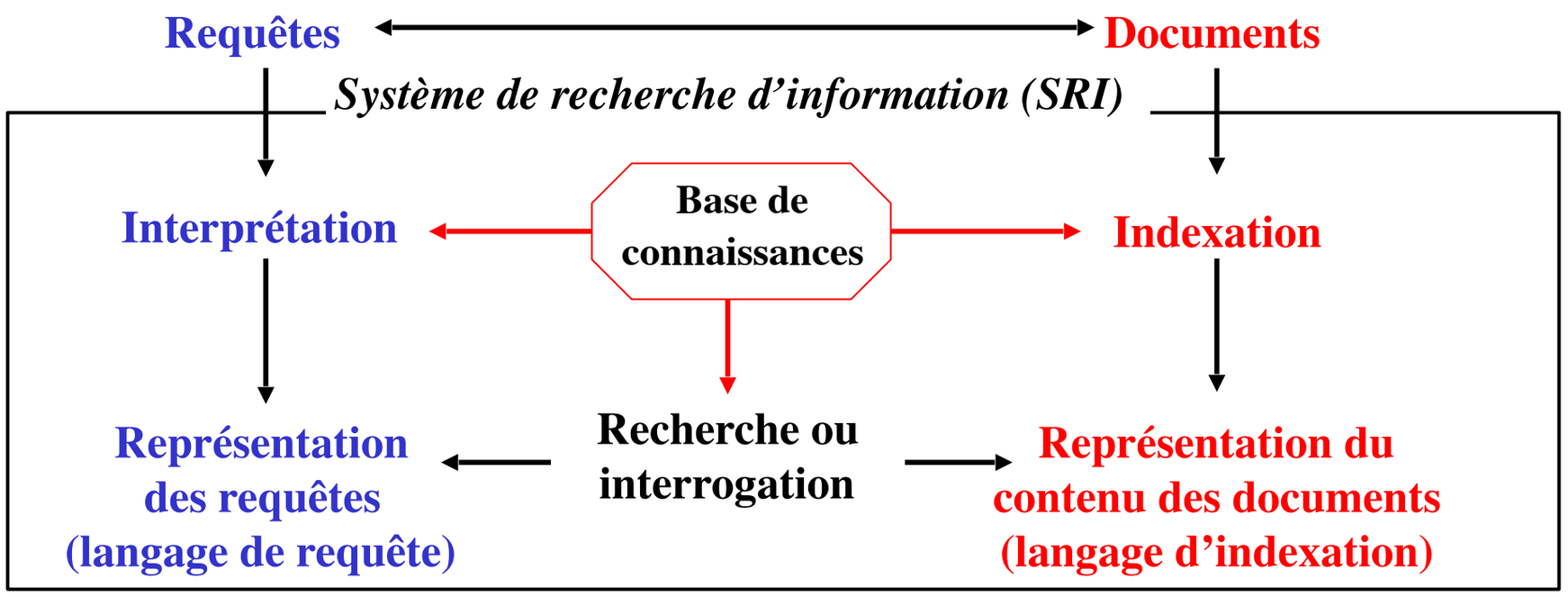
# Plan

1. Introduction
2. Elements clés en recherche d'information (RI)
3. Modèles de RI
4. Systèmes de recherche d'information (SRI)
5. Evaluation des SRI

# Systèmes de recherche d'information

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



# Indexation

Comment choisir les termes à indexer ?

Une propriété souhaitée d'un bon terme d'indexation est sa capacité à distinguer les documents d'une collection les uns des autres

- **Objectif** : on veut trouver les mots qui représentent le mieux le contenu d'un document.

- **Hypothèse** : on admet généralement qu'un mot qui apparaît souvent dans un texte représente un concept important.

- La première approche consiste :

  - à choisir les mots représentatifs selon leur fréquence d'occurrence dans le corpus.

