La Recherche sur Internet

Comment trouver l'information rapidement et efficacement ?

- La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui visent à représenter des documents dans le but d'en récupérer des informations, au moyen de la construction d’index. Avec le développement de l’Informatique, l’Information est devenue numérique.
- la RI est devenu une préoccupation primordiale de l’Informatique.
  Trouver des documents pertinents pour un besoin d'information à partir d'un ensemble de documents volumineux (Big DATA).

- La Recherche d'Information (RI) ou Information Retrieval en anglais (IR), consiste à trouver des documents peu ou faiblement structurés, dans une grande collection, en fonction d'un besoin d'information.
- C'est la science qui étudie la manière de répondre pertinemment à une requête en retrouvant de l'information dans un corpus.
- la RI a pour thème central l'étude de modèles et systèmes d'interaction entre des utilisateurs humains et des corpus de documents numériques, en vue de la satisfaction de leurs besoins d'information. [

RI :information structurée ?

- La RI fut d'abord le domaine des spécialistes des Sciences de l'Information, généralement l'information est non structurée: l'information textuelle.
- l'informatique était une science de l'information structurée, La RI affrontait l'information textuelle, c'est-à-dire non structurée ou semi-structurée, Ce type d'Information est caractérisée par l'ambiguïté et l'incertitude difficiles à gérer.
- En Général, la RI traite de l'information non structurée surtout de l'information « naturelle », avec toute la complexité que cela implique.

RI: Types d'Information

Actuellement avec le développement des approches scientifiques, l'Informatique s'occupe de 3 types d'information :

1. **Information structurée**

   Information structurée: facile à manipuler, sa nature et sa fonction sont bien identifiées.

On la trouve dans les bases de données et les langages informatiques. Nous reconnaissons les informations structurées au fait qu'elles sont disposées de façon à être traitées automatiquement et efficacement par un logiciel.

Information semi structurée

Par exemple, un courriel (message) est transmis sur Internet dans une forme non structurée . Un courriel contient une combinaison pratiquement égale: d'informations non structurées (le corps du message) et d'informations structurées (date, auteur, destinataire, etc.). Une partie du courriel s'adresse à un humain et l'autre, à une machine.

Une page web partage cette caractéristique : une partie de son contenu s'adresse à l'humain, comme le texte (informations non structurées), alors qu'une autre partie est destinée à la machine, comme les balises (informations structurées)

RI : Information non structurée ?

3. Information non structurée: ex: texte .Les Informations non structurée et semi structurée représentent un problème plus complexe et difficile qui constitue Le domaine de la Recherche d'Informations. Actuellement, il y'a un grand développement des outils de gestion des informations non et semi structurés.

RI: Manipulation de l'Information

- Information structurée ----Facile !
- informations non structurées Difficile !
  informations non structurées

1. dans un CV textuel, trouver automatiquement le nombre de postulants ayant un baccalauréat .
2. La recherche de documents : trouver le rapport, écrit en 2001, qui traitait de la nouvelle politique concernant les départs à la retraite .
3. La recherche de textes : trouver la définition de l'expression « abus législatif » dans un ensemble de documents juridiques.

RI: Manipulation de l'Information

4. La recherche d'images : trouver la photographie de mon enfant, qui a été prise le jour de mon dernier anniversaire.

5.La recherche qualitative : trouver la dernière fois que le cours de mes actions a connu une hausse subite.

Qu’est ce que la RI ?

- **Recherche d’information** (RI) est une branche de l’informatique qui s’intéresse à l’**acquisition**, l’**organisation**, le **stockage**, la **recherche** et la **sélection d’information** «**salton1968**»

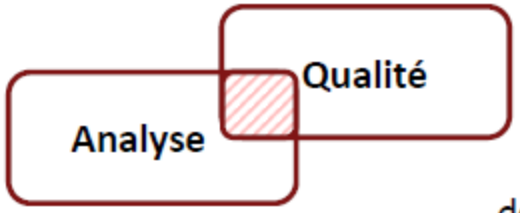
- Ensemble des **méthodes et techniques** pour l’acquisition, l’organisation, le stockage, la recherche et **la sélection d’information pertinente pour un utilisateur**

