Institut Supérieur d'Informatique et des Mathématiques de Monastir

TECHNIQUES D'INDEXATION ET RECHERCHE MULTIMÉDIA

IMEN CHEBBI

Chapitre 2

1.Indexation

Recherche d’information (RI) :

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

Un Système de Recherche d’Information (SRI)

Un système de recherche d'information (RI) est un système qui permet de retrouver les documents pertinents à une requête d'utilisateur, à partir d'une base de documents volumineuse.

Trois notions clés: **documents**, **requête**, **pertinence**.

Requête : exprime le besoin d'information d'un utilisateur

Document : toute unité qui peut constituer une réponse à une requête,
Un document peut être un texte, un morceau de texte, une page Web, une image, une bande vidéo, etc,

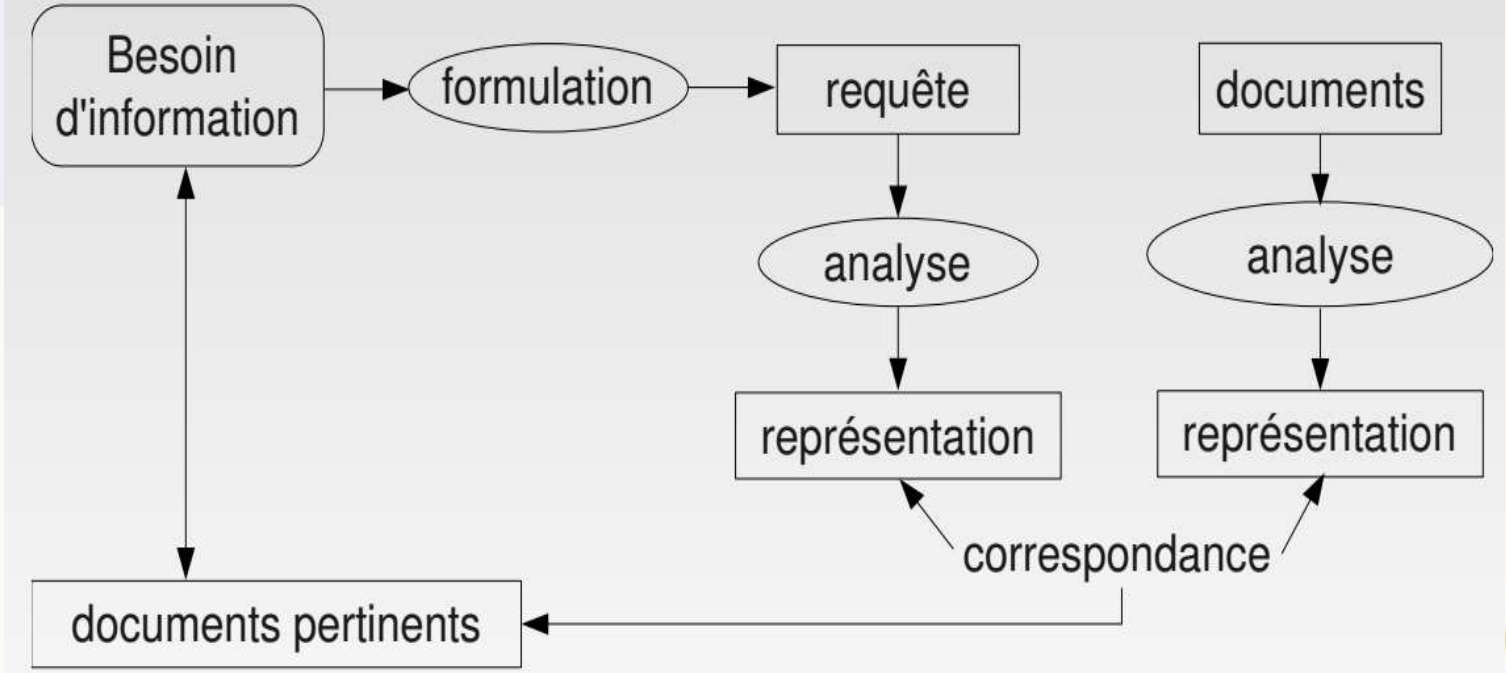
Base de documents : ensemble des documents disponibles

Pertinence : De façon générale, dans document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin. Sur cette notion le système doit juger si un document doit être donné à l'utilisateur comme réponse ou non