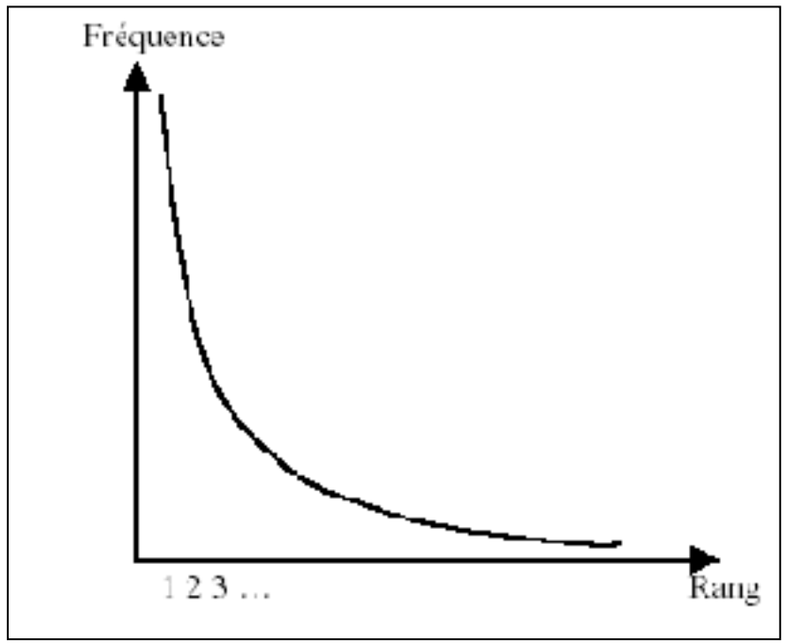
  - à définir un seuil $S_{MIN}$ sur la fréquence : si la fréquence d'occurrence d'un mot dépasse ce seuil, alors il est considéré important pour les documents qui le contiennent.

# Indexation

- Les mots les plus fréquents ou mots outils ne sont pas discriminants

- Loi de Zipf :

- La distribution de mots :




- L'idée peut être alors de garder les termes"utiles" : ni trop rares (place en mémoire), ni trop présents (pas discriminants)…

# Indexation

Quels termes indexer ?

- Utilisation d'un anti-dictionnaire, aussi appelé liste de mots outils

  {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 :

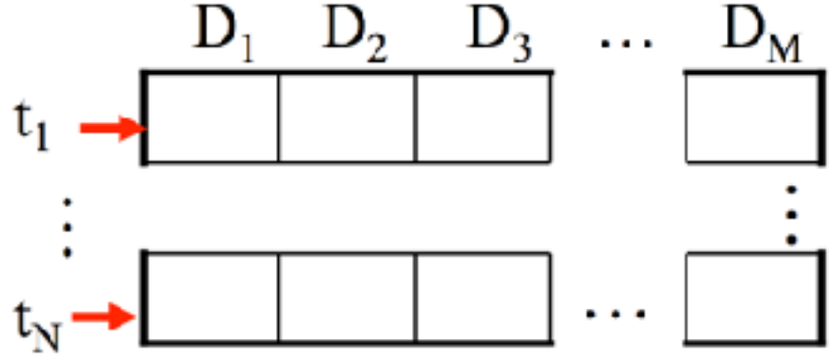
  – Diminuer la taille des index, grouper les termes « similaires »

| Mot initial | Mot tronqué |
|---|---|
| computer | comput |
| computing | comput |
| engineer | engin |
| engineers | engin |
| engineered | engin |

# Indexation

Fichiers inverses - principe

- Sélection des termes : tableau document x termes
- Fichier inverse : tableau terme —> document