Collecter des **informations pertinentes**

Les opérateurs booléens

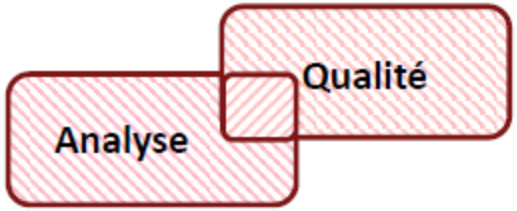
ET

documents correspondant aux deux termes



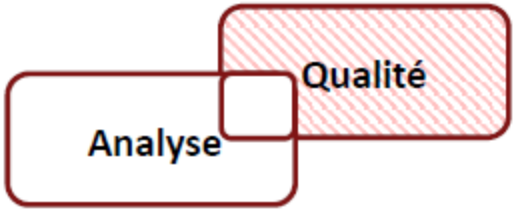
OU

documents avec au moins un des deux termes



SAUF

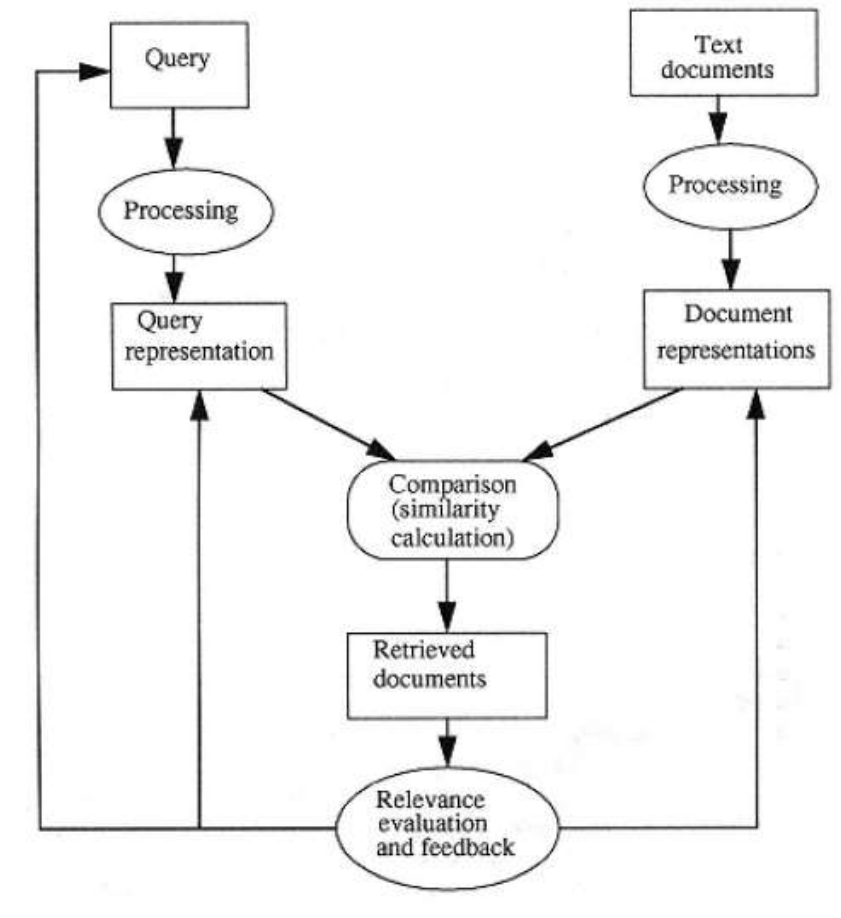
exclusion d'un terme



Analysez vos mots-clés

- Trop de résultats?
  - Soyez plus précis, utilisez un vocabulaire technique
  - Utilisez des opérateurs de recherche
- Pas assez de résultats?
  - Utilisez un vocabulaire plus large, plus général
  - Allégez votre équation

RI


Query
Text documents
Processing
Processing
Query representation
Document representations
Comparison (similarity calculation)
Retrieved documents
Relevance evaluation and feedback

Eléments clés en RI

1. Document
2. Contenu des documents
3. Besoin d'information d'un utilisateur
4. Satisfaction

Eléments clés en RI

1. **Document**

• Texte, Image, vidéo, documents structurés
– Différents types d'information par média
• Texte : livre, article, lettre…
• Image : Images par rayons X, Photographies, Graphiques, …
• Texte : document complet, élément de structure (chapitre , section, paragraphe, phrase), passage (fenêtre de *x mots dans* un texte)
• Vidéo : toute une vidéo, un plan, une image

Eléments clés en RI

2. Contenu des 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 symbolique : information extraite du contenu brut

Eléments clés en RI

3. Besoin d'information d'un utilisateur

Utilisation de requêtes suivant un langage

Requête : une requête correspond à la traduction du besoin d'information de l'utilisateur dans un langage interrogation du **SRI**.
Elle est constituée d'une liste de mots-clés, de terme du langage naturel ou d'un graphique

Eléments clés en RI

4. **Satisfaction**

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

.....Difficile à réaliser

Satisfaction et Pertinence ?

La notion de pertinence peut être appréhendée à deux niveaux :

➢**Niveau utilisateur :** la pertinence correspond à la satisfaction de l'utilisateur par apport à l'ensemble des documents restitués par le SRI. (**pertinence subjective, cognitive)**

➢**Niveau système :** le système mesure un degré de pertinence, une valeur de similitude entre un document et une requête.

Le but d'un SRI est de rapprocher la pertinence système de la pertinence utilisateur.

Bref historique de la RI

- **1940:** Apparition des SRI, focalisation de la RI sur les applications dans des bibliothèques.
- **1950:** Apparition du modèle booléen et l'élaboration de petites expérimentations sur des petites collections de documents.
- **1960 et 1970:** Apparition du système SMART (G. Salton, 1971), développement d'une **méthodologie d'évaluation de système et conception de corpus de test** pour évaluer des systèmes différents.
- **1980:** Développement de l'intelligence artificielle, ainsi on tentait d'intégrer des techniques de l'IA en RI (système expert).
- **1990 et 1995:** L'apparition d'internet, la RI a été modifié et sa problématique plus élargie (traitement des documents multimédia).

RI versus Bases de données

- Base de données ⇒ l'information est structurée par des schémas prédéfinis à l'avance par des relations

  exemple: Auteur(Livre, Nom)

  Information facilement retrouvée par requête:

  select Livre from Auteur where Nom = "DIB"

- RI ⇒ Une partie seulement des spécifications du document est structurée.

  Un **SGBD** peut être utilisé pour rechercher des attributs externes.

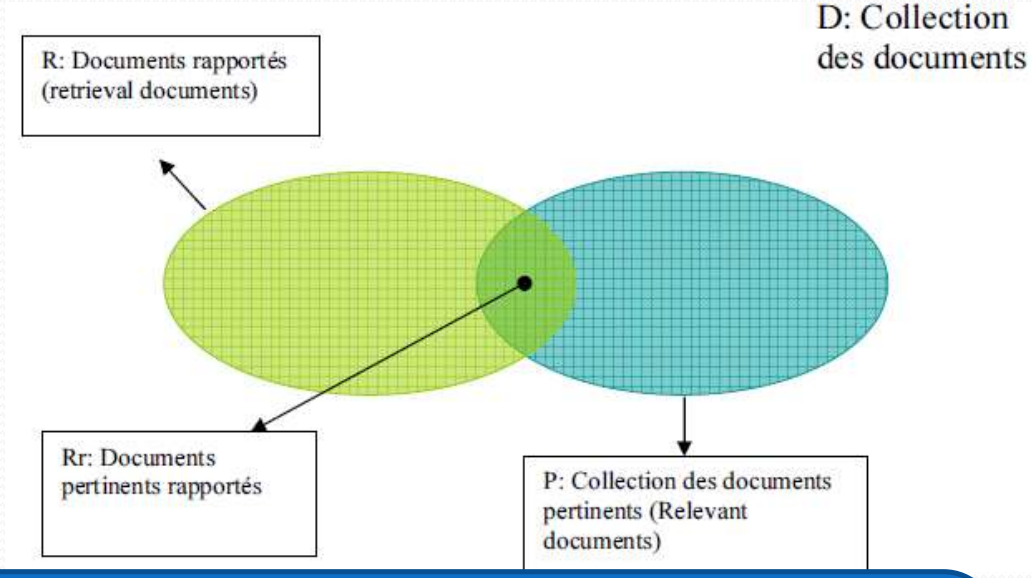
Difficultés dans la RI

- RI est basée essentiellement sur le contenu
- Le contenu est sans structure ….impossible d'extraire une relation

comment mesurer la Pertinence d'un SRI ?

- Précision et Rappel: est un moyen permettant de mesurer la pertinence car la notion de valeur de plausibilité est assez vague.

Précision et Rappel



R: Documents rapportés
P: Collection des documents pertinents
Rr: Documents Pertinents rapportés

Précision = Rr / R **Rappel = Rr / P**

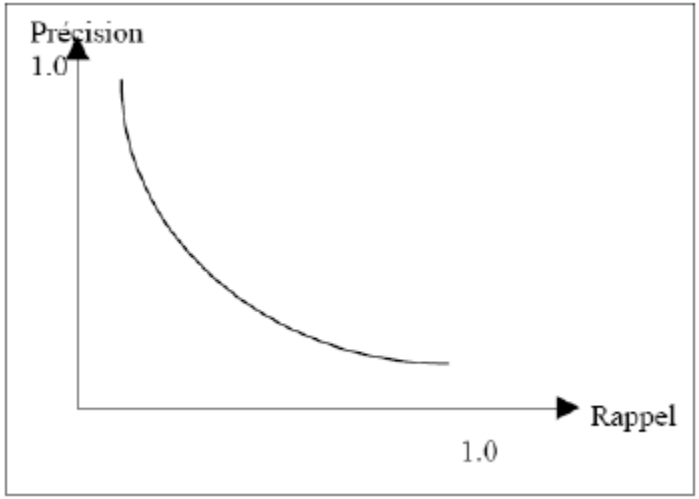
Précision : Un système de RI sera très précis si presque tous les documents renvoyés sont Pertinents,c'est la proportion des documents pertinents parmi l'ensemble de ceux renvoyés par le système.

