UNIVERSITÉ DES SCIENCES ET TECHNOLOGIE HOUARI BOUMEDIENE

FACULTÉ D'INFORMATIQUE - MASTER 2 SII

RECHERCHE D'INFORMATION
INFORMATION RETRIEVAL

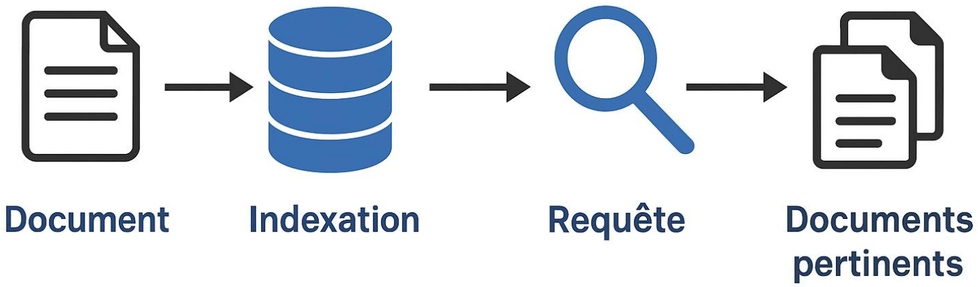
CHAPITRE 2: REPRÉSENTATION DE L'INFORMATION - INDEXATION

29 Septembre 2025

I. INTRODUCTION

La représentation ou l'indexation de l'information est un processus permettant de construire un ensemble d'éléments « clés » permettant de :

- ✓ Bien caractériser le contenu d'un document
- ✓ Réduire la taille d'un document
- ✓ Faciliter le processus de recherche
- ✓ Réduire le temps de recherche.



I.1. ELÉMENTS CLÉS

Les éléments clés d'une information à une autre peuvent être :

Type	Éléments clés	Exemple concret	Remarque
📝 Texte	Mots simples, groupes de mots, phrases	"Programmation", "Programmation avancée"	Utilisé pour les moteurs de recherche, emails, articles scientifiques
🖼️ Image	Couleur, formes, textures, objets détectés	Photo de chat → "chat", "animal"	Utilisé pour la recherche d'images (Google Images)
🎥 Vidéo	Image clé, mouvement, texture dynamique, audio, sous-titres	Clip YouTube → scènes clés + mots des sous-titres	Recherche multimédia, analyse de contenus, recommandations

II. INDEXATION

Basée sur:

- **Vocabulaire contrôlé** : lexique, thésaurus, ontologie ou réseau sémantique.

  Permet la recherche par concepts et la standardisation des termes.
- **Vocabulaire libre – non contrôlé** : éléments extraits directement des

  documents (mots, expressions), utilisé dans l'indexation plein-texte.

II.1. VOCABULAIRE CONTRÔLÉ

- **THÉSAURUS**

Est une liste de mots-clés accompagnée des **relations sémantiques** entre ces mots.
Pour élaborer un thésaurus il faut :

1. Déterminer les termes qui peuvent être pris en compte
2. Associer des relations sémantiques entre ces termes :
   - ✓ Hyperonymie/hyponymie(relation de généralisation ou de spécialisation) (is-a),
   - ✓ antonymie (relation d'opposition)
   - ✓ … etc.

- **ONTOLOGIE** : va un peu plus loin. Elle consiste en une **liste de concepts** – souvent un regroupement de termes partageant une sémantique commune – et définit les **relations entre ces concepts**. Elle permet ainsi de représenter plus finement la signification et les liens entre les notions d'un domaine.

II.2. AVANTAGES ET INCONVÉNIENTS DU VOCABULAIRE CONTRÔLÉ

- **AVANTAGES**

✓ Permet la recherche par concepts (par sujets, par thèmes), plus précis que la recherche par mots simples
✓ Facilite la classification des documents en regroupant ceux qui traitent du même sujet
✓ Fournit une terminologie standard pour indexer et rechercher les documents

- **INCONVÉNIENTS**

✓ Indexation très coûteuse (pour construire le vocabulaire, pour affecter les concepts (termes) aux documents.
✓ Difficile à maintenir (la terminologie évolue, plusieurs termes sont rajoutés tous les jours)
✓ Les utilisateurs ne connaissent pas forcément le vocabulaire utilisé

