Recherche d'Information (RI)
- Fondements -

Philippe Mulhem
LIG
Philippe.Mulhem@imag.fr

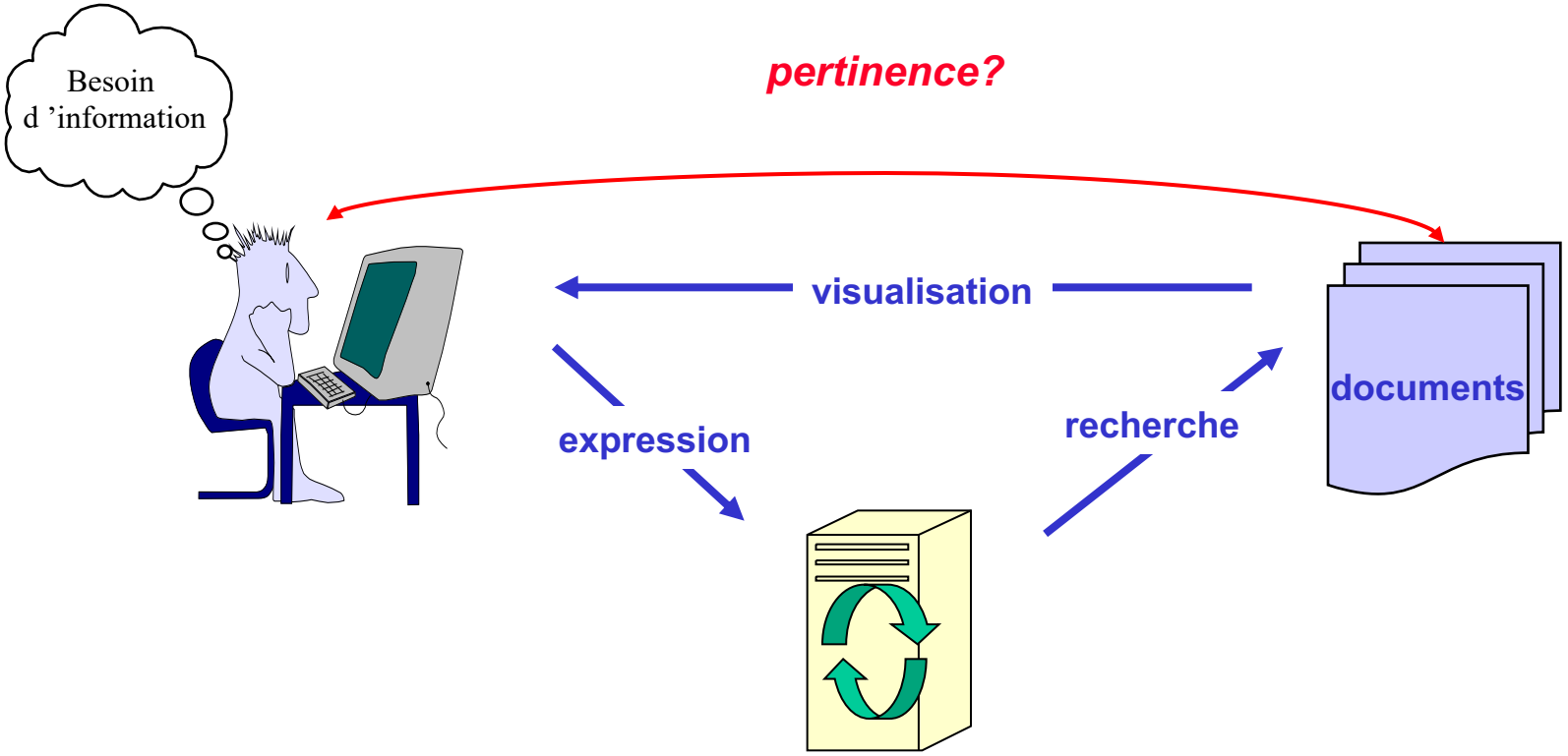
https://rimiashs.imag.fr/

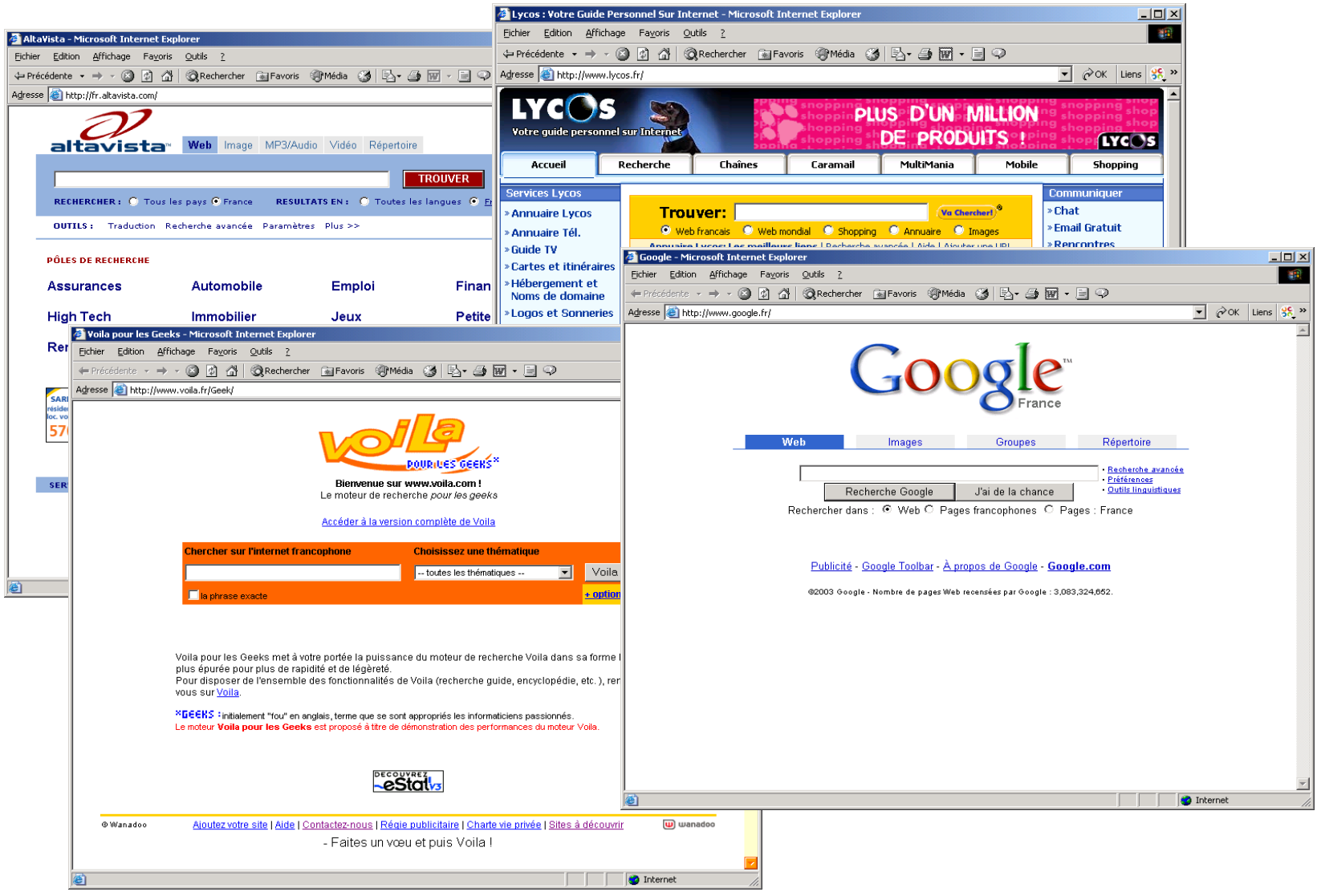
Plan

1. Introduction
2. Eléments clés en RI
3. Modèle de RI
   - Modèle Booléen Pondéré
   - Modèle Vectoriel
4. Systèmes de recherche d'information (SRI)
   - Architecture
5. Evaluation de SRI
6. Conclusion

1. Introduction

- Problématique de la recherche d'information :
  - Accès par le contenu à des documents satisfaisant un besoin d'information d'un utilisateur




AltaVista - Microsoft Internet Explorer
altavista
Web
Image
MP3/Audio
Vidéo
Répertoire
TROUVER
Lycos : Votre Guide Personnel Sur Internet - Microsoft Internet Explorer
LYCOS
Votre guide personnel sur Internet
PLUS D'UN MILLION DE PRODUITS !
Accueil
Recherche
Chaînes
Caramail
MultiMania
Mobile
Shopping
Services Lycos
Annuaire Lycos
Annuaire Tél.
Guide TV
Cartes et itinéraires
Hébergement et Noms de domaine
Logos et Sonneries
Trouver:
Communiquer
Chat
Email Gratuit
Google - Microsoft Internet Explorer
Google
France
Web
Images
Groupes
Répertoire
Recherche Google
J'ai de la chance
Publicité - Google Toolbar - À propos de Google - Google.com
Voila pour les Geeks - Microsoft Internet Explorer
Bienvenue sur www.voila.com !
Le moteur de recherche pour les geeks
Accéder à la version complète de Voila
Chercher sur l'internet francophone
Choisissez une thématique
Voila pour les Geeks met à votre portée la puissance du moteur de recherche Voila dans sa forme
- Faites un vœu et puis Voila !

