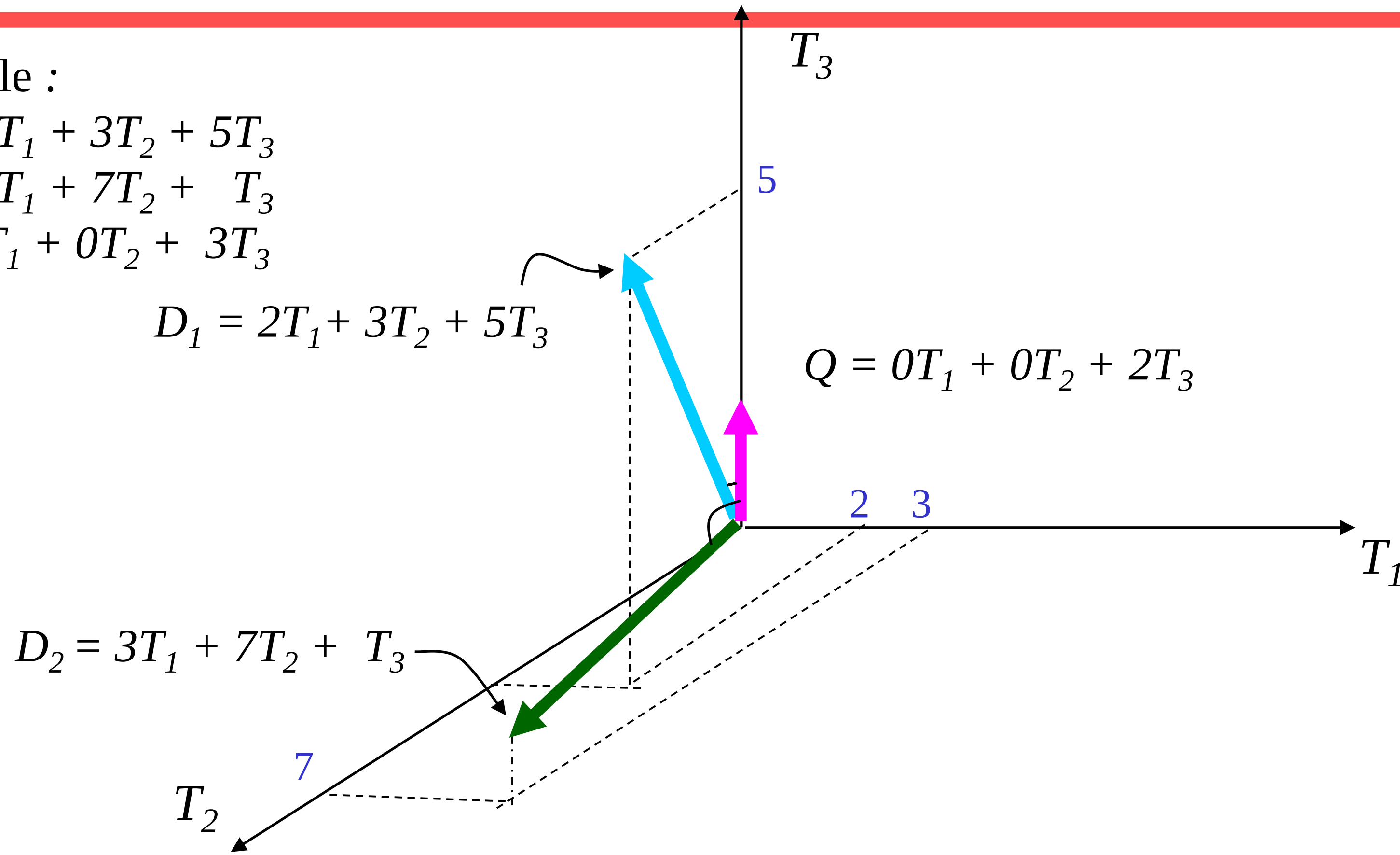
Modèle Vectoriel

Exemple :

$D_1 = 2T_1 + 3T_2 + 5T_3$

$D_2 = 3T_1 + 7T_2 + T_3$

$Q = 0T_1 + 0T_2 + 3T_3$



La pertinence 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.

Le Modèle Vectoriel
mesure de similarité

Inner product	$\|X \cap Y\|$	$\sum x_i * y_i$
Coef. de Dice	$\dfrac{2 * \|X \cap Y\|}{\|X\| + \|Y\|}$	$\dfrac{2 * \sum x_i * y_i}{\sum x_i^2 + \sum y_j^2}$
Mesure du cosinus	$\dfrac{\|X \cap Y\|}{\sqrt{\|X\|} * \sqrt{\|Y\|}}$	$\dfrac{\sum x_i * y_i}{\sqrt{\sum x_i^2 * \sum y_j^2}}$
Mesure du Jaccard	$\dfrac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|}$	$\dfrac{\sum x_i * y_i}{\sum x_i^2 + \sum y_j^2 - \sum x_i * y_i}$

Le Modèle Vectoriel

- Avantages:
  - 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

- Inconvénients:
  - La représentation vectorielle suppose l'indépendance entre termes (?)