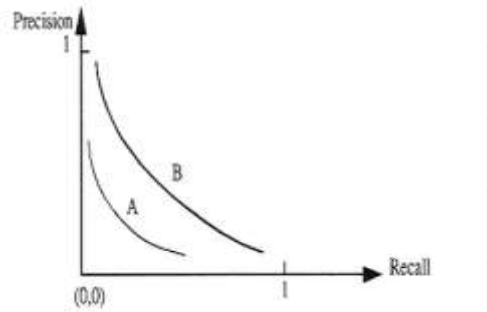
Rappel : Un système de RI aura beaucoup de rappel s'il renvoie la plupart des documents pertinents du corpus pour une question, c'est la proportion de documents pertinents renvoyés par le système parmi tous ceux qui sont pertinents

Présision et Rappel

Il y a une forte relation entre Précision et Rappel: quand Précision ↗ Rappel ↘

Ainsi, pour un système, on a une courbe de précision-rappel qui a en général la forme suivante




Precision
1
B
A
Recall
(0,0)
1

Présision et Rappel

Exemple: Soit une requête qui a 5 documents pertinents dans la base. La liste de réponse du système à cette requête est comme suit

Documents Trouvés	Pertinence
Doc 1	P
Doc 2	X
Doc 3	P
Doc 4	P
Doc 5	X

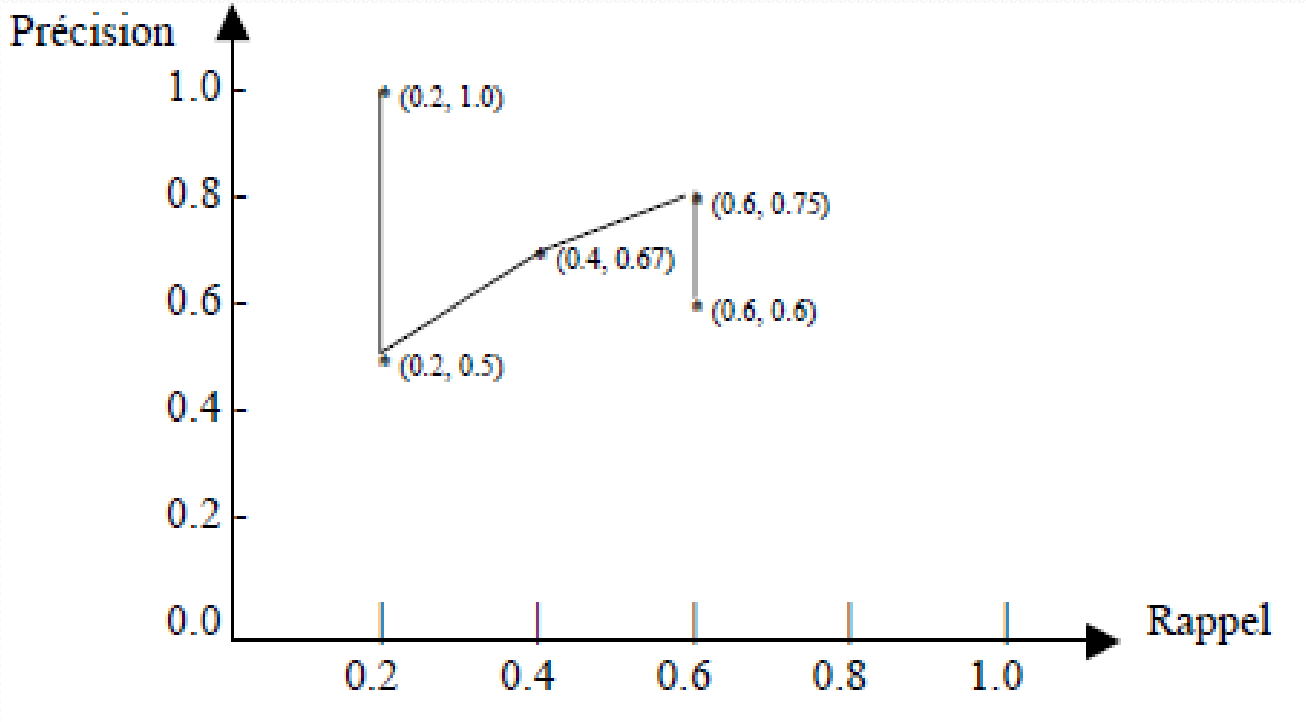
Quelles serait la courbe de Précision et Rappel ?

Précision et Rappel

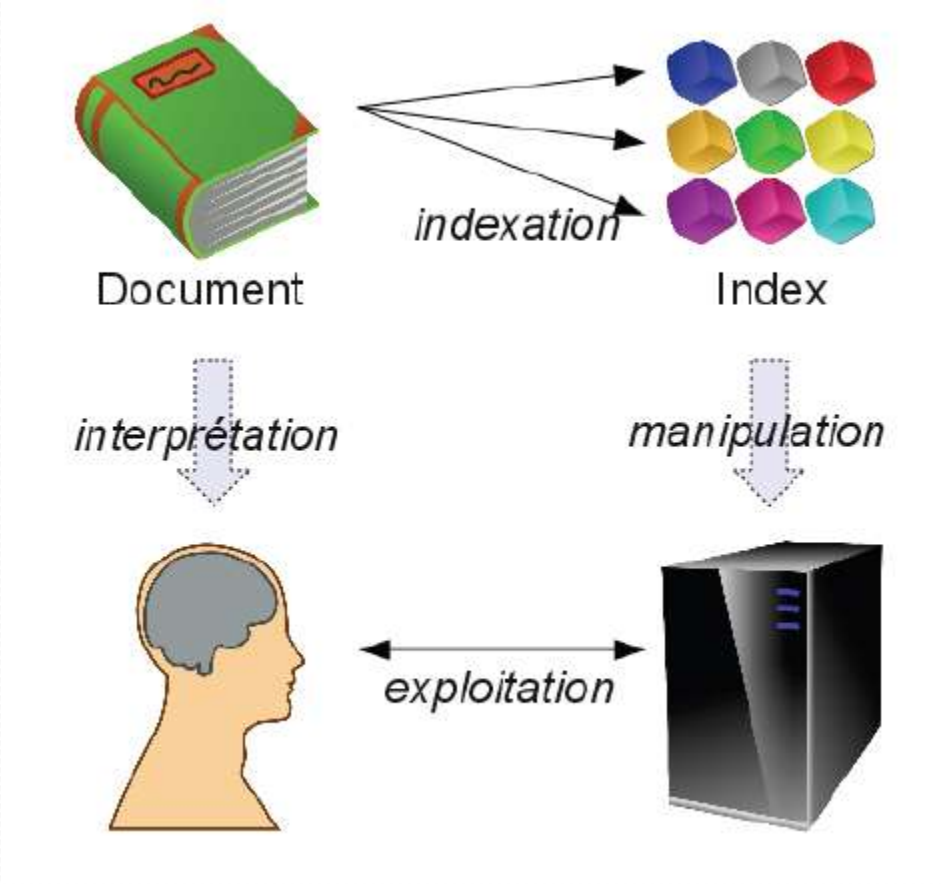
Itération 1:	Rappel=0.2	Précision=1
Itération 2:	Rappel=0.2	Précision=0.5
Itération 3:	Rappel=0.4	Précision=0.67
Itération 4:	Rappel=0.6	Précision=0.75
Itération 5:	Rappel=0.6	Précision=0.6