1. Introduction

- Contextes d'utilisation
  - Bureautique
  - Applications techniques : maintenance de matériel
  - Médecine
  - Génie logiciel
  - Vente en ligne
  - Tourisme
  - Recherche scientifique
  - etc.

2. Eléments clés en RI

- Quels éléments sont centraux pour la Recherche d'Information
  - Documents
  - Contenu des documents
  - Besoin d'information d'un utilisateur
  - Satisfaction

2. Eléments clés en RI

- Documents
  - Différents médias :
    - Texte (livre, article, tweet, …)
    - Image (photo, radio, …)
    - Vidéo
    - Documents structurés

2. Eléments clés en RI

- Documents
  - 2 classes d'information
    - Méta-Information (information à propos du document)
      - Attributs : titre, auteur, date de création, etc.
      - Structure (organisation du contenu) : structure logique, liens, etc.
    - Contenu
      - Contenu brut : le document initial
      - Contenu extrait : information extraite du contenu brut

2. Eléments clés en RI

- Besoin d'information d'un utilisateur

  Utilisation de requêtes suivant un langage fixé

  - Sur la méta information
    - Attributs : « roman écrit par Victor Hugo »
      - attribut de type de document et auteur
    - Structure : « article de football contenant une photographie »
      - Structure de lien entre texte et image
  - Sur le contenu
    - Contenu brut : « lettre avec le texte "Je suis venu, j'ai vu, j'ai vaincu " » : recherche sur des chaînes de caractères
    - Contenu extrait : « documents au sujet de recherche d'information »

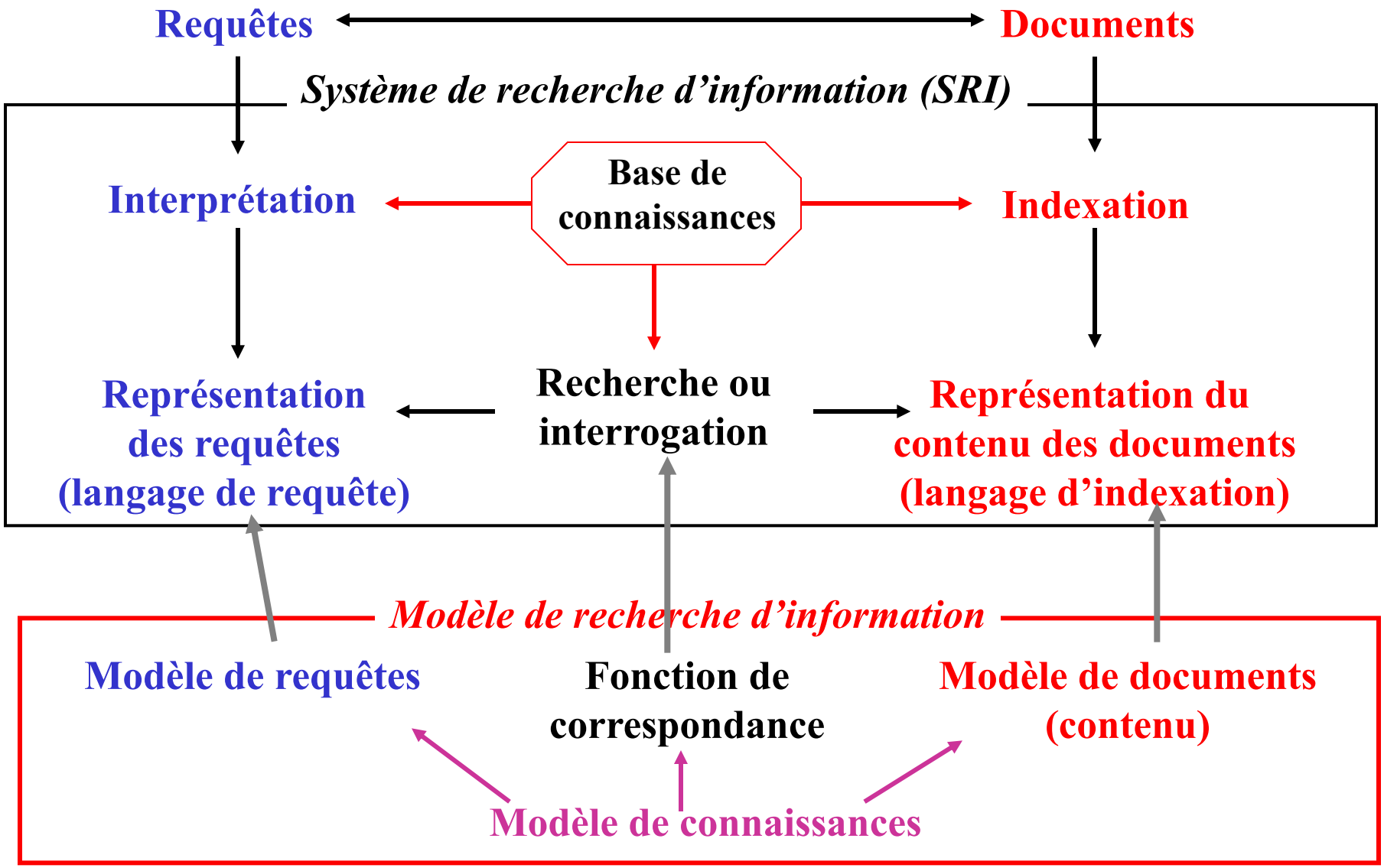
2. Eléments clés en RI

- Satisfaction de l'utilisateur

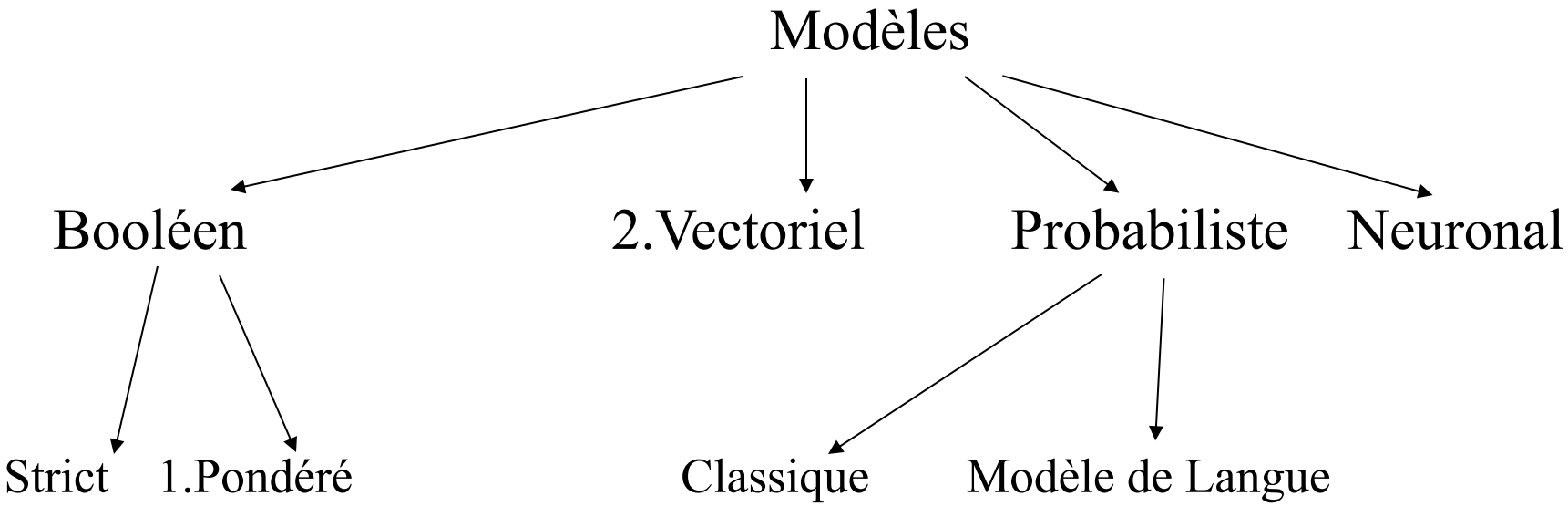
Le système doit

- être simple à utiliser
- fournir les meilleures réponses possibles, et ces réponses doivent être « pertinentes » pour l'utilisateur
  - Pertinence système versus pertinence utilisateur
- fournir un nombre raisonnable de réponses
- fournir des réponses rapides

3. Modèles de RI



3. Modèles de RI



3. Modèles de RI

- Le modèle booléen pondéré (1)
  - Modèle de connaissances : $T = \{t_i\}$, $i \in [1, N]$
    - Termes t_i qui indexent les documents
  - Un document D est représenté par :
    - Une fonction $W_{\mathcal{D}} : T \rightarrow [0,1]$, qui pour chaque terme de T donne le poids de ce terme dans $\mathcal{D}$ (représentation de D). Le poids vaut 0 pour un terme non présent dans le document.
  - Une requête Q est représentée par une formule logique

$$\text{ex.: } Q = (t_1 \wedge t_3) \vee (t_{25} \wedge t_{145} \wedge \neg t_{134})$$