II.3. AVANTAGES ET INCONVÉNIENTS DU VOCABULAIRE NON CONTRÔLÉ

- **AVANTAGES**

✓ Indexation plus rapide a réalisé
✓ Facile à maintenir sa mise à jour
✓ Les utilisateurs connaissent facilement le vocabulaire utilisé

- **INCONVÉNIENTS**

✓ Indexation basée sur des statistiques, ne prend pas en compte le sens des mots
✓ Empêche la recherche des documents par concept, par sujet, par thème

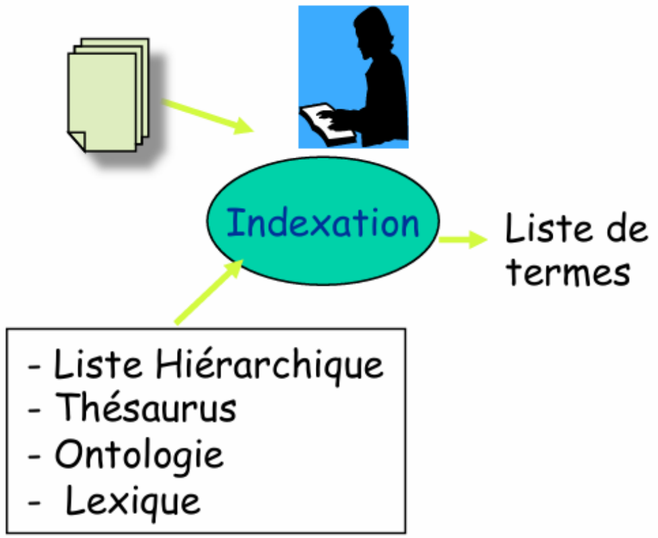
III. PROCESSUS D'INDEXATION

L'indexation peut être :

- **Manuelle** : réalisée par des experts en indexation
- **Automatique** : effectuée des algorithmes efficaces sur un ordinateur
- **Semi-automatique** : combinaison de l'expertise humaine et des outils automatiques.

III.1. Indexation manuelle

- ✓ Choix des mots effectué par les indexeurs (experts)
- ✓ Basée sur un vocabulaire contrôlé
- ✓ Approche utilisée souvent dans les bibliothèques, les centres de documentation
- ✓ Dépend du savoir-faire de l'indexeur (expert)



III. PROCESSUS D'INDEXATION

III.2. Indexation automatique

- ✓ Approche statistique par distribution des mots
- ✓ Approche statistique compréhension du texte (TALN)
- ✓ Approche courante, plutôt statistique avec des hypothèses simples
  - Redondance (fréquence) d'un mot marque son importance
  - Cooccurrence des mots marque des concepts sur le sujet d'un document

III.3. Indexation Manuelle VS. Automatique

L'indexation manuelle nécessite des experts des domaines des sujets de documents, elle est couteuse en temps et en ressource humaine. L'indexation automatique est plus rapide en temps, mais nécessite des approches efficaces. Dans le cadre de ce cours, on s'intéresse à l'**INDEXATION AUTOMATIQUE**

IV. PROCESSUS D'INDEXATION AUTOMATIQUE

Pour indexer les documents de façon automatique, nous utilisons l'approche courante, qui est une approche basée sur des statistiques sur l'apparition des mots dans les documents.

Généralement un mot est une suite de caractères séparés par blanc, signe de ponctuation, caractères spéciaux,…). Donc, il faut d'abord définir la façon de localiser un mot. Puis, plusieurs questions se poses :

1. Est ce qu'on garde dans l'index tous les mots du document, ou bien seulement les mots représentatifs du document. ?
2. Les mots extraits seront utilisés comme ils sont, ou bien ils doivent subir certaines transformations ?
3. Les mots ont la même importance ou bien chaque mot a une certaine importance qu'il faut définir ?