Précision et Rappel

Courbe

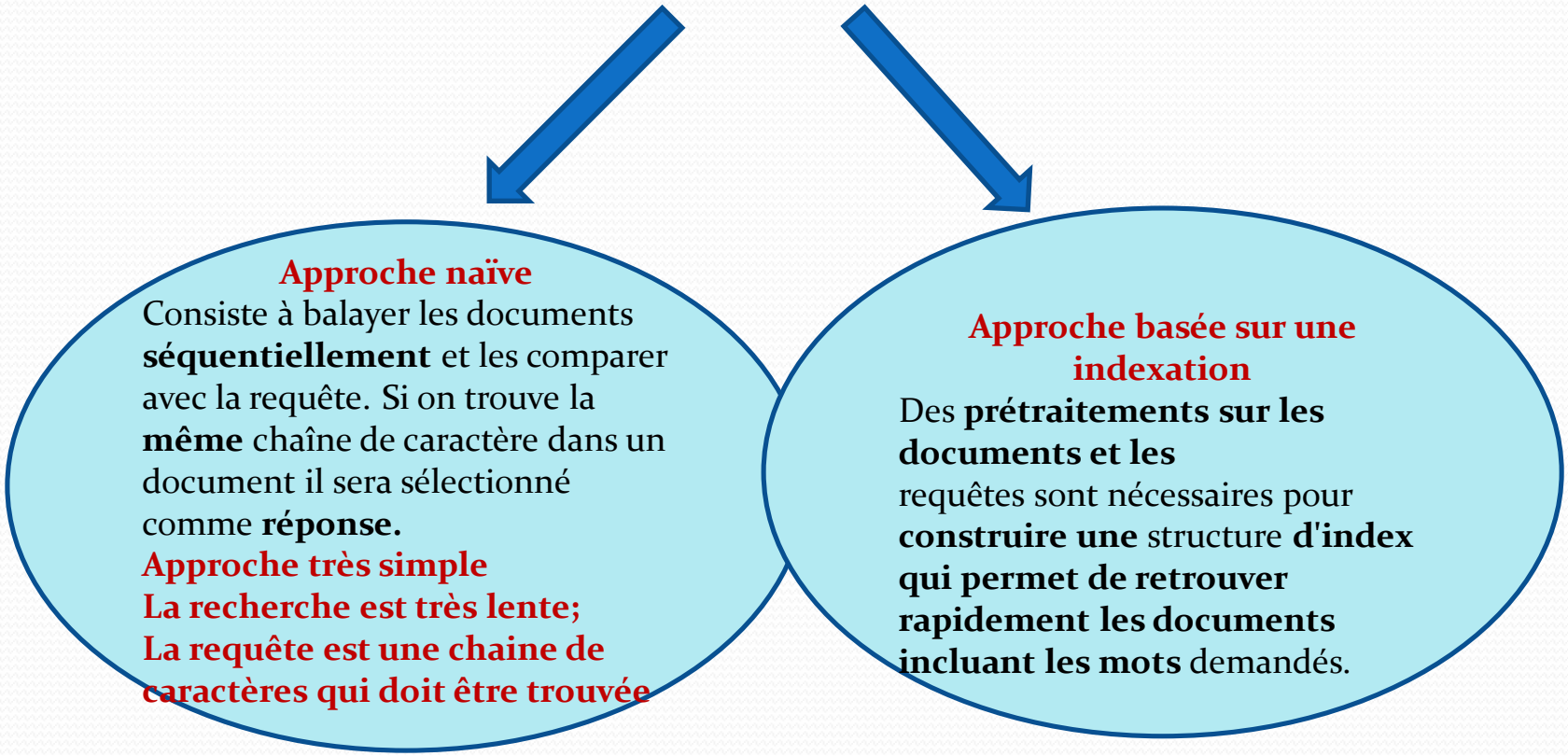


Indexation


indexation
Document
Index
interprétation
manipulation
exploitation

RI: Indexation

2 Approches



+rapide
-La **structure d'index exige un espace de stockage**
important (de 40% à 200% de la taille de collection de documents)

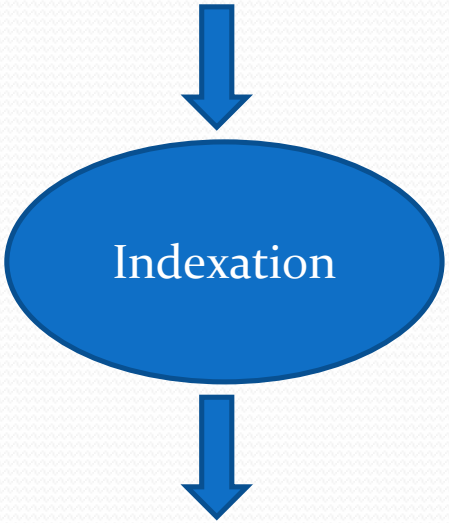
Index ? Définition

- Un index est une structure qui nous donne, pour chaque mot trouvé dans un corpus, la liste des documents où il se trouve et/ou la position des mots dans les documents.
- Un index peut donner la liste des documents où les mots qui apparaissent dans les documents.
- L'indexation peut concerner la représentation des documents ou des requêtes.
- L’indexation a pour rôle d’extraire à partir d’un document ou d’une requête, une représentation paramétrée qui couvre au mieux son contenu sémantique.