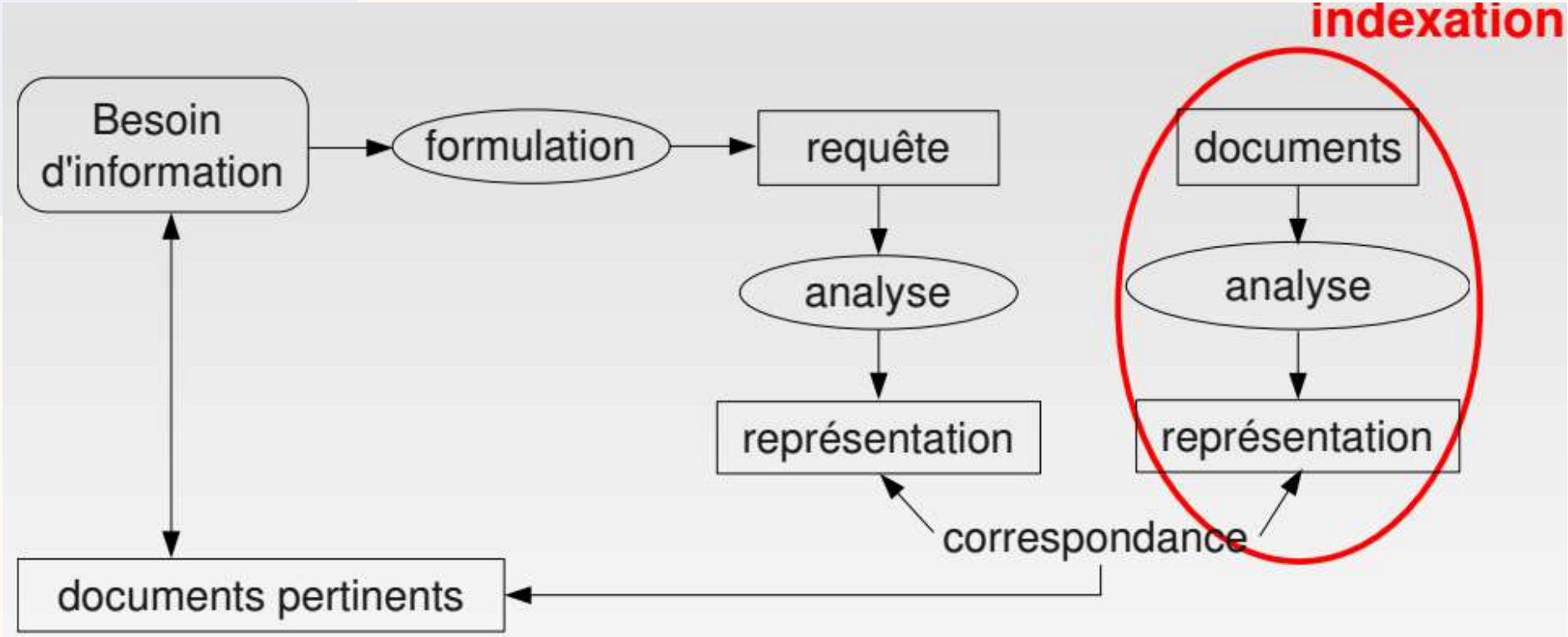
EXEMPLES D'APPLICATIONS

- Outils de recherche dans les mails, dans les fichiers, ...
- Systèmes de RI documentaires,
- Systèmes de RI pour les bases de documents d'une entreprise,
- Systèmes de RI sur le Web tels que google, bing ,,,etc.