Note: $\wedge$ = ET, $\vee$ = OU, $\neg$ = NON

3. Modèles de RI

- Le modèle booléen pondéré (2)
  - Fonction de correspondance notée *Sim*
    - Formules inspirées de la logique floue
    - Requêtes simples (avec a et b des termes t_i quelconques)
      - » $Sim(\mathcal{D}, (a \wedge b)) = \min [W_{\mathcal{D}}(a), W_{\mathcal{D}}(b)]$
      - » $Sim(\mathcal{D}, (a \vee b)) = \max [W_{\mathcal{D}}(a), W_{\mathcal{D}}(b)]$
      - » $Sim(\mathcal{D}, (\neg a)) = 1 - W_{\mathcal{D}}(a)$
    - Requêtes complexes (x et y sont des sous-requêtes):
      - » $Sim(\mathcal{D}, (x \wedge y)) = \min [Sim(\mathcal{D}, x) , Sim(\mathcal{D}, y)]$
    - Limitation : on ne tient pas compte dans la réponse de tous les termes de la requête
      - On a: min(0.3, 0.5) = min(0.3, 1)

3. Modèles de RI

- Le modèle booléen pondéré (3)
  - Ex. avec des poids binaires pour les document

			Sim	
Documents	**a**	**b**	**$a \vee b$**	**$a \wedge b$**
D_1	**1**	**1**	**1**	**1**
D_2	**1**	**0**	**1**	**0**
D_3	**0**	**1**	**1**	**0**
D_4	**0**	**0**	**0**	**0**

3. Modèles de RI

- Le modèle booléen pondéré (4)
  - Ex. avec des poids non-binaires pour les documents

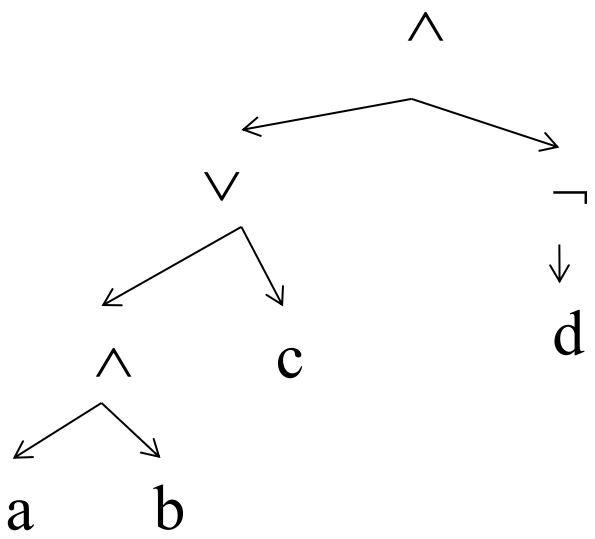
			Sim	
Documents	**a**	**b**	**a ∨ b**	**a ∧ b**
D_1	**1**	**1**	**1**	**1**
D_2	**0.8**	**1**	**1**	**0.8**
D_3	**0**	**0.5**	**0.5**	**0**
D_4	**0.8**	**0**	**0.8**	**0**

3. Modèles de RI

- Le modèle booléen pondéré (5)
  - Traitement de requête (totalement parenthésée)
    - Exemple : (((a $\wedge$ b) $\vee$ c) $\wedge \neg$ d)
  - Etape 1 : génération de l'arbre de requête
    - Une parenthèse (on a un opérateur binaire $\wedge$ ou $\vee$) :
      - un nœud non-feuille avec 2 fils, décoré avec l'opérateur
      - Fils gauche : sous-arbre avec partie gauche
      - Fils droit : sous-arbre avec partie droite
    - Un terme : noeud feuille avec le terme
    - Un $\neg$: un nœud non-feuille unaire avec sous-arbre fils
  - Génération des opérations

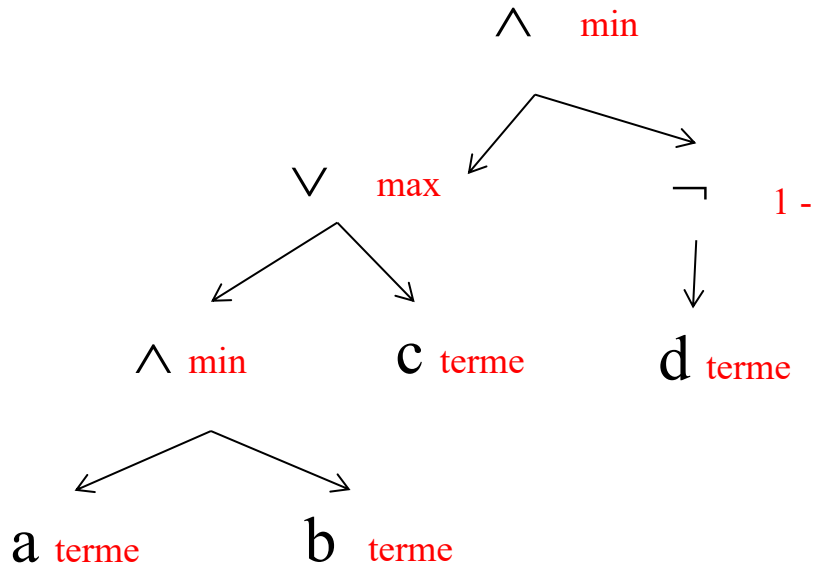
3. Modèles de RI

- Le modèle booléen pondéré (6)
  - Arbre de requête pour (((a $\wedge$ b) $\vee$ c) $\wedge$ $\neg$ d)



3. Modèles de RI

- Le modèle booléen pondéré (7)
  - Les opérations :



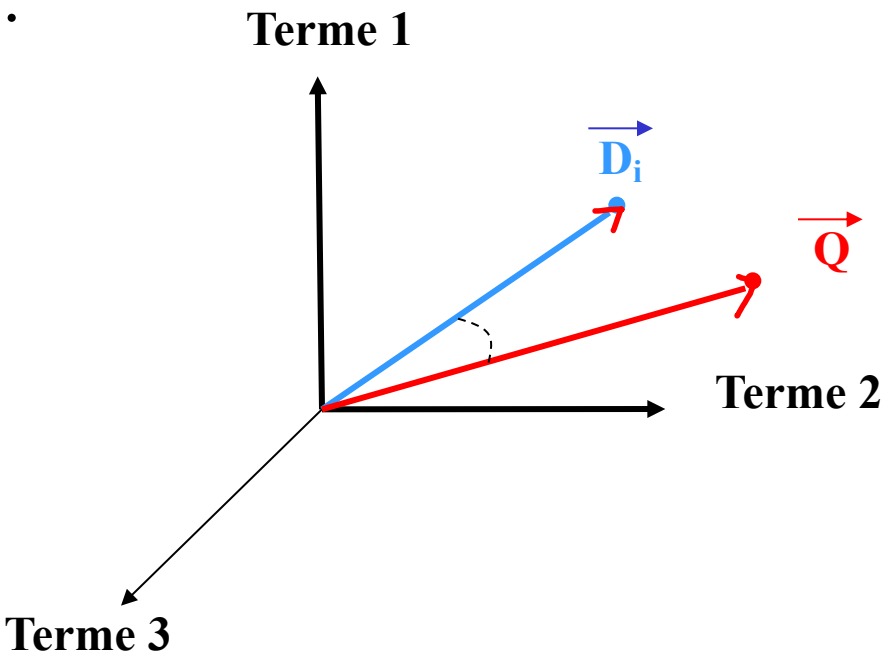
→ Opérations effectuées :
$\min(\max(\min(W_{\mathcal{D}}(a), W_{\mathcal{D}}(b)), W_{\mathcal{D}}(c)), 1-W_{\mathcal{D}}(d))$

3. Modèles de RI

- Le modèle vectoriel (1)
  - Modèle de connaissances : $T = \{t_j\}, j \in [1, N]$
  - Tous les documents sont décrits suivant ce vocabulaire
  - Un document Di est représenté par un vecteur $\overrightarrow{Di}$ décrit dans l'espace vectoriel $\mathbb{R}^N$ défini par T :
    - $\overrightarrow{Di} = (w_{i,1}, w_{i,2}, \ldots, w_{i,j}, \ldots, w_{i,N})$, avec w_{kl} le *poids* d'un terme pour un document
  - Une requête Q est représentée par un vecteur $\overrightarrow{Q}$ décrit dans l'espace vectoriel $\mathbb{R}^N$ défini par T :
    - $\overrightarrow{Q} = (w_{Q,1}, w_{Q,2}, \ldots, w_{Q,j}, \ldots, w_{Q,N})$