DÉMARCHE COURANTE DE L'INDEXATION

Afin de faire une indexation qui permet une représentation réduite, représentative et facile à utiliser dans la phase de la recherche, il faut la faire en 3 étapes :

- **Étape 1** : extraction de mots représentatifs
- **Étape 2** : normalisation des mots extraits
- **Étape 3**: pondération des mots normalisés, pour mesurer leurs importances

IV.1. ETAPE 1 : EXTRACTION DES MOTS À UTILISER

- Cette étape s'appelle aussi « tockenization/Segmentation ». Généralement un mot-terme est une suite de caractères séparés par blanc, signe de ponctuation, caractères spéciaux,…).
- Ce sont les index utilisés lors de la recherche
- Dépend de la langue
  - Langue française
    - L'ensemble → un terme ou deux termes ?
      - L ? L' ? Le ?
  - Langue Allemande les mots composés ne sont pas segmentés
    - *Lebensversicherungsgesellschaftsangestellter*, qui signifie *employé d'une compagnie d'assurances vie*.

IV.1. ETAPE 1 : EXTRACTION DES MOTS À UTILISER

– Langue arabe

- L'arabe s'écrit de droite à gauche, sauf certains éléments comme les chiffres qui restent de gauche à droite.
- Les mots sont bien séparés, mais les lettres sont liées entre elles dans un mot.
- Problème d'orthographe des noms propres : un même nom peut s'écrire de plusieurs façons.

  Exemple : Einstein → انشتاين, نيتشناي, اينشتاين

IV.1. ETAPE 1 : EXTRACTION DES MOTS À UTILISER

Donc :

1. Définir la façon de localiser un mot.
2. Extraire les mots
3. Définir les mots non utiles (liste de mots vides / stoplist / common words).

Ex. • Anglais : the, or, a, you, I, us, …

• Français : a, le , la, de , des, je, tu, …

Attention à :

- US : «USA » ; « give us information »
- a de (vitamine a) • ….. Etc.
- Donc, il faut bien définir cette liste de mots vides !

4. Supprimer (ignorer) les mots vides, et garder que les mots représentatifs.

IV.2. ETAPE 2 : NORMALISATION

sert à ramener les mots à une forme de base pour éviter la dispersion entre différentes variantes.

Il existe plusieurs façons pour la normalisation :

1. **Normalisation par Lemmatisation : (radicalisation) / (stemming)**
   C'est un processus morphologique permettant à réduire les mots à leur radical.

Ex.

- Français : économie, économiquement, économiste → econom
- Anglais : retrieve, retrieving, retrieval, retrieved, retrieves → retriev

IV.2. ETAPE 2 : NORMALISATION

2. **Normalisation par l'utilisation de règles de transformations**

- Règles de type : condition action
  - L'algorithme le plus connu est : Porter (utilisé pour l'anglais)
  - Plusieurs implémentations sont accessibles

    http://www.tartarus.org/~martin/PorterStemmer/

- Analyse grammaticale
  - Utilisation de lexique (dictionnaire)
  - Un outil gratuit et connu est **TreeTagger :** qui identifie la catégorie grammaticale du mot (verbe, nom, adjectif, etc.) et en donne sa forme de base.

- Troncature : sert à couper directement une partie du mot pour garder seulement la racine

IV.2. Normalisation : Algorithme Porter Stemmer

- Est basée sur la mesure m de séquences voyelles-consonnes pour décider si une règle s'applique.

  Ex.
  - *tree* → pas d'alternance → **m = 0** **seule séquence de voyelles**
  - *tr**oub**le* → une alternance → **m = 1** **alternance V/C**
  - *tr**oubl**es* → deux alternances → **m = 2** **deux alternances V/C**

- L'algorithme applique des règles étape par étape pour supprimer ou changer les **suffixes** :