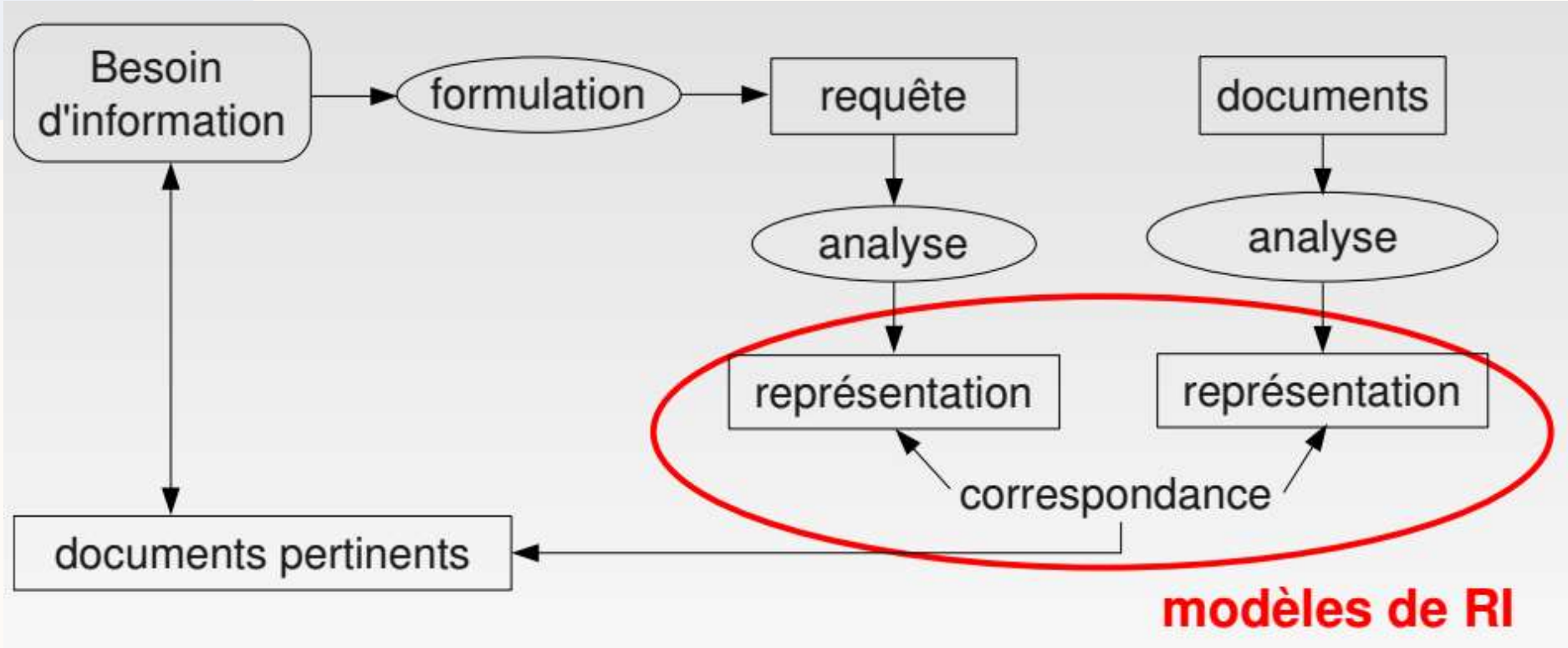
APPROCHE CLASSIQUE DE LA RI



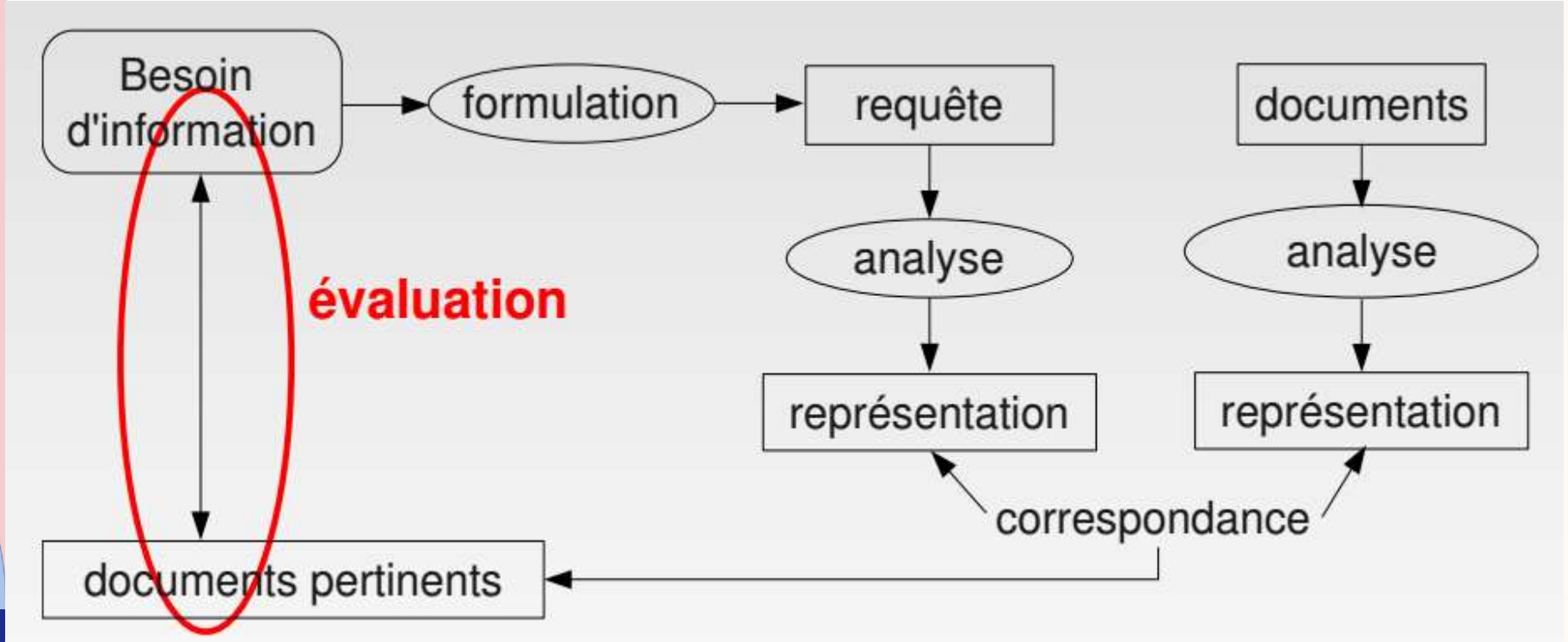
APPROCHE CLASSIQUE DE LA RI



APPROCHE CLASSIQUE DE LA RI



APPROCHE CLASSIQUE DE LA RI



INDEXATION

base
de
données

INDEXATION - POURQUOI UTILISER LES INDEX ?