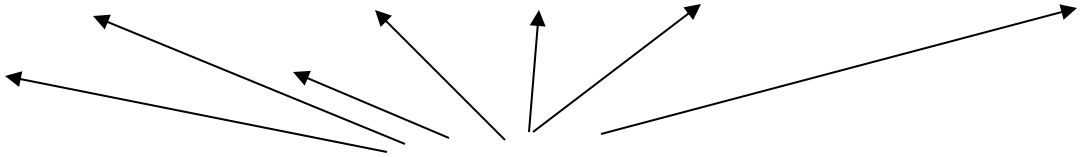
3. Modèles de RI

- Modèle vectoriel (2)
  - Plus les vecteurs représentant les documents/requêtes sont « proches », plus les documents/requêtes sont similaires :



3. Modèles de RI

- Modèle vectoriel (3) – Poids des termes
  - Un document
    - « Un violon est composé de bois précieux comme l'érable, le palissandre, l'ébène... »
  - Pour indexer, la première idée est de compter les mots les plus fréquents excepté les termes non significatifs comme « de », « avec », « comme »…
    - « Un violon est composé de bois précieux comme l'érable, le palissandre, l'ébène... »



Termes retenus et comptés

3. Modèles de RI

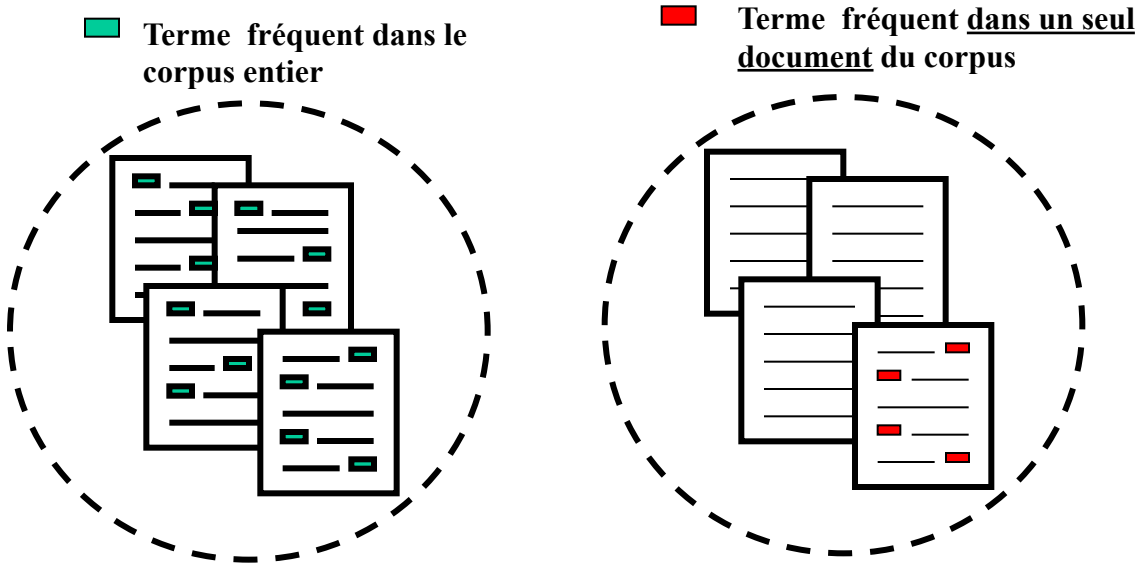
- Modèle vectoriel (4) – Poids des termes
  - On définit la "fréquence" d'un terme (term frequency)
    - caractérise le terme dans un document

    - **$tf_{i,j}$: la "fréquence" du terme t_j dans le document D_i est égale au nombre d'occurrences de t_j dans D_i.**

    - Exemple : si violon apparaît 5 fois dans le document D_3, avec violon=t_{23}, alors $tf_{3,23} = 5$

3. Modèles de RI

- Modèle vectoriel (5) – Poids des termes
  - On tient compte du corpus (base de documents) entier, un terme qui apparaît beaucoup ne discrimine pas nécessairement les documents :



3. Modèles de RI

- Modèle vectoriel (6) – Poids des termes
  - On définit la "fréquence" documentaire d'un terme
    - Caractérise le terme dans le corpus
    - df_j : la fréquence dans le corpus du terme t_j est le nombre de documents du corpus où t_j apparaît
  - On utilise l'

    inverse de la fréquence documentaire, idf_j

    - Définition simple : $idf_j = 1 / df_j$
    - Définition la plus utilisée : $idf_j = \log(N_D / df_j)$, avec N_D le nombre de documents du corpus.

    (idf : inverse document frequency)

3. Modèles de RI

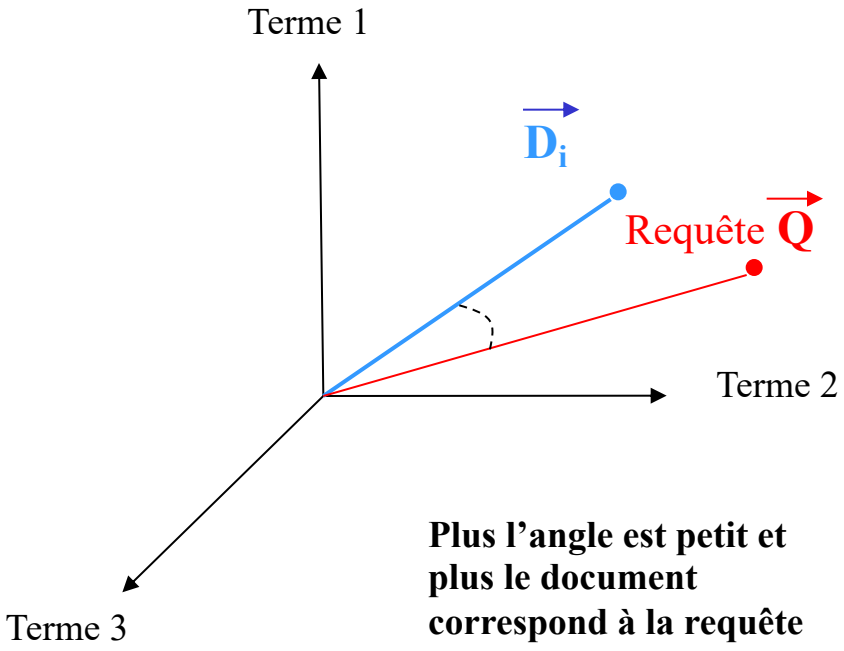
- Modèle vectoriel (7) – Poids des termes
  - Combinaison du tf et de l'idf pour un vecteur document:
    - Le poids d'un terme dénote la capacité du terme à discriminer les documents et à décrire un document
    - Exemple le plus courant

$$w_{i,j} = tf_{i,j} \cdot idf_j$$

  - Utilisation tf.idf aussi pour une requête

3. Modèles de RI

- Modèle vectoriel (8) – Poids des termes
  - Fonction de correspondance : fonction de l'angle entre le vecteur requête $\vec{Q}$ et le vecteur document $\vec{D}_i$



3. Modèles de RI

- Modèle vectoriel (9) :
  - Fonction de correspondance : le *cosinus* de l'angle entre le vecteur requête et le vecteur document.

$$\mathrm{Sim}(\overrightarrow{D_i}, \vec{Q}) = \frac{\sum_{k=1}^{N}(w_{i,k} \cdot w_{q,k})}{\sqrt{\sum_{k=1}^{N} {w_{i,k}}^2} \cdot \sqrt{\sum_{k=1}^{N} {w_{q,k}}^2}}$$

$$= \frac{\overrightarrow{D_i} \circ \vec{Q}}{\|\overrightarrow{D_i}\| . \|\vec{Q}\|}$$

  - Note : les $w_{q,k} = 0$ et $w_{i,k} = 0$ n'impactent pas le numérateur (utile pour les *fichiers inverses*)

3. Modèles de RI

- Modèle vectoriel (10)
  - On peux aussi normaliser en amont les vecteurs documents et requêtes (gain en vitesse)