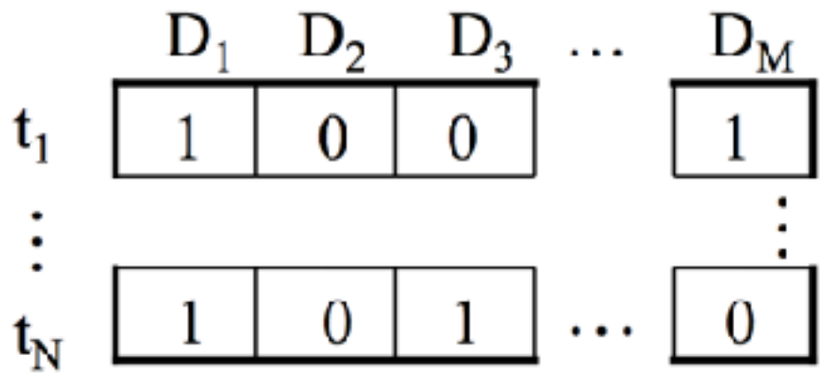



Avantage : rapidité lors du traitement de requête, car pas de traitement séquentiel des documents

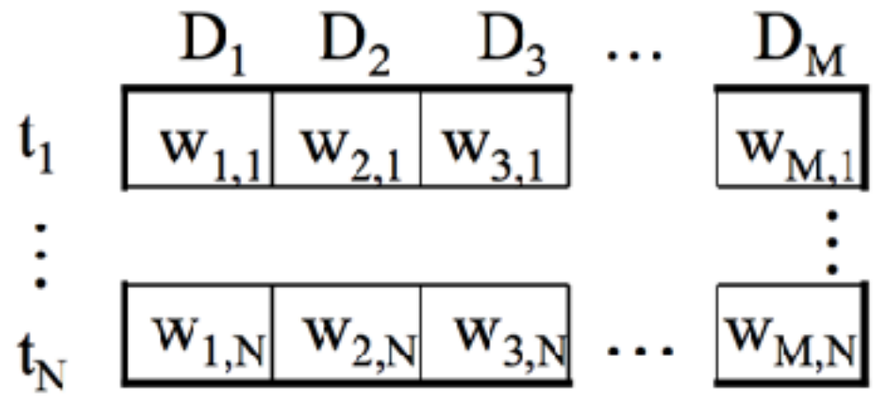
- Création en au moins 2 passes : la première permet de déterminer tous les termes, et la seconde construit le tableau

# Indexation

Variété des index selon les modèles de RI :




Modèle booléen strict



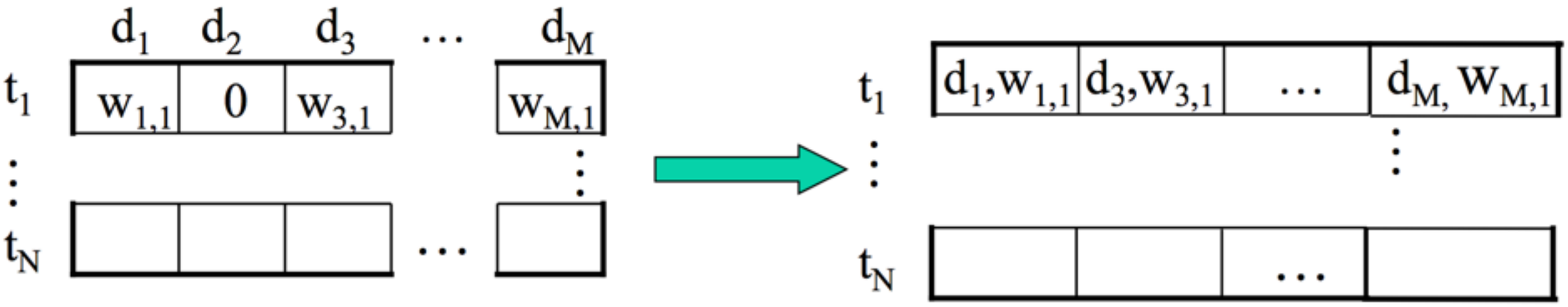

Modèles pondérés
(booléen, vectoriel)

Pour un modèle vectoriel avec tf.idf, la première étape, en plus de trouver les termes, calcule l'idf de ces termes.

# Indexation

- En réalité, les fichiers inverses ne stockent pas toutes les valeurs, car il y a beaucoup de valeurs nulles (> 90% des cases du tableau)

- On utilise des représentations de tableaux creux (tableau avec tailles de lignes différents, listes chaînées)




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.