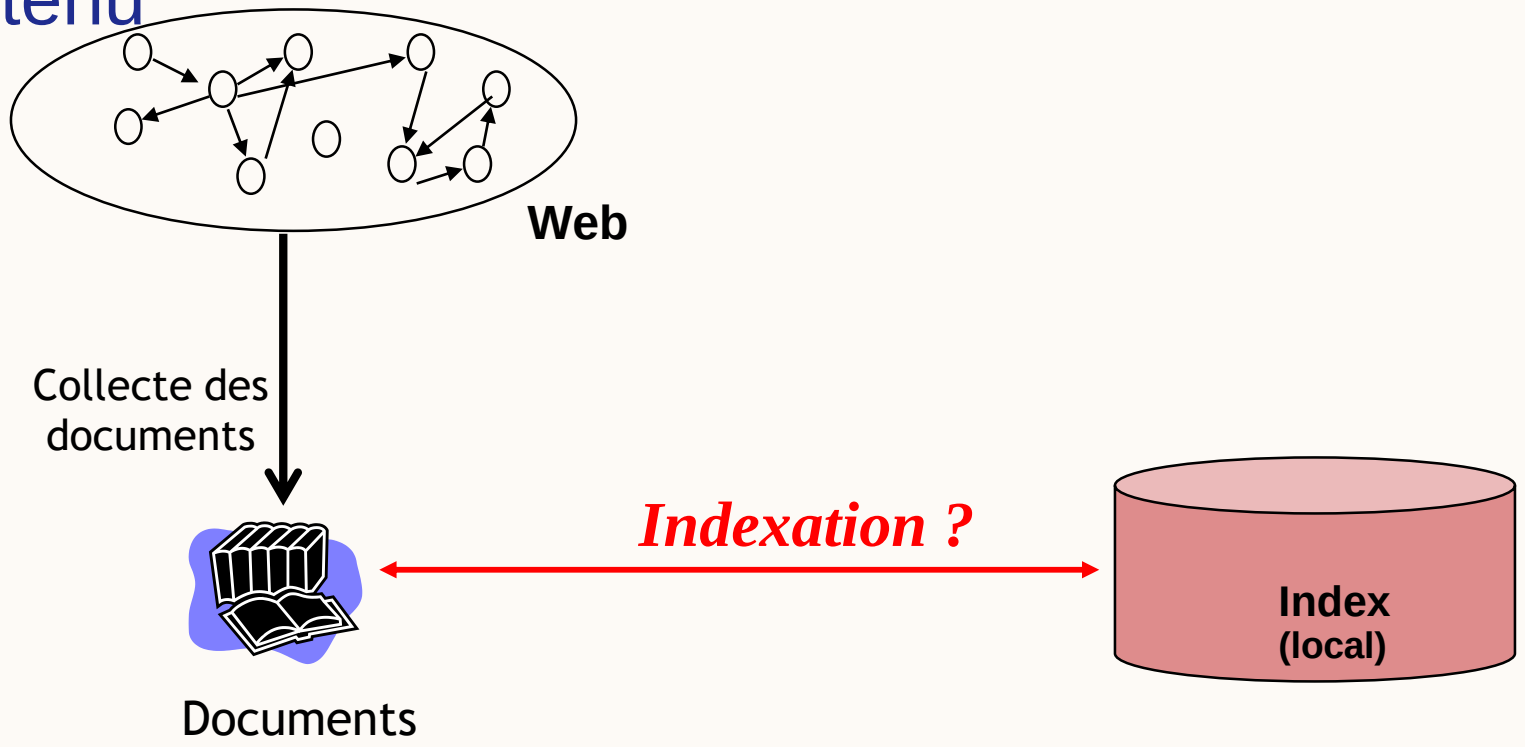
- Imaginez un moteur de recherche qui ne dispose pas d'une base d'index

- Pour chaque requête, il doit
  - **accéder au Web (faire un tour complet)**
  - **analyser les documents un par un**
  - **juger l'importance de chaque document par rapport à la requête en question**
  - **« fabriquer" la réponse en fonction des pertinences des documents**
  - **afficher le résultat**

=> une base d'index est indispensable

INDEXATION

- Analyse du document et interprétation de son contenu



INDEXATION

- Un index contient une "interprétation" du document au lieu du document entier

- Il contient
  - les termes représentatifs d'un document
  - les poids (l'importance) des termes dans chaque document

- Chaque moteur possède un index inverse
  - transformation de

    "quels mots apparaissent dans la page ?" en **"dans quelles pages (URL) apparaît le mot X?"**

INDEXATION

- Peut être
  - Manuelle (expert en indexation)
  - Automatique (ordinateur)
  - Semi-automatique (combinaison des deux)
- Basée sur
  - Un langage contrôlé (lexique/thesaurus/ontologie/réseau sémantique)
  - Un langage libre (éléments pris directement des documents)

INDEXATION MANUELLE

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

INDEXATION MANUELLE: AVANTAGE DU VOCABULAIRE CONTRÔLE

- Permet la recherche par concepts (par sujets, par thèmes), plus intéressante que la recherche par mots simples.
- Permet la classification (regroupement) de documents (par sujets, par thème).
- Fournit une terminologie standard pour indexer et rechercher les documents