Extension du modèle Booléen

Introduction

- Prendre en compte l'importante des termes dans les documents et/ou dans la requête
- Possibilité d'ordonner les documents séléctionnés
- Comment étendre le modèle booléen ?
  - Interpréter les conjonctions et les disjonction
- Deux modèles :
  - Modèle flou- fuzzy based model (basé sur la logique floue)
  - Modèle booléen étendu- extended boolean model

Ensembles flous (1.)

- Théorie des ensembles flous
  - Un cadre pour représenter les ensembles dont les bornes ne sont pas bien définis
  - L'objectif principal est l'introduction de la notion de degré d'appartenance d'un élément à un ensemble
  - Contrairement à la théorie des ensembles ou un élément est dans l'ensemble ou ne l'est pas,
  - …dans les ensembles flous, l'appartenance est mesurée par un degré variant entre 0 et 1
    - 0 $\longrightarrow$ non appartenance
    - 1 $\longrightarrow$ appartenance complète

Ensembles flous (2.)

- Définition
  - Un sous ensemble A d'un univers de discours U est caractérisé par une fonction d'appartenance
    - μ_A: $U \rightarrow [0,1]$
    - qui associe à chaque élément u de U un nombre $\mu_A(u)$ dans [0,1]
  - Soient A et B deux sous-ensembles flous de U
    - Complément $\mu_A(u)$ $$\mu_{\overline{A}}(u) = 1 - \mu_A(u)$$
    - Union $$\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$$
    - Intersection $$\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$$

Modèle flou de RI

- Un document est un ensemble de termes
- chaque terme à un poids qui mesure à quel point le terme caractérise le document
- Ces poids sont dans [0, 1]. (dans le booléen standard un terme est soit présent 1 ou absent 0 dans un document)
- On pourrait écrire :

$$\mu_d(t) = w_{dt}$$

Modèle flou de base, requête non pondérée

- Soient :
  - Termes: $t_1, t_2, \ldots, t_n$
  - Document: $d(w_1, w_2, \ldots, w_n)$
- *Requête disjonctive :* $q_{or} = (t_1 \vee t_2 \vee \ldots \vee t_n)$
  - $RSV(q_{or},d): = \max(w_1, w_2, \ldots, w_n)$
- Requête conjonctive : $q_{and} = (t_1 \wedge t_2 \wedge \ldots \wedge t_n)$
  - $RSV(q_{and}, d) = \min(w_1, \ldots, w_n)$
- **Généralisation**
  - ***$RSV(d,q1 \wedge q2) = \min(RSV(d,q1), RSV(d,q2))$***
  - ***$RSV(d,q1 \vee q2) = \max(RSV(d,q1), RSV(d,q2))$***
  - ***$RSV(d, \text{not } q) = 1- \max(RSV(d,q)$***

Exemple

	théorie des ensemble				ensemble. flous			
	$t1$	$t2$	$t3$	$t3$	$t1$	$t2$	$t3$	$t4$
q	1	1	0	0	0.5	0.5	0	0
d	1	0	1	0	0.7	0	0.7	0
$q \cap d$	1	0	0	0	0.5	0	0	0
$q \cup d$	1	1	1	0	0.7	0.5	0.7	0

Modèle booléen étendu

- Combinaison des modèles booléen et vectoriel
  - Document : liste de termes pondérés
  - Requête booléenne
  - Utilisation des distances algébriques pour mesurer la pertinence d'un document vis-à-vis à d'une requête

Modèle booléen étendu appariement

- Considérons
  - $d_j\ (w_{1j}, w_{2j}, \ldots\ w_{tj})$
  - q : requête à deux termes

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

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

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

Modèle booléen (*pnorm*)étendu appariement

- Généralisation
  - Distance euclidienne à plusieurs dimensions
  - Utilisation de la **p-norm**
- Considérons :
  - un document dj ($w_{1j}, w_{2j}, \ldots w_{tj}$) et q ($t_1$, t_2, ..t_m) : une requête composée de **m** termes non pondérés:
  - Soit p un poids associé aux opérateurs logiques:

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

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

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

Modèle booléen(*pnorm*) étendu appariement

- 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

Modèle booléen (*pnorm*) étendu appariement

- **Généralisation** :
  - **document pondéré**
  - **requête pondérée**

Si la requête et les documents sont pondérés

- $q(q_1, q_2, .., q_m)$
- $d_j\ (w_{1j}, w_{2j}, \ldots\ w_{tj})$

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

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

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

Modèle booléen étendu

- Modèle puissant
- Calcul complexe
- Problème de distributivité
  - $q_1=(t_1 \text{ OU } t_2) \text{ ET } t_3$
  - $q2=(t_1 \text{ ET } t_3) \text{ OU } (t_2 \text{ ET } t_3)$
  - $RSV(q_1,d) <> RSV(q_2,d)$

Exercice

- Exemple :
  - Ensemble des termes d'indexation = (document, web, information, recherche, image,contenu)
  - d1=(document 1,web 0,5)
  - q1 = (document OU web)
  - q2= (web ET document)
  - q3= ((web OU document) ET image)