L'Indexation

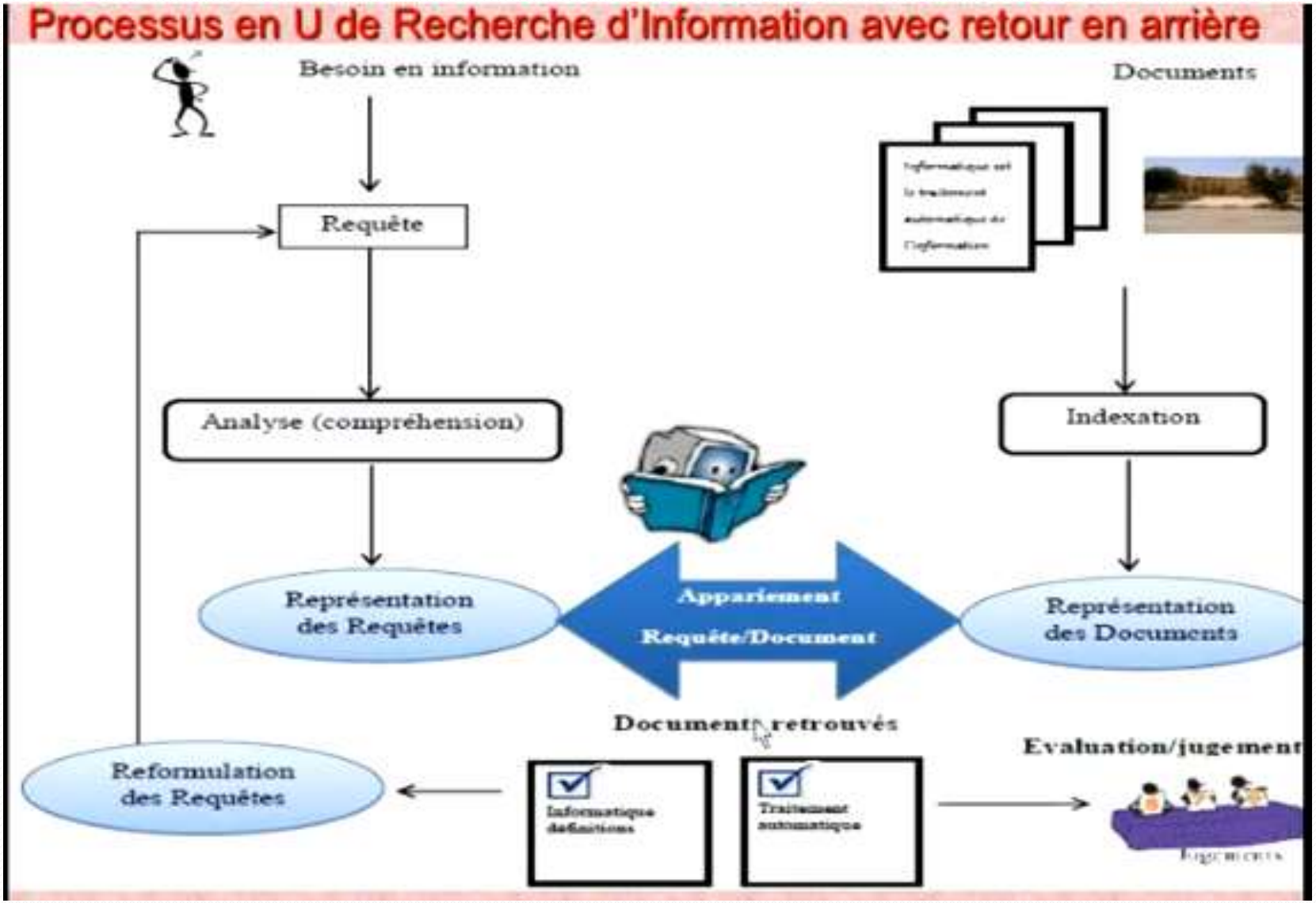
L'indexation est une Représentation Structurée de l'information.

Description Brute (Document ou Requête)



Description Structurée (Documents ou Requête) : des descripteurs (éléments « clés » du document ou de requête).

Indexation


Processus en U de Recherche d'Information avec retour en arrière
Besoin en information
Documents
Requête
Analyse (compréhension)
Indexation
Représentation des Requêtes
Appariement
Requête/Document
Représentation des Documents
Documents retrouvés
Evaluation/jugement
Reformulation des Requêtes
Informatique
Traitement automatique

Indexation

- La qualité de la recherche dépend en grande partie de la qualité de l'indexation

- Si l'information est un texte, le descripteur est : une liste de termes ou groupe de termes significatifs pour l'unité textuelle correspondante.

RI: Indexation

L'indexation doit permettre de sélectionner des représentants du document (mots clés).

Ces représentants permettent de décrire le contenu (la sémantique) du document et de la requête de façon assez précise.

RI: Indexation

L'indexation de documents permet d'utiliser des méthodes pour organiser un ensemble de documents afin de faciliter ultérieurement la recherche .

La diversité des types de documents (textes, images, son, vidéo, Web) nécessite des approches très différentes , en termes de représentation des données.

Ces méthodes ont un point commun : l'**extraction** des caractéristiques, de l'information

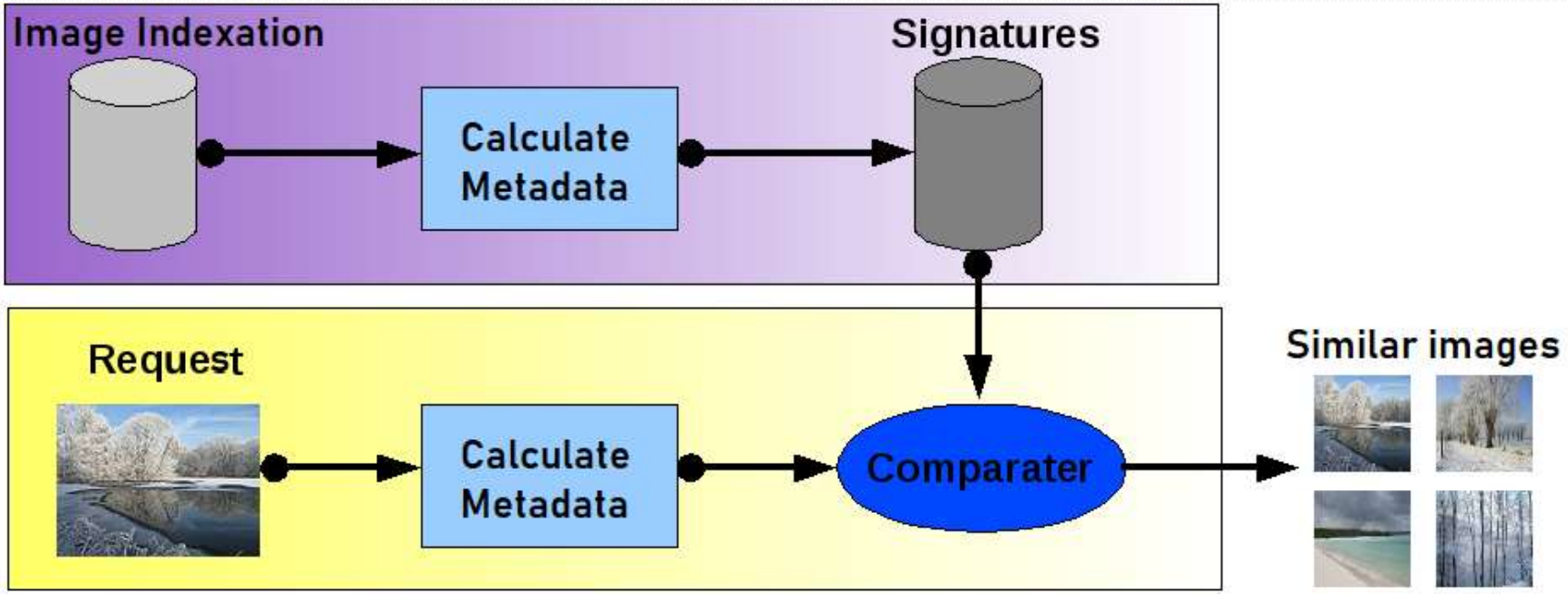
RI: Indexation(images)