INDEXATION MANUELLE: INCOVÉNIENT DU VOCABULAIRE CONTRÔLE

Indexation très coûteuse

– Pour construire le vocabulaire
– Pour affecter les concepts (termes) aux documents (**imaginer cette opération sur le web**)

Difficile à maintenir

– La terminologie évolue, plusieurs termes sont rajoutés tous les jours

Processus humain donc subjectif

– Des termes différents peuvent être affectés à un même document par des indexeurs différents

Les utilisateurs ne connaissent pas forcément le vocabulaire utilisé

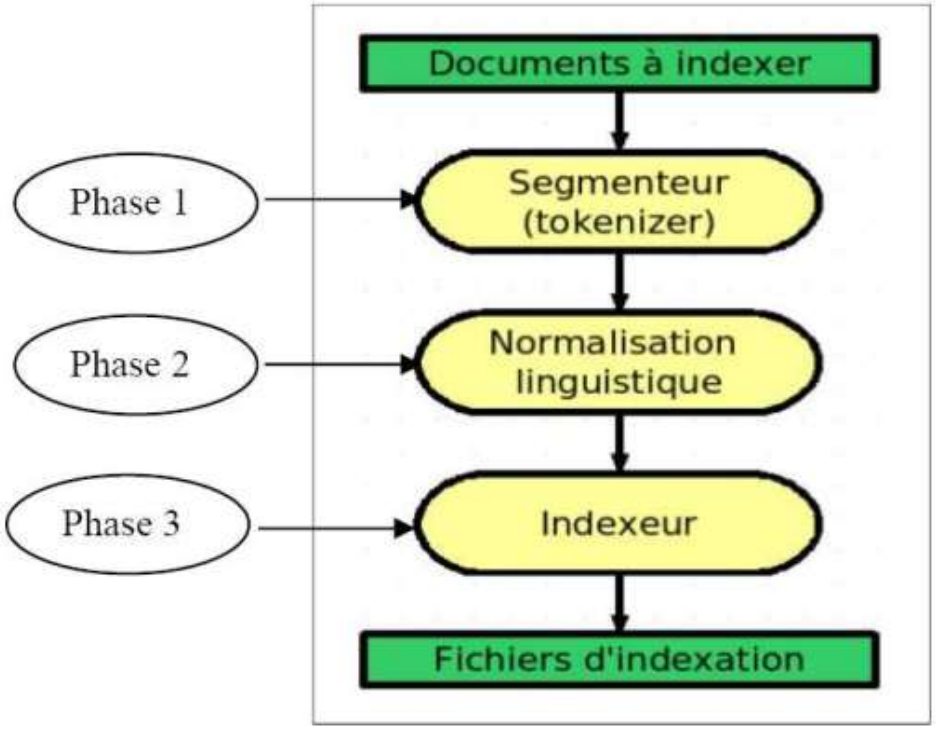
par les indexeurs

INDEXATION AUTOMATIQUE

- **Approches basées sur**
  - Statistique (distribution des mots) et/ou TALN (compréhension du texte)
- **Approche repose sur des hypothèses simples:**
  - Redondance d’un mot marque son importance
  - Cooccurrence des mots marque le sujet d’un document

INDEXATION AUTOMATIQUE: PROCESSUS

- Le processus de l'indexation se constitue de :



INDEXATION AUTOMATIQUE
ETAPE 1 : EXTRACTION DES MOTS

- Extraire les termes (tockenization)
  – terme = suite de caractères séparés par (blanc ou signe de ponctuation, caractères spéciaux,…), Nombres
- Ce sont les index utilisés lors de la recherche
- Dépend de la langue

INDEXATION AUTOMATIQUE
ETAPE 2 : NORMALISATION

- Cette phase peut contenir plusieurs étapes.
- Les étapes les plus importantes et les plus utilisées :
  1. *Elimination des mots vides*
  2. *La racinisation* (« stemming » en anglais)
  3. *La lemmatisation*
  4. *Extraction des mots composés*
  5. *Extraction des entités nommées,*

NORMALISATION: ELIMINATION DES MOTS VIDES

- Les mots vides (article, proposition, conjonction, etc.) sont des mots non significatifs dans un document, car ils ne traitent pas le sujet du document.
- On distingue deux techniques pour éliminer les mots vides :
  - L'utilisation d'une liste préétablie de mots vides (*stop-words*),
  - L'élimination des mots ayant une fréquence qui dépasse un certain seuil dans la collection.
- L'élimination des mots vides réduit la taille de l'index, ce qui améliore le temps de réponse du système.

NORMALISATION: STEMMING

- La normalisation consiste à représenter les différentes variantes d'un terme par un format unique appelé lemme ou racine. Ce qui a pour effet de réduire la taille de l'index.

- Plusieurs stratégies de normalisation sont utilisées :
  -la table de correspondance,
  -l'élimination des affixes **(l'algorithme de Porter),**
  -la troncature,
  -l'utilisation des N-grammes.