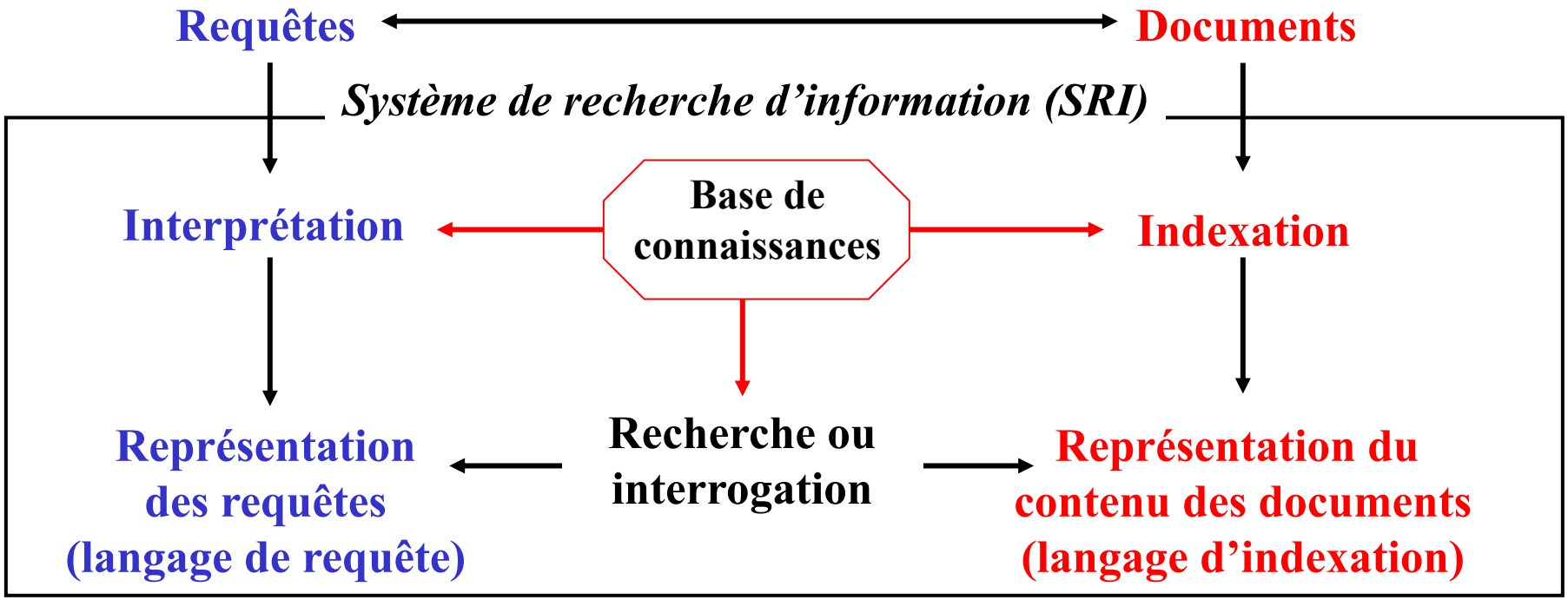
$$\mathrm{Sim}\left(\overrightarrow{D_i}, \vec{Q}\right) = \frac{\sum_{k=1}^{N}(w_{i,k} \cdot w_{q,k})}{\|\overrightarrow{D_i}\| \cdot \|\vec{Q}\|} = \sum_{k=1}^{N}\left(\frac{w_{i,k}}{\|\overrightarrow{D_i}\|} \cdot \frac{w_{q,k}}{\|\vec{Q}\|}\right)$$

$$= \overrightarrow{D'_i} \circ \overrightarrow{Q'}$$

- Le résultat est exactement celui du cosinus initial

4. Systèmes de recherche d'information

- Un SRI est un système informatique qui instancie un modèle de recherche d'information



– Le système doit intégrer les problèmes de vitesse.

4. Systèmes de recherche d'information

- Indexation (1)
  - Choix des termes
    - Une propriété souhaitée d'un bon terme d'indexation est sa capacité à distinguer les documents d'une collection les uns des autres
    - Comment faire?

4. Systèmes de recherche d'information

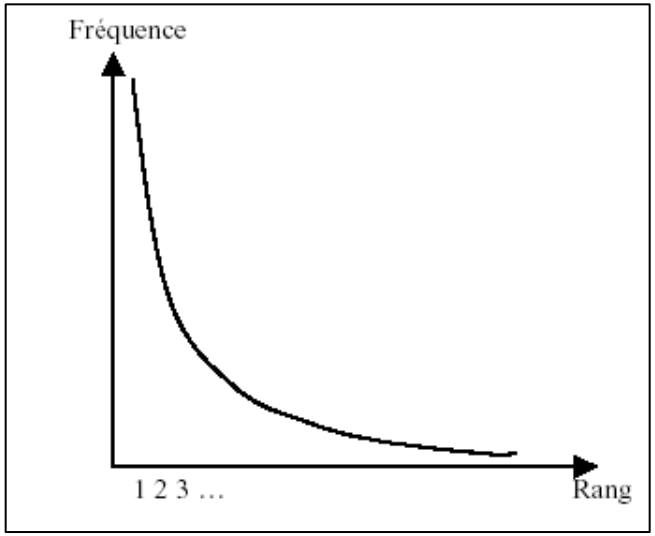
- Indexation (2)
  - Choix des termes – Occurrences (1)
    - **Hypothèse :** Un mot qui apparaît souvent dans un texte représente un concept important.
    - MAIS, on s'aperçoit que les mots les plus fréquents sont des mots fonctionnels (ou mots outils, mots vides). En français, les mots "de", "un", "les", etc. sont les plus fréquents. En anglais, ce sont "of", "the", etc.
    - Ce phénomène est décrit par **la loi de Zipf**.

4. Systèmes de recherche d'information

- Indexation (3)
  - Choix des termes – Occurrences (2)
    - **La loi de Zipf**
    - Si on classe les mots dans l'ordre décroissant de leur fréquence, et on leur donne un numéro de rang (1, 2, …), alors: Rang * fréquence ≈ constante.

Rang	Mot	Fréquence	Rang* Fréquence
1	the	69 971	69 971
2	of	36 411	72 822
3	and	28 852	86 556
4	to	26 149	104 596
5	a	23 237	116 185
6	in	21 341	128 046
7	that	10 595	76 165

    - La distribution de mots suit la courbe :



    - les termes "utiles" : ni trop rares (place en mémoire), ni trop présents (pas discriminants)…
    - Calcul coûteux si on le fait à chaque collection considérée

4. Systèmes de recherche d'information

- Indexation (4)
  - Choix des termes – Listes a priori
    - Anti-dictionnaire (liste de mots à ne pas garder) inspiré de Zipf
      - au, aux, avec, ce, ces, dans, de, des, du, elle, en, etc, et, eux, il, je, la, le, les, leur, lui ...
      - Ne garder que des termes qui ont du sens, diminuer la taille des index
    - Extraction de troncatures des mots du texte :
      - Algorithme de Porter (anglais) :

        - Règles (exemples)
        » s → /
        » ed → /
        » ing → /
        » er → /
        » e → /

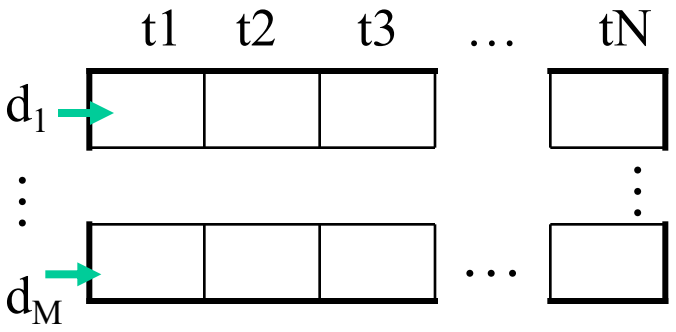
Mot initial	Mot tronqué
engineered	engin
engineer	engin
engineers	engin
informing	inform
computer	comput
computing	comput

      - Diminuer la taille des index, grouper les termes « similaires »

4. Systèmes de recherche d'information

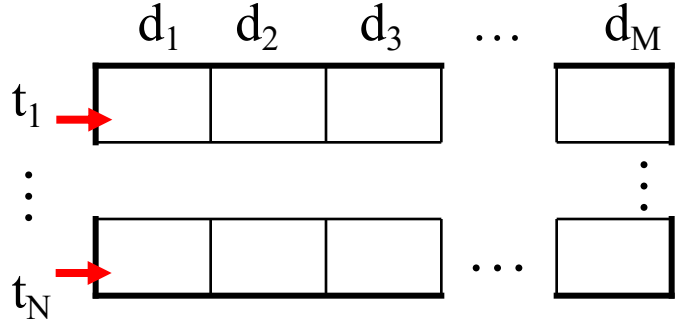
- Indexation (5) – préparation de recherche rapide
  - Fichiers inverses - principe
    - Par analyse des documents d'un corpus, on obtient un tableau document x termes
    - Utilisation en tableau direct

      « document -> terme »



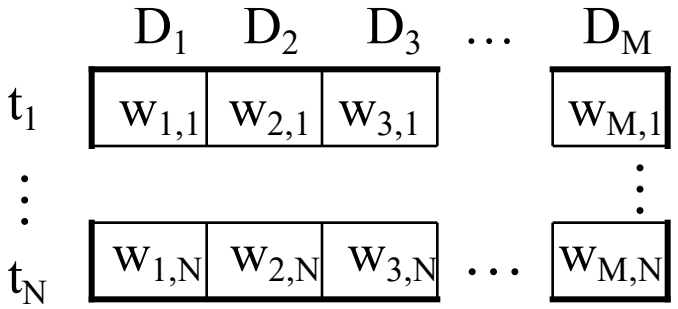
    - Génération d'un tableau inverse « terme -> document » (appelé fichier inverse)

      – Avantage : rapidité lors du traitement de requête, car pas de traitement séquentiel des documents