Indexation par le contenu

- L ’indexation est réalisée sur le contenu graphique de l'image, c'est-à-dire les formes, les couleurs, les textures et il s'agit d'une indexation d'image par le contenu

- L'indexation d'images par le contenu consiste, après analyse de tous les pixels . Les images sont classiquement décrites comme rendant compte de leur texture, couleur, forme.

Indexation par le contenu: images



RI: Indexation texte

Exemple: Recherche d'un livre

- Un livre peut être spécifié comme suit:

  ISBN: 178-0254828211
- Auteur: Thomas H.Carmen, Charles E.Leiserson
- Titre: Introduction à l'algorithmique
- Editeur: DUNOD
- Date: 2002

OU

- Contenu: <Texte du livre>

RI: Indexation

Qui fait l'indexation ?

Manuelle / Semi-automatique / Automatique

RI: Indexation

3 Types d'indexation

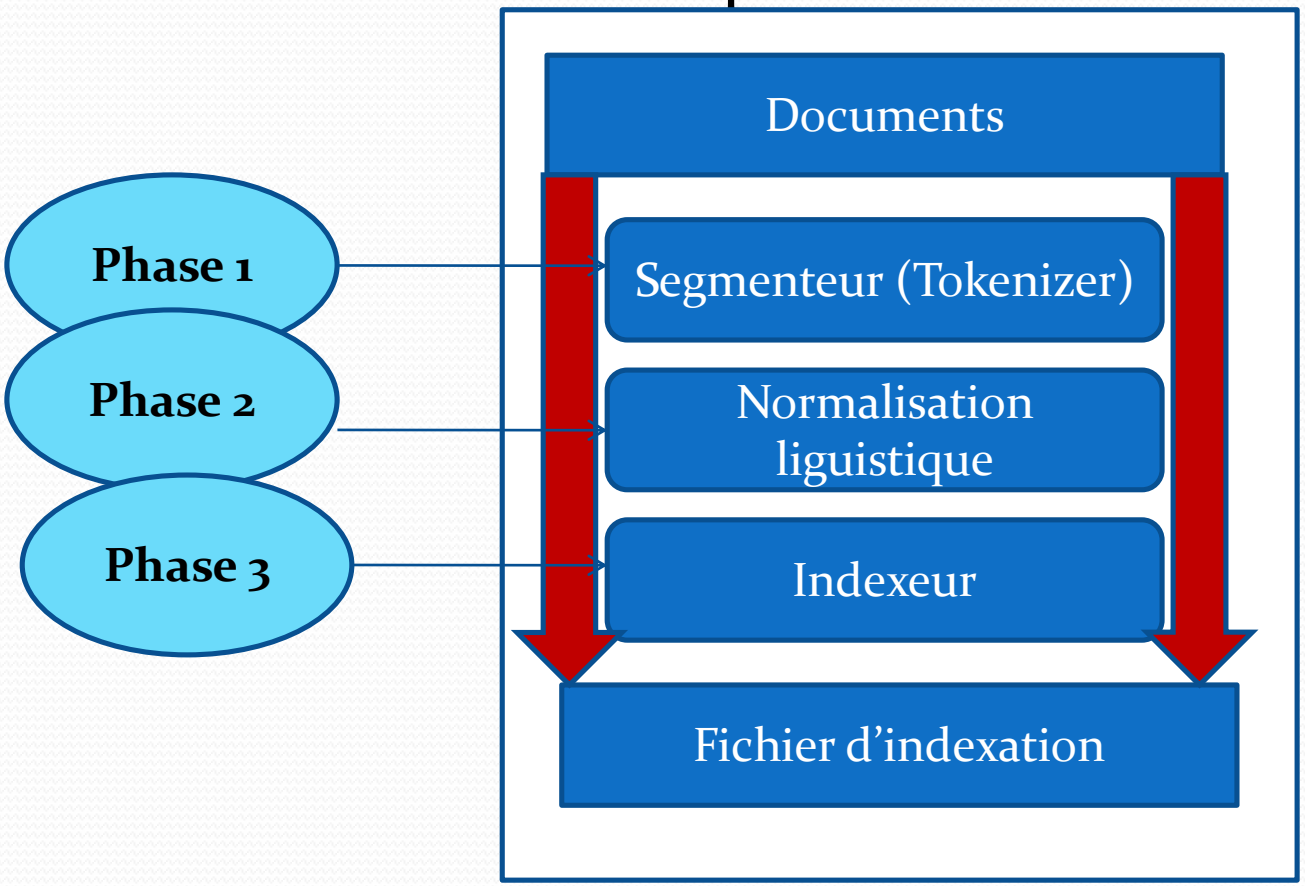
1. Indexation Manuelle: Chaque document est analysé par un documentaliste ou un spécialiste du domaine étudié. Il utilise un vocabulaire contrôlé basé sur le thésaurus, le lexique,…)

   Exige un effort et prend beaucoup de temps

2. Indexation semi-automatique: Un premier processus automatique permet d'extraire les termes du document ,ensuite un spécialiste du domaine choisira les termes significatives de la sélection précédente.
3. Indexation automatique: Processus complètement automatisé à l'aide d'un processus adéquat.

RI: Indexation

Etapes d'indexation



Indexation: Phase 1 (Segmentation)

- Consiste à diviser un texte en unités lexicales (token) élémentaires.
- « localise» les chaînes de caractères entourées de séparateurs (caractères blancs, ponctuations), et les identifie comme étant des mots. Il permet de procéder à une correction des fautes d'orthographe et des erreurs de saisie .