NORMALISATION: STEMMING

- Exemple:

  - économie, économiquement, économiste, ➔ économ
  - pour l'anglais : retrieve, retrieving, retrieval, retrieved, retrieves ➔ retriev
- L'inconvénient majeur de cette opération est qu'elle supprime dans certains cas la sémantique des termes originaux,

INDEXATION AUTOMATIQUE: INDEX INVERSÉ

- Une fois les documents indexés :
  – chaque document aura donc un descripteur (une liste de mots souvent simples): à Sac de mots (Bag of Words)
  – Ces termes sont ensuite stockés dans une structure appelée fichier inverse.

INDEXATION AUTOMATIQUE: INDEX INVERSÉ

- Dans sa forme la plus simple, l'index inversé d'une collection de documents est essentiellement une structure de données qui relie chaque terme distinct à une liste de tous les documents qui le contiennent.

INDEX INVERSÉ

- Un index inversé se compose de deux parties:

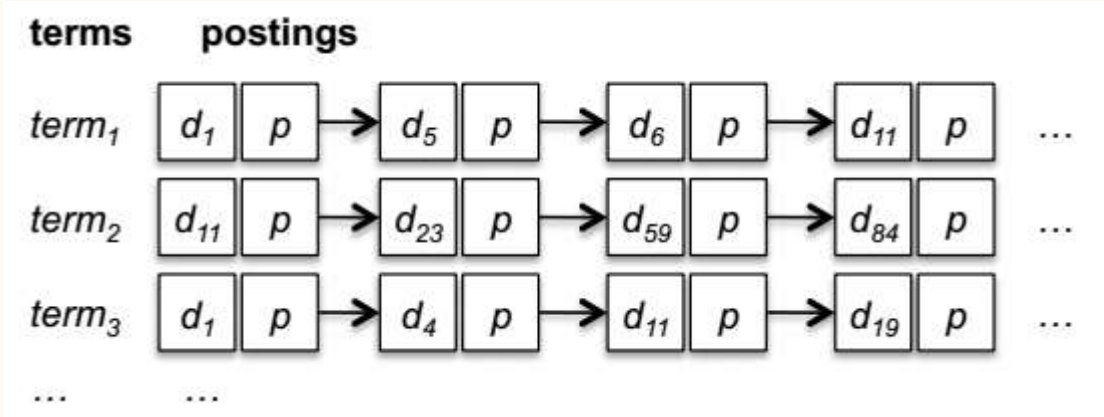
1. Un vocabulaire V, contenant tous les termes distincts de l'ensemble de documents, et
2. pour chaque terme distinct, une liste inversée de publications.

INDEX INVERSÉ

- Chaque enregistrement stocke l'ID (désigné par id_j) du document d_j qui contient le terme t_i et d'autres informations sur ce terme dans ce document.
- Selon le besoin de l'algorithme de recherche ou de classement, différentes informations peuvent être incluses.

INDEX INVERSÉ

- Pour chaque terme, nous avons une liste qui enregistre dans quels documents le terme se produit.
- Chaque terme de la liste est appelé classiquement **Posting** (publication).
- Un Posting est un tuple de la forme(t_i,d_j), où t_i est un identificateur de terme et d_j est un identifiant de document.
- La liste est appelée liste de posting (ou liste inversée)



INDEX INVERSÉ

- Les index inversés sont indépendants du modèle IR adopté (Modèle booléen, modèle d'espace vectoriel, etc.)
- Chaque Posting contient généralement:
  - L'identifiant du document lié.
  - La fréquence d'apparition du terme dans le document
  - La position du terme pour chaque document (facultatif)
    - Exprimé en nombre de mots depuis le début du document, le nombre d'octets, etc.
- Pour chaque terme est également généralement stocké la fréquence d'apparition du terme dans l’ensemble des documents.

INDEX INVERSÉ: CONSTRUCTION

- Extraire les termes de chaque document dans un fichier (1fichier par document) ou un fichier pour plusieurs documents)

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

INDEX INVERSÉ: CONSTRUCTION

- Trier le fichier termes-documents:

Trier le fichier par ordre alphabétique des termes et par document

Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Term	Doc #
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

INDEX INVERSÉ: CONSTRUCTION

- Pour chaque terme,
  - on dispose de la liste de documents qui le contient
  - Le nombre de documents comportant ce terme