Étape	But	Règles principales	Exemples
STEP 1A	Supprimer les pluriels	- sses → ss - ies → i - ss → ss - s → ""	caresses → caress , ponies → poni cats → cat
STEP 1B	Supprimer participes passés	- (m>0) eed → ee - (*v*) ed → "" - (*v*) ing → "" (règles spéciales après suppression : ajout de e, suppression double consonne, etc.)	agreed → agree plastered → plaster motoring → motor
STEP 1C	Remplacer y final	- (*v*) y → i	happy → happi sky → sky
STEP 2	Réduction de suffixes longs (si m>0)	- ational → ate - tional → tion - enci → ence - anci → ance	relational → relate conditional → condition valenci → valence
STEP 3	Transformation de suffixes (si m>0)	- icate → ic - ative → "" - alize → al - iciti → ic	triplicate → triplic formalize → formal electriciti → electric
STEP 4	Suppression de suffixes (si m>1)	- al → "" - ance → "" - ence → "" - er → "" - ion → "" (si précédé de s ou t)	revival → reviv allowance → allow inference → infer airliner → airlin
STEP 5	Nettoyage final	- (m>1) e → "" - (m=1 and not *o) e → "" - (m>1 and *d and *L) → supprimer L doublé	probate → probat cease → ceas , controll → control, toe→toe

Exemple de normalisation par Porter Stemmer

- **Texte original :**

marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales

Étape 1 : Tokenization (segmentation en mots)

On découpe le texte en mots séparés :

marketing, strategies, carried, out, by, U.S., companies, for, their, agricultural, chemicals, report, predictions, for, market, share, of, such, chemicals, or, report, market, statistics, for, agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted, sales, market, share, stimulate, demand, price, cut, volume, of, sales

Étape 2 : Suppression des mots vides (stopwords)

On enlève les mots très fréquents qui n'apportent pas de sens : (ex. *out, by, for, their, of, such, or, the…*)

Exemple de normalisation par Porter Stemmer

Il reste :
marketing, strategies, U.S., companies, agricultural, chemicals, report, predictions, market, share, chemicals, report, market, statistics, agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted, sales, market, share, stimulate, demand, price, cut, volume, sales

Etape 3 : Stemming avec Porter

On réduit chaque mot à sa racine :

- *marketing → market*
- *strategies → strategi*
- *carried → carri*
- *companies → compani*
- *agricultural → agricultur*
- *chemicals → chemic*
- *predictions → prediction*
- *statistics → statistic*
- *agrochemicals → agrochem*
- *pesticide → pesticid*
- *herbicide → herbicid*
- *fungicide → fungicid*
- *insecticide → insecticid*
- *fertilizer → fertiliz*
- *sales → sale*
- *stimulate → stimulat*

Étape 4 : Comptage des fréquences

Market 4, strategi 1, carri 1, U.S. 1, compani 1, agricultur 1, chemic 2, report 2, prediction 2, share 1, statistic 1, agrochem 1, pesticid 1, herbicid 1, fungicid 1, insecticid 1, fertiliz 1, sale 2, stimulat 1, demand 1, price 1, cut 1, volum 1

IV.2. ETAPE 2 : NORMALISATION

3. Normalisation par troncature

Consiste à Tronquer les mots à X caractères

- Tronquer plutôt les suffixes
- Ex. Troncature à 7 caractères • économiquement : économi

Quelle est la valeur optimale de X ? : **7 caractères pour le Français**

IV.3. ETAPE 3 : PONDÉRATION DES MOTS

Pour mesurer l'importance d'un mot dans un document, il faut lui attribuer une valeur.

La 1 ère valeur qu'on peut exploiter pour le moment c'est la fréquence du mot dans le document.

Exemple par troncature :

- **Texte :**

un système de recherche d 'informations (document) (SRI, base de données documentaires, recherche documentaire) permet d 'analyser, d 'indexer et de retrouver les documents pertinents répondant à un besoin d'un utilisateur en information.

- **Extraction des mots et suppression des mots vides :**

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

- **Normalisation par troncature à 7 caractères et mettre tout en miniscule :**

système, recherc, informa, documen, sri, base, données, documen, recherc, documen, analyse, indexer, retrouv, documen, pertine, reponda, besoin, utilisa, informa

- **Pondération des termes par fréquences :**

système 1, recherc 2, informa 2, documen 4, sri 1, base 1, données 1, analyse 1, indexer 1, retrouv 1, pertine 1, reponda 1, besoin 1, utilisa 1

IV.3. INCONVÉNIENTS DE LA NORMALISATION

- ✓ Les algorithmes de "Stemers" sont souvent difficiles à comprendre et à modifier
- ✓ Peut conduire à une normalisation "agressive du mot" - perdre le sens du mot
  - Exemple de normalisation par Porter : « general » devient « gener »
  - Exemple de normalisation par troncature : « Internet » devient « Interne »

Note : Il existe des techniques (analyse de corpus) pour réduire ces effets négatifs.