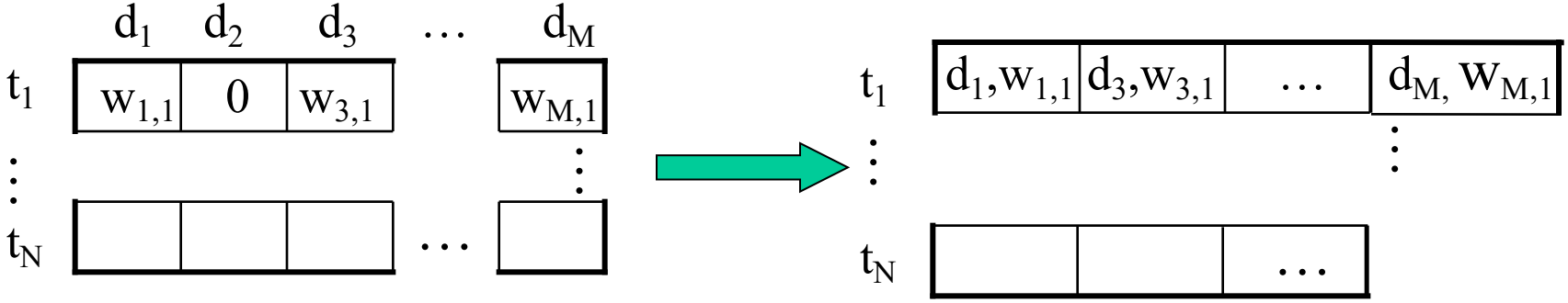
4. Systèmes de recherche d'information

- Indexation (6)
  - Fichier inverse avec les modèles pondérés (booléen pondéré, vectoriel) : version simple



4. Systèmes de recherche d'information

- Indexation (7)
  - En fait il y a beaucoup de valeurs nulles dans le fichier inverse (> 90% des cases du tableau) :
    - Représentation optimisée possible : utiliser des représentations de tableaux creux (tableau avec tailles de lignes différents, listes chaînées)



4. Systèmes de recherche d'information

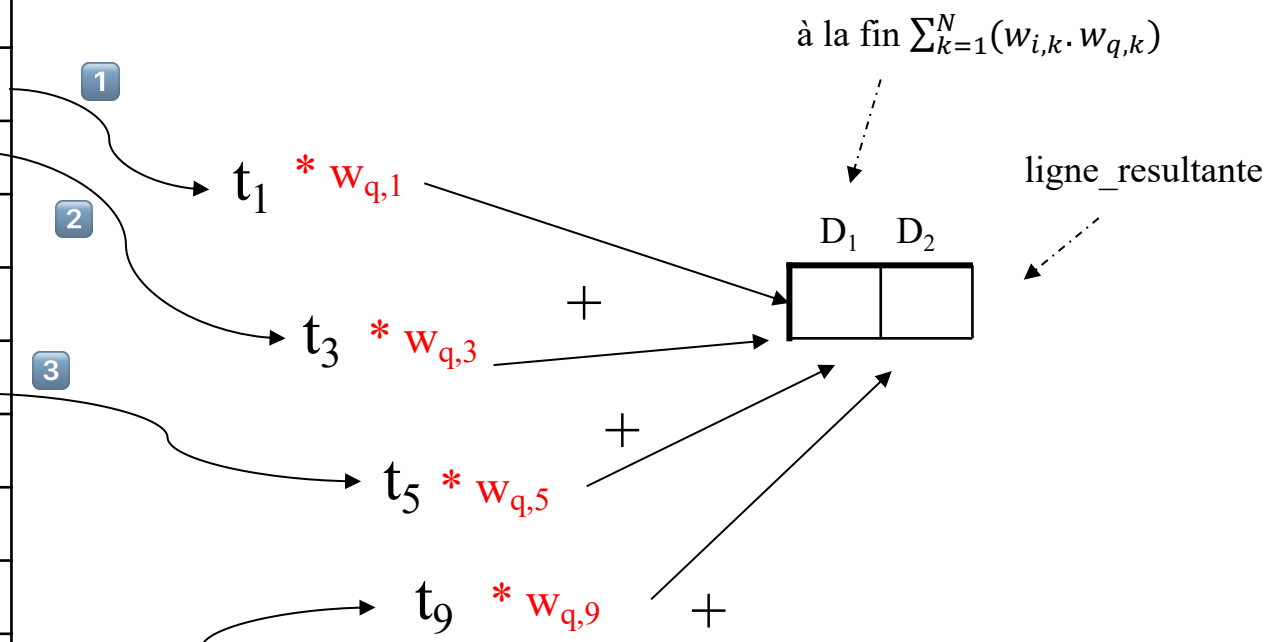
- Indexation (8)
  - Pour une implantation d'un modèle vectoriel, on stocke aussi pour chaque terme du vocabulaire son idf (utilisé pour les requêtes) : utilisation lors de la définition du vecteur requête.

4. Systèmes de recherche d'information

- Recherche (1) par fichier inverse (vectoriel)
  - Ex. : Q contient les termes t_1, t_3, t_5 et t_9, les poids w_q non-nuls sont $w_{q,1}$, $w_{q,3}$, $w_{q,5}$ et $w_{q,9}$.

Fichier inverse (non-optimimsé)

Terme	D1	D2
t1	0,1	0,2
t2	0	1
t3	0,9	0
t4	0,2	0,3
t5	0,5	0,3
t6	0	0
t7	0,7	0
t8	0,6	0
t9	0	0,1
t10	0,8	0,9



4. Systèmes de recherche d'information

- Recherche (2)
  - Implantation simple du modèle vectoriel par fichier inverse
    - Requête Q = ($w_{q,1}$ … $w_{q,N}$)
    - On garde les termes t_i tels que $w_{q,i} \neq 0$
    - Boucle pour chaque terme t_i et pour les documents D_j 1…
      - ligne_resultante[j] += $w_{j,i}$ * $w_{q,i}$
        (utilisation du fichier inverse pour les $w_{j,i}$)
    - Calcul final : resultat[j] = ligne_resultante[j] / ($||D_j||$. $||Q||$)
    - Tri des résultats par ordre décroissant et affichage

$$\text{Sim}(\vec{D_i}, \vec{Q}) = \frac{\sum_1^N (w_{i,k}, w_{q,k})}{\sqrt{\sum_1^N w_{i,k}^2} \cdot \sqrt{\sum_1^N w_{q,k}^2}}$$

4. Systèmes de recherche d'information

- Recherche (3)
  - Bouclage de pertinence (dans le modèle vectoriel)
    - Idée : l'utilisateur peut montrer ce qui est pertinent pour lui-même s'il ne peut pas l'exprimer explicitement
    - Mise en œuvre
      - Une requête initiale Q0 fournie par l'utilisateur renvoie une réponse
      - L'utilisateur indique dans la liste des réponses les documents pertinents et non pertinents pour lui
      - Le système génère une nouvelle requête Q1 qui prend en compte Q0 et les pertinents et les non pertinents indiqués

4. Systèmes de recherche d'information

- Recherche (9)
  - Bouclage de pertinence dans le modèle vectoriel
    - Formule de Rocchio

$$Q1 = \alpha Q0 + \beta\left(\frac{1}{|D_R|}\sum_{d \in D_R} d\right) - \gamma\left(\frac{1}{|D_N|}\sum_{d' \in D_N} d'\right)$$

      - Avec
        - » D_R l'ensemble des docs marqués pertinents par l'utilisateur
        - » D_N l'ensemble des docs marqués non-pertinents par l'utilisateur
        - » $\alpha \geq \beta \geq \gamma$
        - » valeurs possibles α =1 β =0.4 γ =0.2, ou même 1 1 0

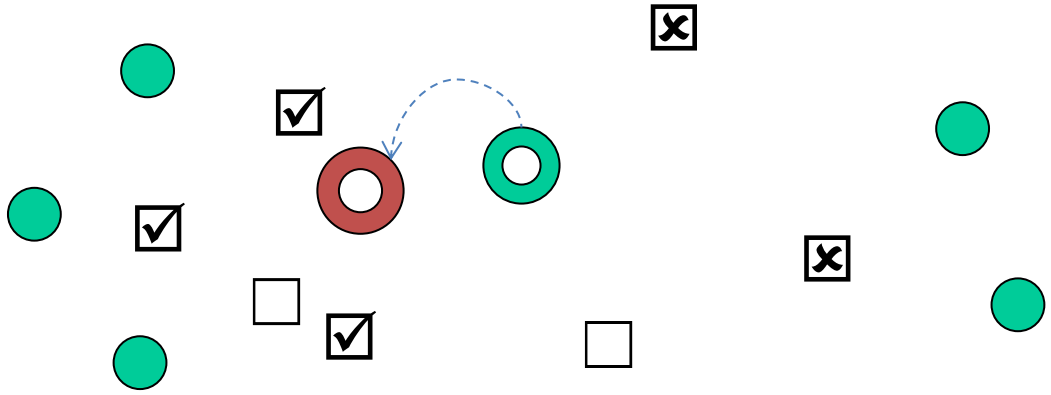
    - Le bouclage de pertinence donne de très bons résultats.