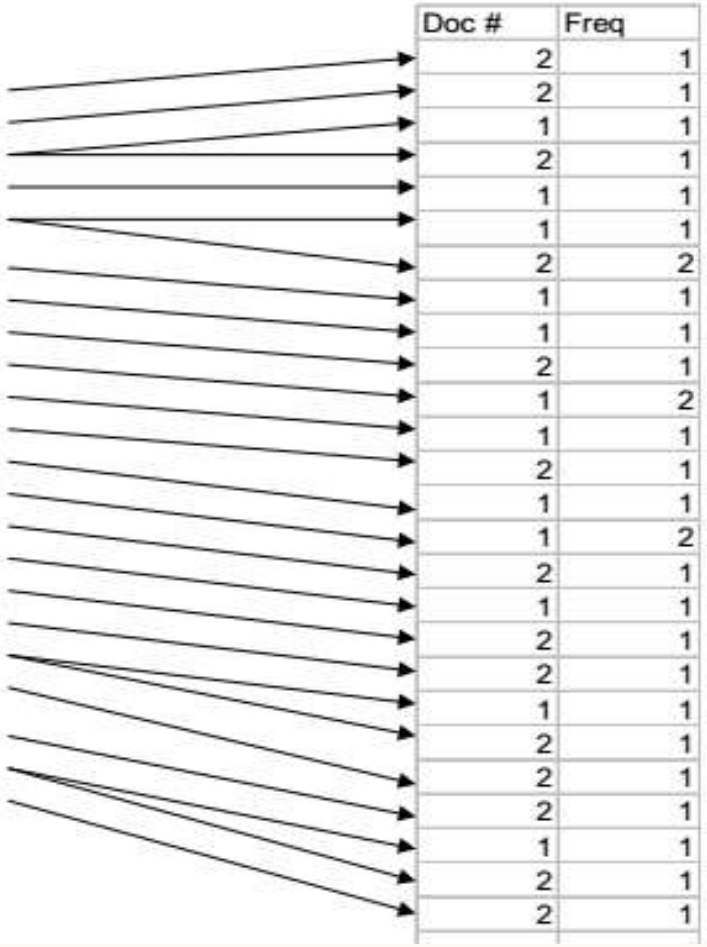
Term	Doc #
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

Term	Doc #	Freq
ambitious	2	1
be	2	1
brutus	1	1
brutus	2	1
capitol	1	1
caesar	1	1
caesar	2	2
did	1	1
enact	1	1
hath	2	1
I	1	2
i'	1	1
it	2	1
julius	1	1
killed	1	2
let	2	1
me	1	1
noble	2	1
so	2	1
the	1	1
the	2	1
told	2	1
you	2	1
was	1	1
was	2	1
with	2	1

INDEX INVERSÉ: CONSTRUCTION

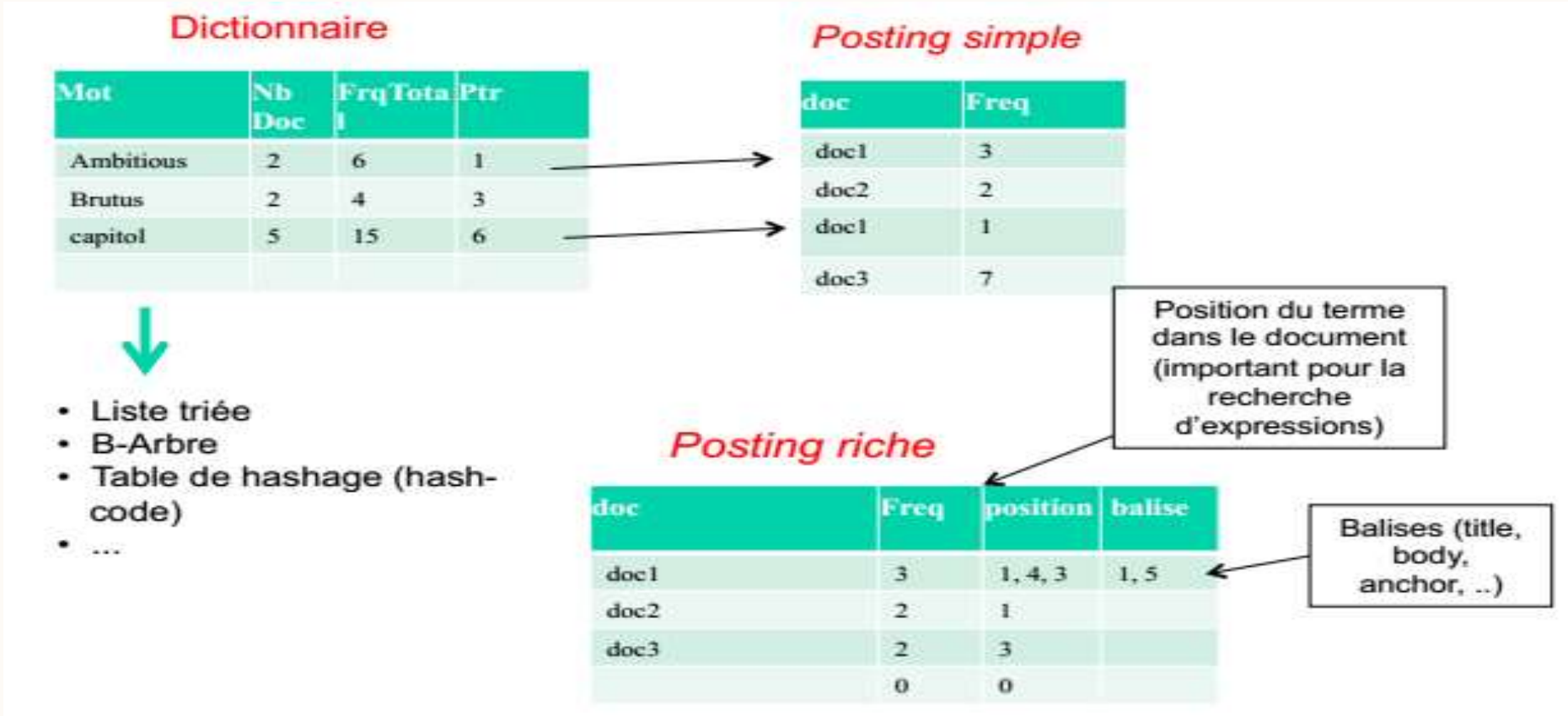
Term	Doc #	Freq
ambitious	2	1
be	2	1
brutus	1	1
brutus	2	1
capitol	1	1
caesar	1	1
caesar	2	2
did	1	1
enact	1	1
hath	2	1
I	1	2
i'	1	1
it	2	1
julius	1	1
killed	1	2
let	2	1
me	1	1
noble	2	1
so	2	1
the	1	1
the	2	1
told	2	1
you	2	1
was	1	1
was	2	1
with	2	1