# Implémentation du modèle vectoriel

- Soit une requête Q = ($w_{q,1}$ … $w_{q,N}$)
- On garde les termes $t_i$ tels que $w_{q,i} \neq 0$
- Pour chaque terme $t_i$ et pour les documents $D_j$
  - 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] / (|Q| . $|D_j|$)
- Tri des résultats par ordre décroissant et affichage

# Implémentation du modèle vectoriel (avancé)

- normalisation en amont w => w' pour les requêtes et le documents

$$Sim(\vec{D}_i, \vec{Q}) = \frac{\sum_{k=1}^{N} (w_{i,k} . w_{q,k})}{\|\vec{D}i\| . \|\vec{Q}\|} = \sum_{k=1}^{N} (\frac{w_{i,k}}{\|\vec{D}i\|} \cdot \frac{w_{q,k}}{\|\vec{Q}\|})$$

$$= \sum_{k=1}^{N} (w'_{i,k} . w'_{q,k})$$

- On stocke dans le fichier inverse les $w'_{i,k}$ , et donc :
  - Calcul final : resultat[j] = ligne-resultante[j]

# Plan

1. Introduction
2. Elements clés en recherche d'information (RI)
3. Modèles de RI
4. Systèmes de recherche d'information (SRI)
5. Evaluation des SRI

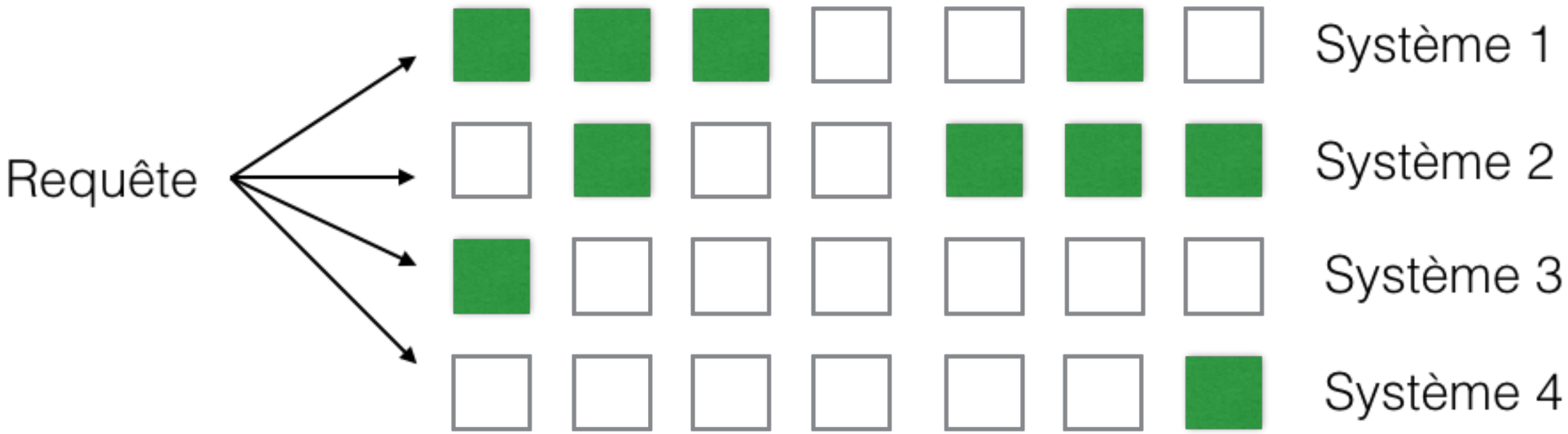
# Evaluation d'un SRI

- 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

# Evaluation d'un SRI




Evaluation à l'échelle du système :