4. Systèmes de recherche d'information

- Recherche (10)
  - Exemple de Bouclage de pertinence

Requête initiale : 
Document : 
Document retrouvé : ☐

Requête modifiée : 
Document retrouvé pertinent : ☑
Document retrouvé non pertinent : ☒



5. Evaluation des SRI

- Rappel : Un système de recherche d'information doit satisfaire un besoin d'information d'un utilisateur

pertinence utilisateur ≠ pertinence système

- Pertinence utilisateur : satisfaction de l'utilisateur
- Pertinence système : estimation du système

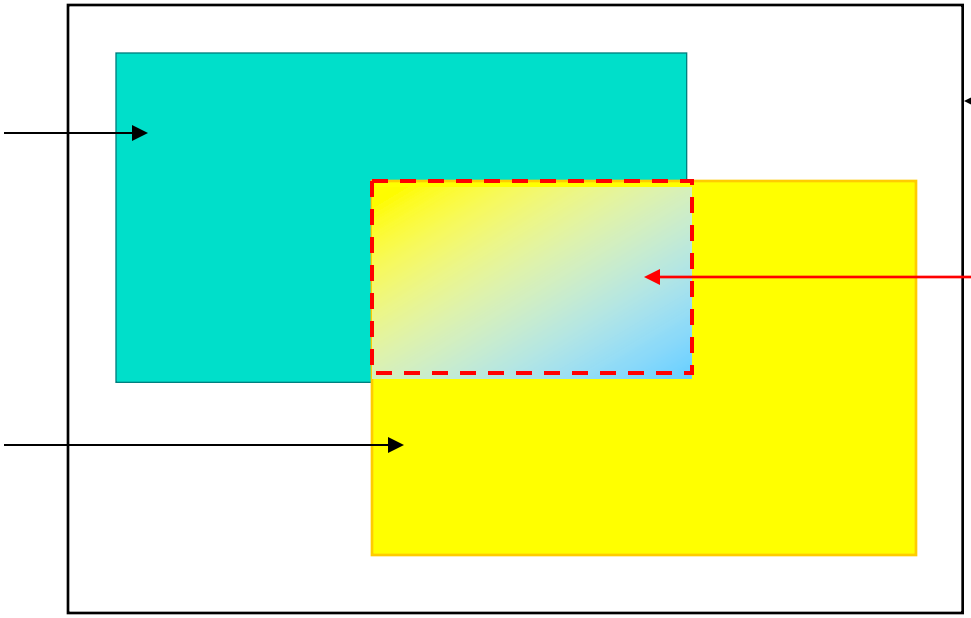
5. Evaluation des SRI

- Objectifs :
  - Déterminer si mon système est bon
  - Déterminer si mon système est meilleur qu'un autre

- Evaluation de type « boîte noire » en comparant les résultats du système testé par rapport à des réponses idéales

5. Evaluation des SRI

- Objectif :
  - Rapprocher pertinence système et utilisateur POUR UNE REQUETE



Ensemble des documents trouvés, pertinents pour **le système : R**

Ensemble des documents convenant à **l'utilisateur: P**

Ensemble des documents du corpus : **C**

Ensemble des documents **retrouvés et convenant** à l'utilisateur, pertinents pour **le système et l'utilisateur : P∩R**

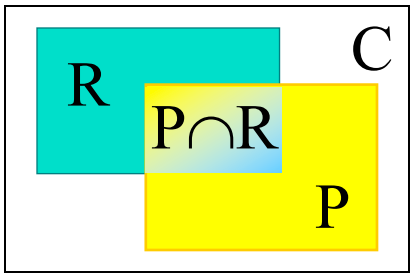
5. Evaluation des SRI

- Les critères essentiels sont :
  - Le rappel : capacité du système à fournir en réponse tous les documents pertinents pour l'utilisateur
  - La précision : capacité du système à ne fournir que des documents pertinents pour l'utilisateur en réponse.

  - Ces deux critères sont antagonistes dans la réalité…

5. Evaluation des SRI

- Le rappel est le rapport du nombre de documents trouvés par le système et convenant à l'utilisateur au nombre total de documents convenant à l'utilisateur

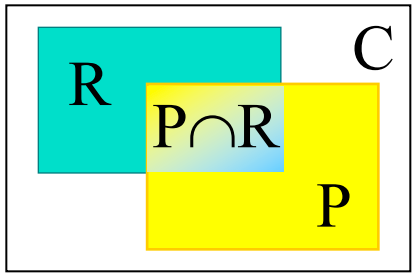


$$rappel = \frac{|P \cap R|}{|P|} \qquad \in [0,1]$$

5. Evaluation des SRI

- La précision est le rapport du nombre de documents trouvés par le système et convenant à l'utilisateur au nombre de documents retrouvés par le système

$$précision = \frac{|P \cap R|}{|R|} \qquad \in [0,1]$$

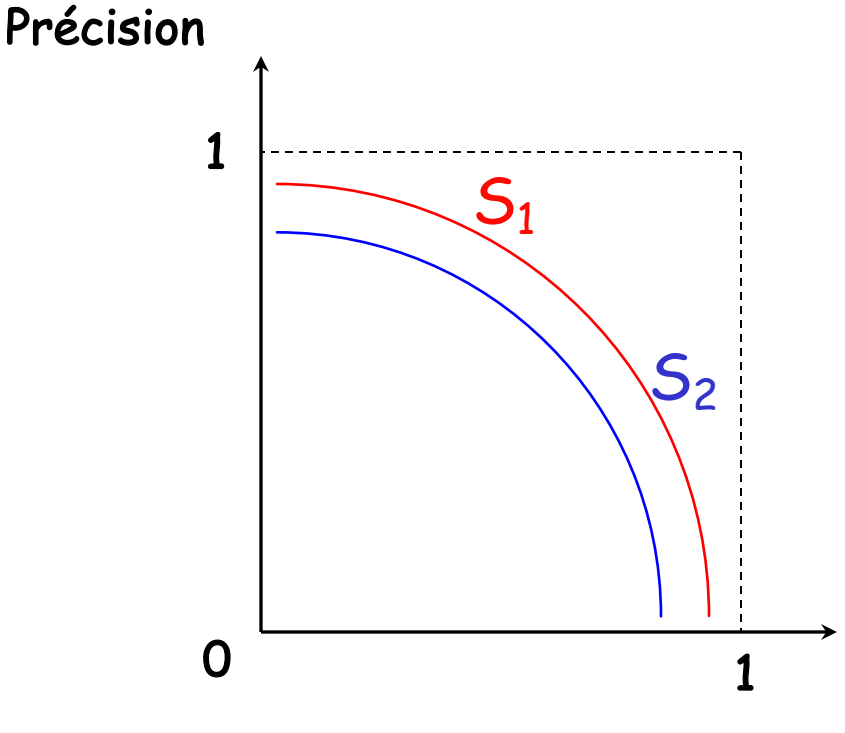


5. Evaluation des SRI

- Pour une requête et un système : 2 valeurs réelles
  - Exemple : un système retourne 5 documents, parmi lesquels 3 sont pertinents, sachant qu'il y a 10 documents pertinents dans le corpus :
    - Rappel = 3 / 10
    - Précision = 3 / 5
- Il faut des analyses plus fines des résultats
  - Courbes de rappel/précision

5. Evaluation des SRI

- Courbes de rappel/précision
  - Comparaison de 2 systèmes S1 et S2



- S_1 est meilleur que S_2 en précision et en rappel

5. Evaluation des SRI

- Courbes de rappel/précision
  - Représente l'évolution de la précision et du rappel avec des résultats triés
  - Méthode :
    - Pour chaque document retrouvé, on calcule la précision et le rappel obtenus en considérant seulement le premier document comme réponse, puis les deux premiers, puis les trois premiers etc., jusqu'à la réponse totale du système.

      → Ceci donne un tableau de rappel/précision non-normalisé

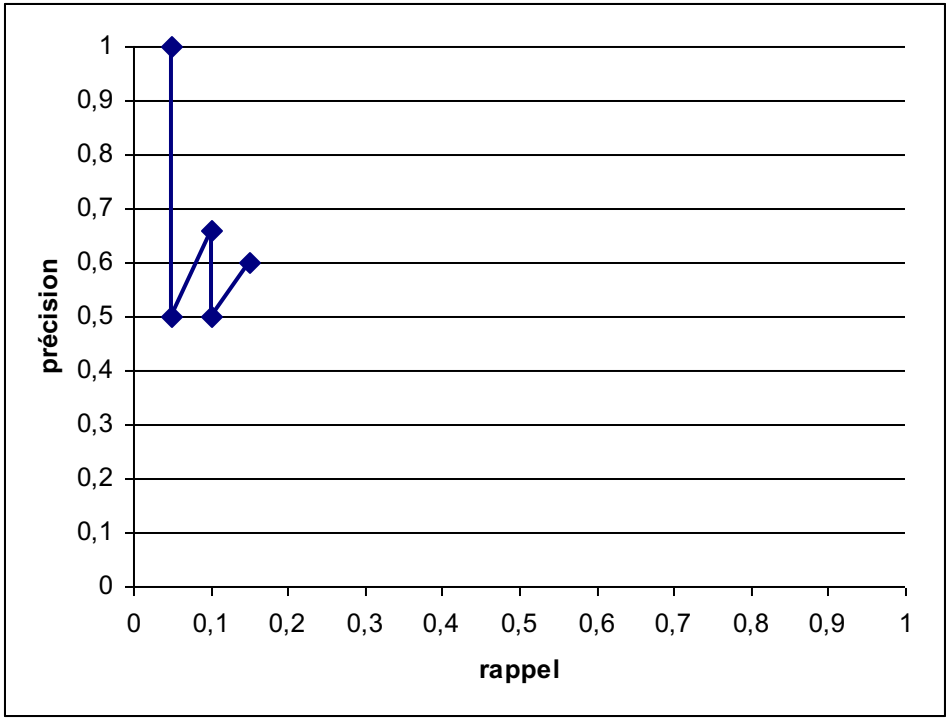
5. Evaluation des SRI

- Courbes de rappel/précision
  - Exemple de tableau non normalisé (hypothèse : 20 documents pertinents)