Term	N docs	Tot Freq
ambitious	1	1
be	1	1
brutus	2	2
capitol	1	1
caesar	2	3
did	1	1
enact	1	1
hath	1	1
I	1	2
i'	1	1
it	1	1
julius	1	1
killed	1	2
let	1	1
me	1	1
noble	1	1
so	1	1
the	2	2
told	1	1
you	1	1
was	2	2
with	1	1



Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

INDEX INVERSÉ: ORGANISATION



INDEX INVERSÉ: STOCKAGE

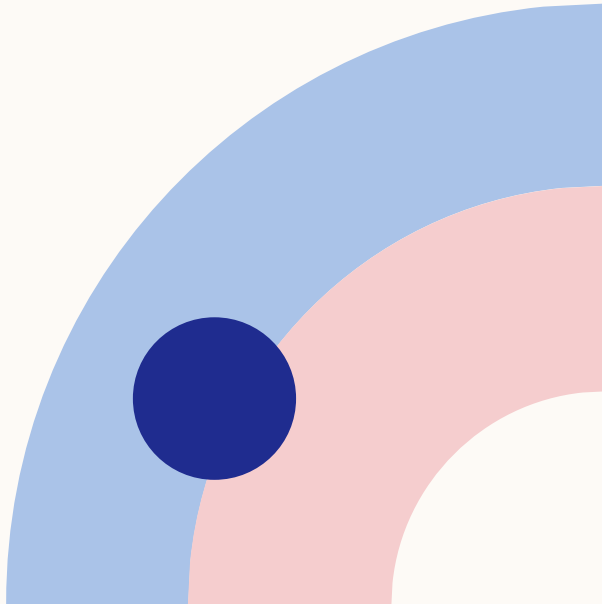
Les termes se trouvent généralement dans un certain nombre de documents:

- Les index inversés réduisent les besoins de stockage de l'index,
- fournir la base pour une recherche efficace
- cette structure d'index inversé est essentiellement sans rivaux comme la structure la plus efficace pour supporter la recherche de texte.

INDEX INVERSÉ: STOCKAGE

Les listes liées généralement préférées aux tableaux:

- Allocation dynamique de l'espace
- L'insertion de termes dans les documents est facile
- Frais généraux des pointeurs
- Espace requis

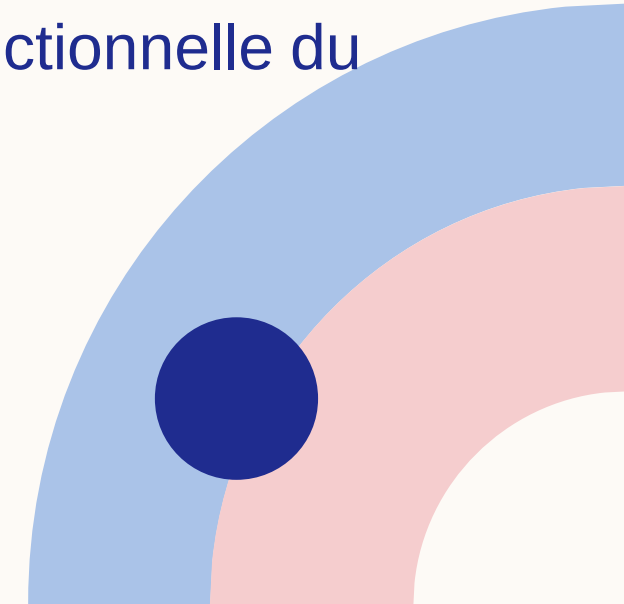


INDEXATION DISTRIBUÉE

Pour de très larges collections (Web).

Un serveur principal dirige le tout

- Il divise la tâche d'indexation en un ensemble de tâches parallèles
- Il assigne chaque tâche à une machine libre et fonctionnelle du réseau



MERCI