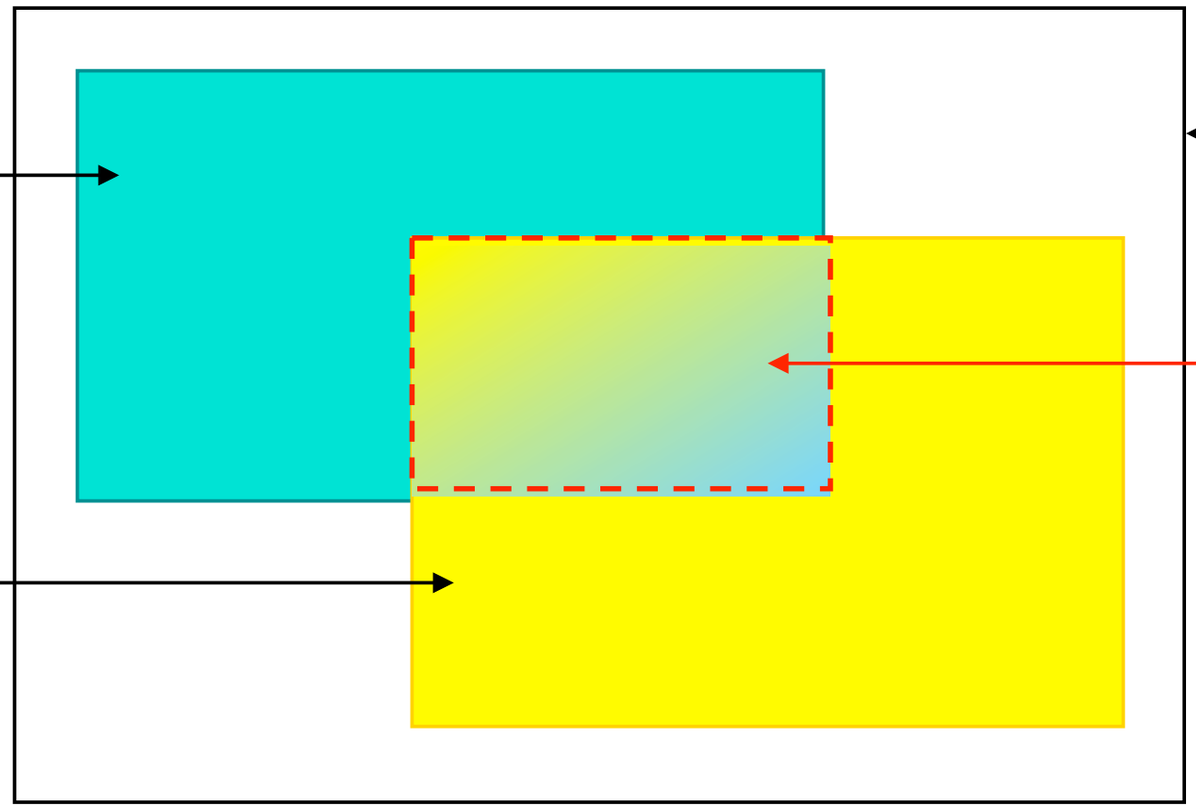
- Quel est le pire système ?
- Quel est le meilleur ?

# Evaluation d'un SRI

Objectif : rapprocher pertinence système et utilisateur

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**

# Evaluation d'un SRI

Les critères essentiels sont :

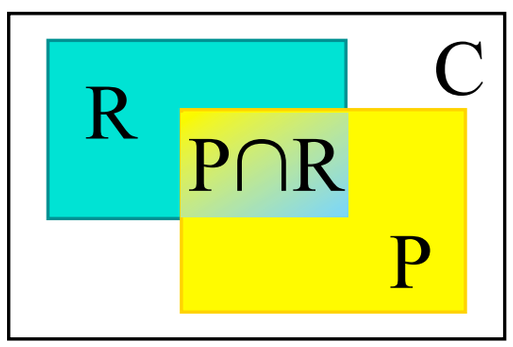
- Le rappel : capacité du système à fournir en réponse tous les documents pertinents

- La précision : capacité du système à ne fournir que des documents pertinents en réponse.