Doc	P?
D_{23}	Oui
D_{12}	Non
D_5	Oui
D_3	Non
D_7	Oui

Rap.	Préc.
0.05	1
0.05	0.5
0.1	0.66
0.1	0.5
0.15	0.6

Tableau non normalisé



5. Evaluation des SRI

- Courbes de rappel/précision
  - Problème, comment faire pour fusionner les résultats de plusieurs requêtes pour un système?
  - On normalise la courbe de chaque requête par <u>la règle du maximum</u> :
    1. On fixe une valeur de rappel normalisée r (dans [0, 0.1, 0.2,…, 0.8, 0.9, 1]);
    2. On garde la valeur de précision max. pour les valeurs de rappels du tableau non normalisé ≥ r, et on la met dans le tableau normalisé pour le rappel r.

Tableau non normalisé →

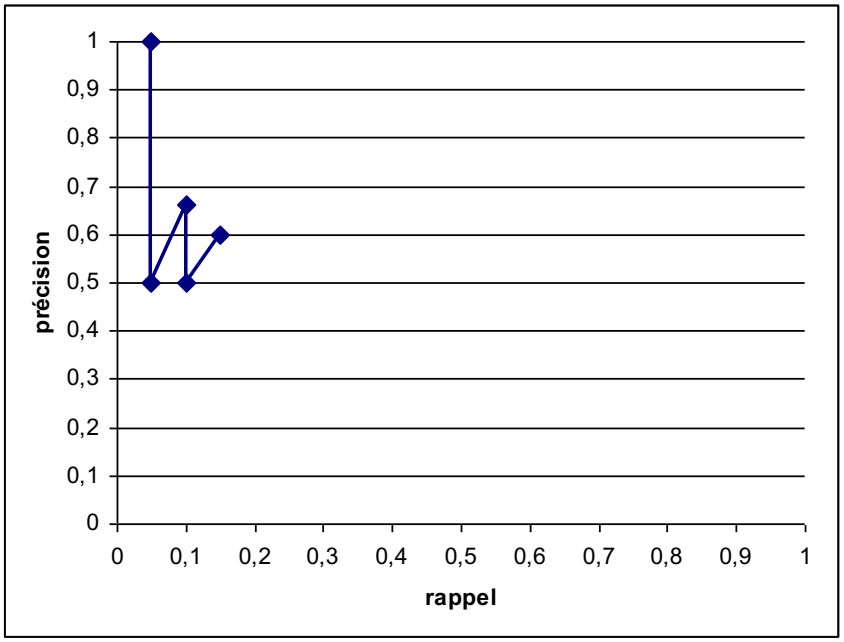
Rap.	Préc.
0.05	1
0.05	0.5
0.1	0.67
0.1	0.5
0.15	0.6

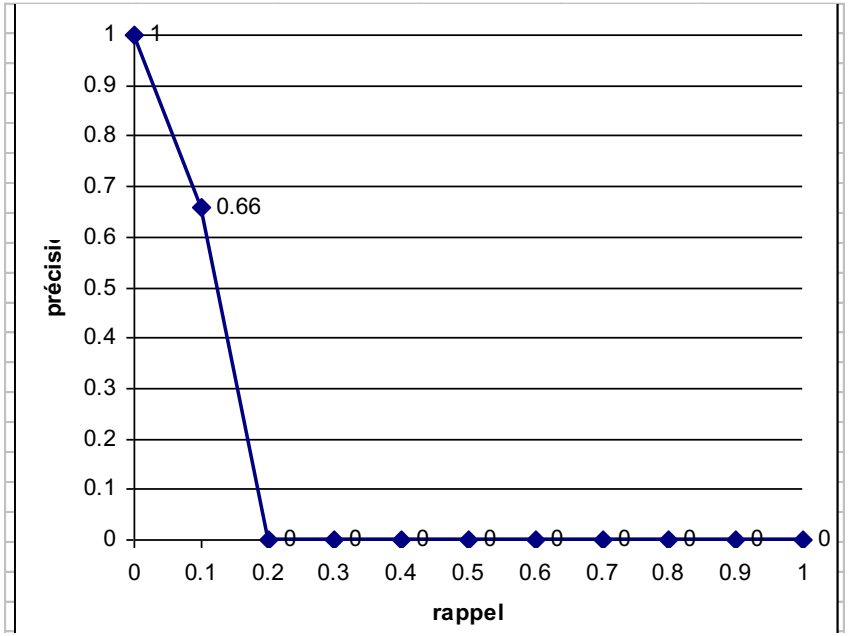
Tableau normalisé →

Rap.	Préc.
0	1
0.1	0.67
0.2	0
0.3	0
0.4	0
0.5	0
0.6	0
0.7	0
0.8	0
0.9	0
1	0

5. Evaluation des SRI

- Courbes de rappel/précision
  - au niveau graphique



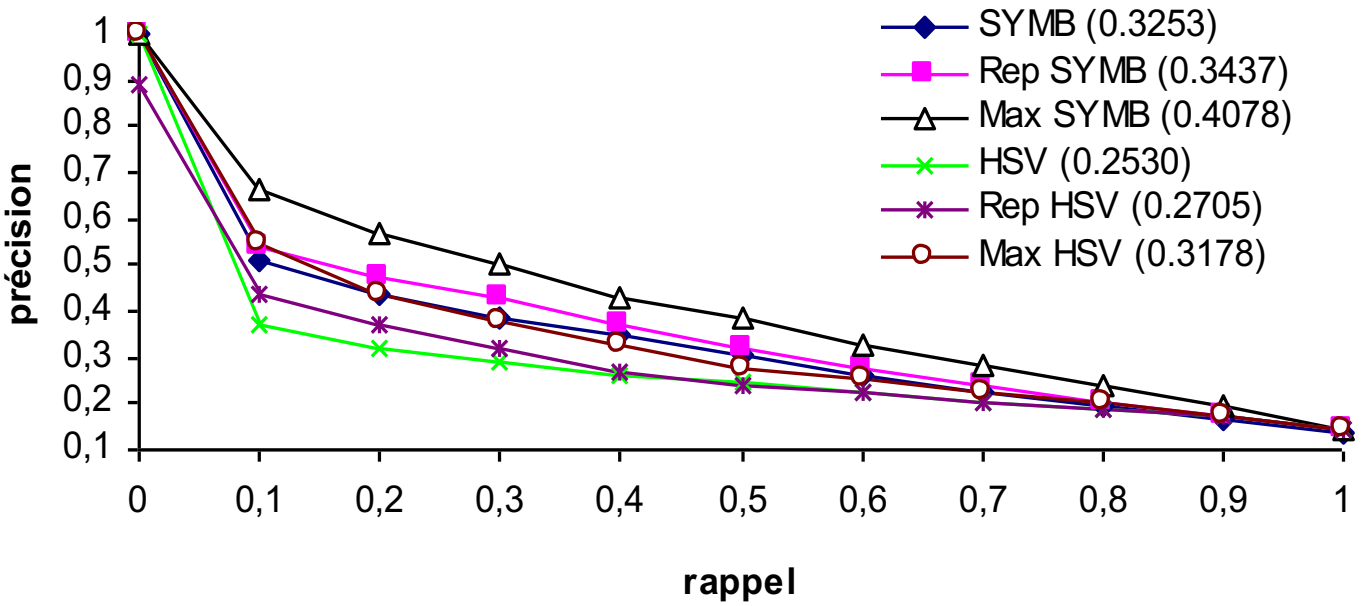


5. Evaluation des SRI

- Courbes de rappel/précision
  - Pour traiter l'évaluation sur plusieurs requêtes, on calcule la moyenne aux points de rappels standards. (moyenne au point 0, au point 0.1, etc.)
- Précision moyenne à *x* documents
  - Il est également courant de calculer le taux de précision après un nombre de documents *x* fixé pour une requête, puis de faire la moyenne sur toutes les requêtes
- Il existe des programmes qui génèrent les tableaux pour les courbes de rappel/précision et les précision moyennes à 5, 10, 20, 50 et 100 documents. (trec_eval)

5. Evaluation des SRI

- Exemple de « vraie » courbe de rappel/précision



  - Question : quel système est le meilleur pour les courbes ci-dessus?

6. Conclusion

- Nous avons traité la description de modèles de recherche d'information.
- Nous avons étudié les principes de base des systèmes basés sur ces modèles
- Nous avons compris comment réaliser une évaluation d'un système de recherche d'information.