Indexation: Phase 2

- Elimination des mots vides :(les articles, les conjonctions de coordination, les verbes auxiliaires, etc……le ; la, et ,ou, est ,….) cependant certains mots peuvent avoir plusieurs sens ex: **or**.
- Racinisation: Permet de retrouver tous les

  documents dans lesquels apparaissent différentes formes du même mot ». **Exemple** : écologie,écologiste, écologique sont "racinisés" par un seul mot : **écologie**
- Lemmatisation: Consiste à remplacer un terme par son lemme **exemple**: (doit, devrait, devions) ont tous le lemme le verbe **devoir**

Indexation: Phase 2

Les lemmes sont les formes canoniques d'un mot(infinitif pour les verbes, singulier pour les noms, etc.)

- Extraction des mots composés : Reconnaître les mots composés en tant qu'une seule unité. Par exemple :

« canne à pêche » ou « pomme de terre ».

Indexation: Phase 3

Pondération

Pour chaque mot, on doit faire la statistique de sa fréquence d'occurrence dans le document.

A chaque nouvelle occurrence d'un mot, on ajoute 1 dans sa fréquence.

Le terme qui est le plus fréquent aura le poids le plus fort

Indexation: Exemple

✓ D : un système de recherche d ’informations (document) (SRI, base de données documentaires) permet d ’analyser, d ’indexer et de retrouver les documents pertinents répondant à un besoin d ’un utilisateur.

Phase 1 : Extraire les termes et suppression des mots vides

✓ système, recherche, informations, document, SRI, base, données, documentaires, analyser, indexer, retrouver, documents, pertinents, répondant, besoin , utilisateur

Phase 2 : Normalisation

✓ systeme, recherc, informa, documen, sri, base, donn, documen, analyse, indexer, retrouv, documen, pertine, reponda, besoin, utilisa

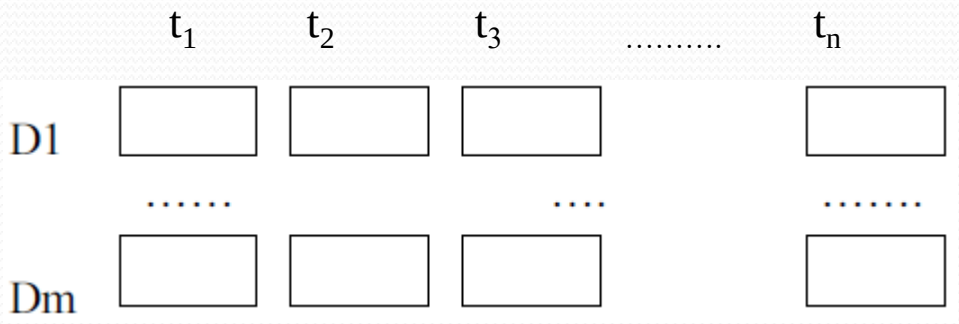
Phase 3 : Indexeur

✓ systeme 1,recherc 1, informa 1, documen 3, sri 1, base 1, donnee 1, analyse1, indexer 1, retrouv 1, pertine 1, reponda 1, besoin 1, utilisa *1*

Indexation: Fichier inverse

- Après analyse de documents d'un corpus, on obtient un tableau :

Document-- > **X** termes

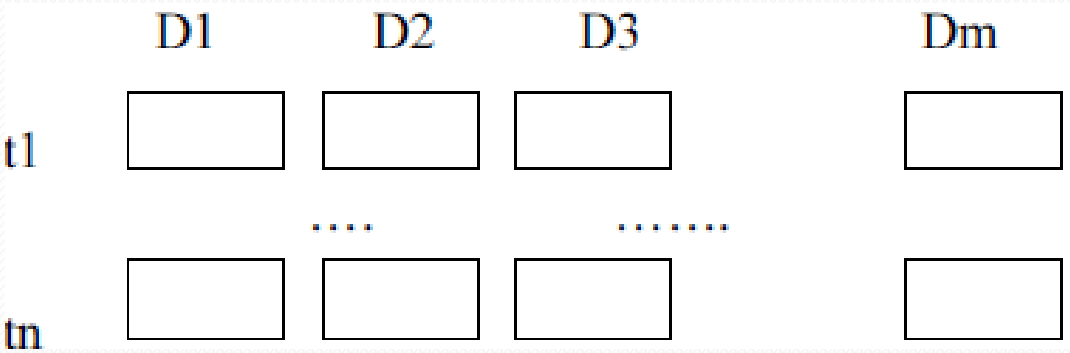


Utilisation en tableau direct « document -> terme »

Ces mots sont ensuite stockés dans une structure appelée **fichier inverse**

Indexation: Fichier inverse

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



Indexation :Phase 3 (indexeur)

- Permet de sélectionner les index et de leur associer une pondération qui assignera aux termes leur degrés d'importance dans les documents.

Il existe 3 approches pour le choix des index :

1. ***Approche basée sur la fréquence d'occurrences (loi de Zipf)***
2. ***Approche basée sur la valeur de discrimination***
3. ***Approche basée sur tf*idf***

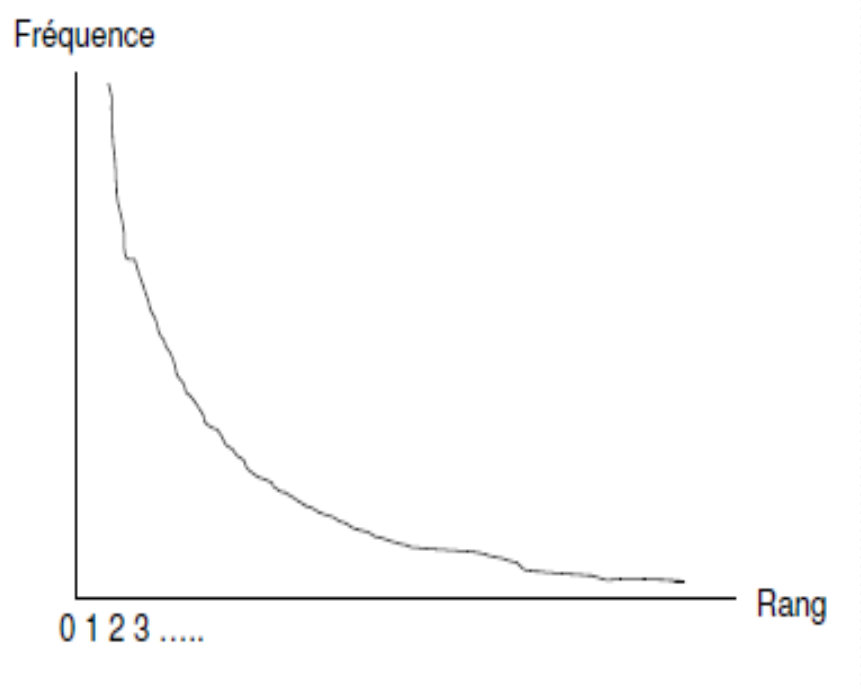
Indexation: Phase 3 (loi de ZIPF)

[Zipf, 1949]