Ces deux critères sont antagonistes dans la réalité…

# Evaluation d'un SRI

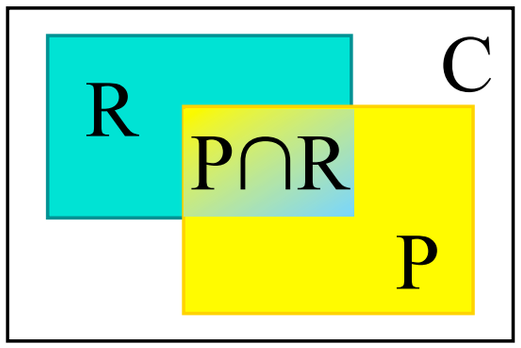
Le rappel est le rapport du nombre de documents trouvés par le système et convenant à l'utilisateur au nombre total de documents convenant à l'utilisateur




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

# Evaluation d'un 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




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

# Evaluation d'un 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

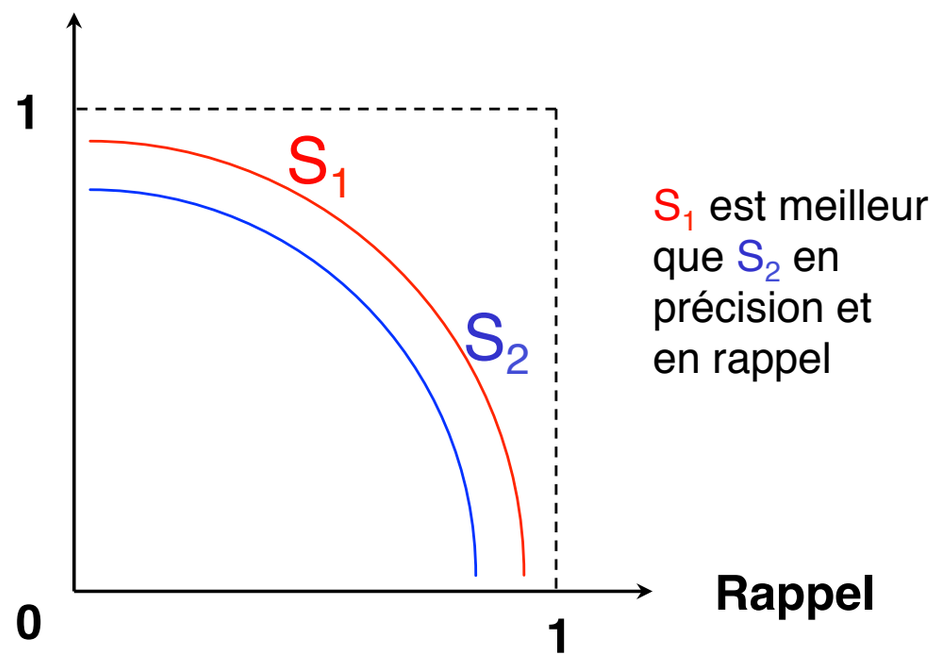
Il faut des analyses plus fines des résultats

- Courbes de rappel/précision

# Evaluation d'un 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 non normalisé)



# Evaluation d’un SRI

Quelles données permettent d'évaluer les SRI ?

1. Observation des logs (pour les moteurs de recherche commerciaux)

   - Recherches sous formes de session (mono ou multi-requêtes)

   - Utilisation de règles (interprétation de la navigation)

   - Ne permet pas de calculer la précision et le rappel, puisque la taille de l'ensemble des documents pertinents pour une requête n'est pas connue

# Evaluation d’un SRI

Quelles données permettent d'évaluer les SRI ?

2. Utilisation de corpus d'évaluation

    - Collections de test créées manuellement, reproduisant des recherches

    - Composées de : requêtes, documents, et jugement de pertinence (permettant de lier les requêtes aux documents qui leur sont pertinents)

    - Les requêtes et les jugements de pertinence sont souvent créés spécifiquement pour une tâche

    - Permet de pouvoir comparer les performances de systèmes sur les mêmes données