V. LA MÉTHODE DES N-GRAMMES

- Définition : un n-gram est une succession de n lettres, Généralement n = 1,2,3
- Utilisée pour le chinois car ses termes ne sont pas séparés par des espaces
- Le SRI doit retrouver des documents pertinents sans avoir réalisé une tokenisation
- Intéressant pour la radicalisation
- Exemple : retrieval
    - 1-gram : r, e, t, r, i, e, v, a, l
    - 2-gram : re, et, tr, ri, ie, ev, va, al
    - 3- gram : ret, etr, tri, rie, iev, eva, val

Comparer deux mots par n-grammes

Exemple : retrieve et retrieval par 3-gram

- A=retrieve : ret, etr, tri, rie, iev, eve
- B=retrieval : ret, etr, tri, rie, iev, eva, val

Sim(A,B)=2x5/6+7≈0.77

$$Sim(A,B) = \frac{2 \times nb\_comm}{nb_A + nb_B}$$

- **nb_A** = number of n-grams in word A
- **nb_B** = number of n-grams in word B
- **nb_comm** = number of n-grams that A and B have **in common**

VI. RÉSUMÉ DU PROCESSUS D'INDEXATION

Le processus d'indexation se déroule généralement en **trois grandes étapes** :

1. **Extraction des mots** : identification des termes significatifs dans le document, suppression des mots vides (stop words) et mise en minuscule pour uniformiser. *(étape obligatoire)*
2. **Normalisation** : réduction des mots à une forme canonique (lemmatisation, stemming, règles de transformation, troncature, n-grammes, etc.). *(étape facultative)*
3. **Pondération des mots** : attribution d'un poids ou d'une importance à chaque terme, souvent basée sur sa fréquence dans le document. *(étape obligatoire)*

VII. FICHIER INVERSE

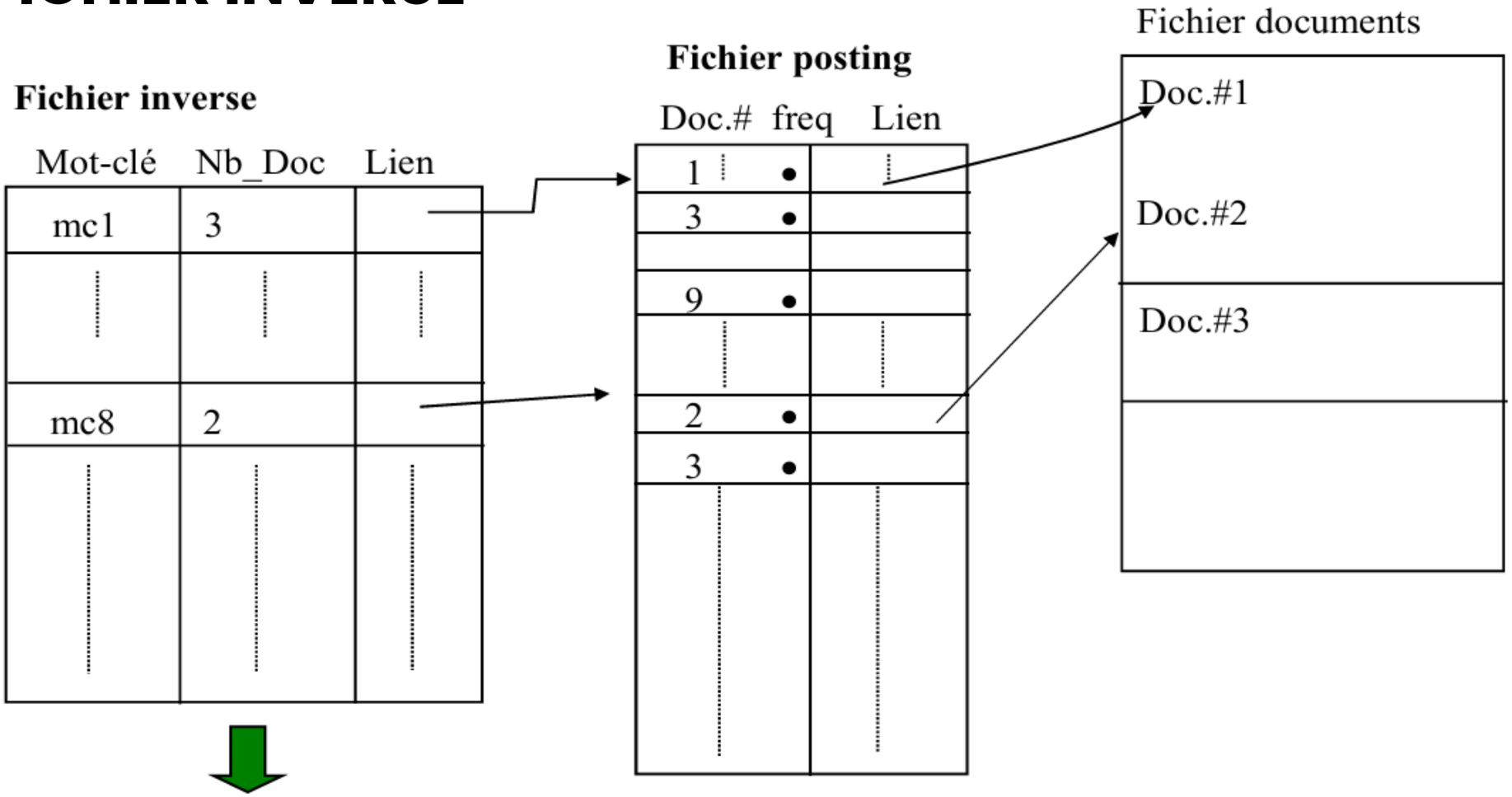
Une fois les documents sont indexés, chaque document est représenté par un **descripteur**.