- Les mots sont classés dans l'ordre décroissant de leurs fréquences, puis on leur affecte un numéro de rang (1, 2, ...), alors: rang * fréquence ≅ constante

Rang	Mot	Fréquence	Rang * Fréquence
1	le	69971	69971
2	de	37411	74822
3	et	25852	77556
4	à	20149	80596
5	un	16237	81185
6	dans	13341	80046
7	que	10595	74165

Indexation: Phase 3 (loi de ZIPF)



$P(n) = C\,N / n$

la probabilité d'apparition du nième mot le plus fréquent dans une collection de n'importe quelle langue est inversement proportionnelle à n

N: Nombre des rangs

la loi de Zipf est utilisée pour déterminer les mots qui représentent au mieux le contenu d'un document

Indexation: Phase 3 (loi de ZIPF)

Conjecture de Luhn

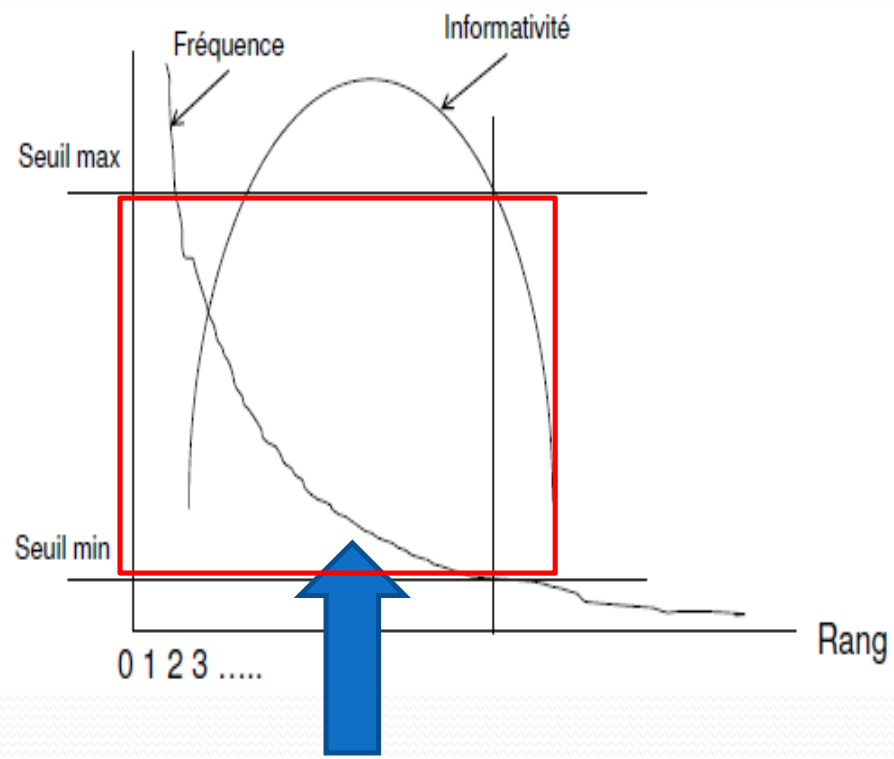
- La conjecture de Luhn est basée sur la loi de Zipf. Elle mesure l' Informativité :

Les termes de rang faible (très fréquents) **ne sont pas pertinents**

Les termes de rang élevés (très rares) **ne sont pas pertinents**

Les descripteurs **pertinents** sont les termes de rang intermédiaire

Indexation: Phase 3 (loi de ZIPF)



il ne faut pas garder tous les mots les plus fréquents.

un **seuil maximal est défini et permet** d'éliminer les mots de très grande fréquence.

Ce seuil maximal est réglé selon **l'informativité de** mots.

L'informativité est la quantité de sens que porte un mot.

Cette notion **n'est pas définie très précisément** dans la RI. Elle est utilisée de façon **intuitive.**

les mots qui ont des fréquences entre les deux seuils sont considérés ayant l'informativité est la plus élevée

Indexation: Phase 3 (Approche basée sur la valeur de discrimination)

- Un terme est dit discriminant s'il distingue bien un document des autres.
- Un terme qui apparaît dans tous les documents n'est pas discriminant.

- Soit le document d_i : < p_{i1} p_{i2} p_{i3} ……..$p_{in,}$> où p_{ij} *est le poids du terme* t_j *dans le document* d_i

- ***n*** *est le nombre de termes du corpus.*

Indexation: Phase 3 (Approche basée sur la valeur de discrimination)

- Le calcul de la valeur de discrimination d'un terme t_k *se* fait comme suit:

1. Calculer le vecteur centroïde **V** du corpus:

Le poids du terme t_j ***dans V est la moyenne de ses poids*** dans les documents:

$$p_j = \frac{\sum_{i=1}^{N} p_{ij}}{N}$$

N est le nombre de documents

Mesure de similarité

Sim(d_i,V)=1- $\sqrt{\frac{\sum_{j=1}^{M}|d_{ij}-V_j|^2}{MaxS}}$

MaxS=M*max $\left(d_{ij:1\leq i\leq M\,,1\leq j\leq N}\right)^2$

M:nombre de termes dans chaque document
N: nombre de documents

Indexation: Phase 3 (Approche basée sur la valeur de discrimination)

2. Calculer l'uniformité du corpus ***U1 (similarité moyenne)***

C est une constante de normalisation = 1/N

sim(di, V) : la similarité entre le document d_i *et V.*

$$U_1 = C \times \sum_{i=1}^{N} sim(d_i, V)$$

Indexation: Phase 3 (Approche basée sur la valeur de discrimination)

3. Uniformiser le poids du terme en question t_k à 0.
4. répéter les deux étapes précédentes pour calculer ***U2***
5. Calculer la valeur de discrimination: v_k =***U2-U1***

Exemple: *Soit le corpus suivant*

d1: 6 2 3 6 2

d2: 6 1 2 0 2 Le vecteur V : 6 2.667 2 2 1.333

d3: 6 5 1 0 0 MaxS=5*36=180

Indexation: Phase 3 (Approche basée sur la valeur de discrimination)

Calcul des similarités:

Sim(d_1, V)=0,685

Sim(d_2, V)=0,800

Sim(d_3, V)=0,739

Calculer l'uniformité du corpus ***U***

U=1/3*(0,685+ 0,800+ 0,739)= 0.741

Indexation: Phase 3 (Approche basée sur la valeur de discrimination)

Normaliser t_k à 0, le vecteur V devient alors

V1: 0 2.667 2 2 1.333

Calcul des similarités:

Sim(d_1, V1)=0,685

Sim(d_2, V1)=0,800

Sim(d_3, V1)=0,739

Calculer l'uniformité du corpus ***U1***

U1=1/3*(0,685+ 0,800+ 0,739)= 0.741

Indexation: Phase 3 (Approche basée sur la valeur de discrimination)

Calcul de la valeur de discrimination: v1 =U_1-U=o

(la suite à faire en TD)