Un descripteur se compose de :

- ✓ Liste de mots (ou termes) extraits du document
- ✓ Fréquence d'apparition de chaque mot
- ✓ Exemple : système 1, recherc 1, informa 1, documen 3, sri 1, base 1, donnée 1, analyse1, indexer 1, retrouv 1, pertine 1, reponda 2, besoin 3, utilisa 1

Ces mots sont ensuite stockés dans une structure appelée fichier inverse, pour pouvoir l'utiliser dans la phase de recherche (appariement matching document/requête)

VII. FICHIER INVERSE



- Liste triée
- B-Arbre
- Table de hashage (hash-code)
- ...

EXEMPLE

Pour une collection de 3 documents, on a les descripteurs suivants :

- D1 (sri 1 , recherche 1, information 2, document 3)
- D2 (document 1, information 4, vidéo 2)
- D3 (document 2, automatique 1)

La représentation théorique du fichier inverse de cette collection est :

Terme	D1	D2	D3
sri	1	0	0
recherche	1	0	0
information	2	4	0
document	3	1	2
vidéo	0	2	0
automatique	0	0	1

DÉMARCHE DE CONSTRUCTION D'UN FICHIER INVERSE

La construction d'un fichier inverse est une étape très importante en RI, elle peut prendre énormément de temps, mais ce temps n'a pas d'influence sur le processus de RI, car elle se fait avant la phase de recherche. Pour construire un fichier inverse il faut :

1. Faire toutes les étapes de l'indexation
2. Choisir la bonne structure à utiliser, qui permet le stockage de chaque mot, en le reliant avec son poids (fréquence pour l'instant) et son document.
3. Stocker chaque mot, avec son poids et son document

RÉCAPITULATIF DE L'INDEXATION DE L'INFORMATION EN RI

- ✓ A l'issue de cette opération, chaque document sera représenté par une liste de termes pondérés. (sera détaillé dans le chapitre suivant)
- ✓ Le poids est fondamental et a une grande influence dans toutes les approches (modèles) de RI
- ✓ L'ensemble des termes extraits de tous les documents est stocké dans une structure spécifique appelée : fichier inverse
- ✓ Ce fichier permet de retrouver pour un terme donné tous les documents qui contiennent ce terme, avec son poids d'apparition.