

Plan

- Fondements de la Recherche d'information (RI)
 - Définition et contours de la RI
 - Problématique de la RI
 - Ouvrir la boîte d'un système de RI
- Quelques thématiques de recherche en RI
- Conclusion

Qu'est ce que la RI ?

- 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**



Qu'est ce que la RI ?



- Information retrieval (IR)
 - is the science of **searching for documents**, for **information within documents**, and for metadata about documents, as well as that of searching relational databases and the World Wide Web.
- IR is interdisciplinary,
 - **based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics, and physics.**
 - are used to reduce what has been called "information overload".
 - Many universities and public libraries use IR systems to provide access to books, journals and other documents. **Web search engines are the most visible IR applications.**

Qu'est ce que la RI ?

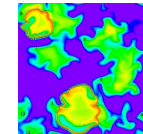


.. Mais pas seulement

- Plusieurs domaines d'application
 - Internet (Web, Forum/Blog search, news)
 - Entreprises (entreprise search)
 - Bibliothèques numériques «digital library»
 - Domaine spécialisé (médecine, droit, littérature, chimie, mathématique, brevets, software, ...)
 - Nos propres PC (Yahoo! Desktop search)

Origine de l'information

Enterprise Apps



Scientific data

Web Apps



Social Media Data

Device explosion



Machine Data



Information est partout → Gros volume

- Average Number of Tweets Sent Per Day:
500 million
 - 2 billions queries per day on twitter
- Every minute 510,000 posted comments
FaceBook
- 45 milliards (Google), 25 milliards (Bing)
- 672 Exabytes - 672,000,000,000
Gigabytes (GB) of accessible data.



Le problème ..

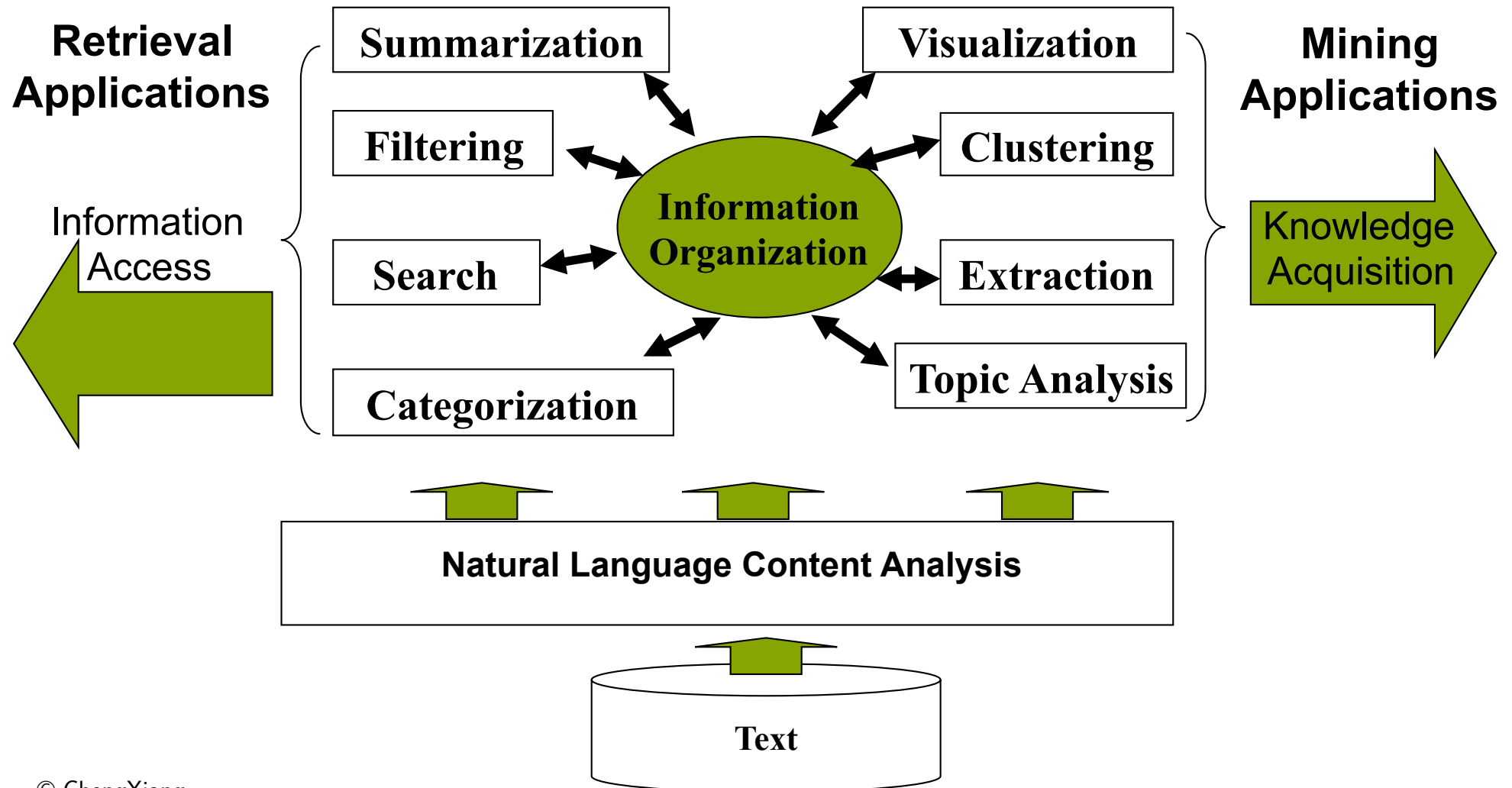
- n'est pas tant la disponibilité de l'information
MAIS
- sa sélection, son identification → arriver à trouver au bon moment l'information utile



RI et domaines liés

- Recherche d'information s'intéresse à l'accès à des textes (documents textuels)
- Domaines liés
 - Fouille de textes (Text Mining)
 - Détection des « patterns » dans un texte (latent topics, ...)
 - Extraction d'information
 - Représentation des connaissances
 - Traitement automatique de la langue (NLP)
 - Compréhension d'un texte
 - Représentation sémantique d'un texte
 - Bases de données
 - Gestion de données structurées
 - Requêtes précises

Tâches de RI : vue large



- Recherche adhoc (Search)
 - Je cherche des infos (pages web) sur un sujet donné
 - Je sou mets une requête → retour liste de résultats
 - Requête «recherche d'info» → SRI → renvoie une liste de documents traitant de la » recherche d'information »
 - Plusieurs types de RI adhoc
 - Recherche adhoc (tâches spécifiques)
 - Domaine spécifique (médical, légal, chimie, ...)
 - Recherche d'opinions (Opinion retrieval) (sentiment analysis)
 - Recherche d'événements
 - Recherche de personnes (expert)

Tâches de RI

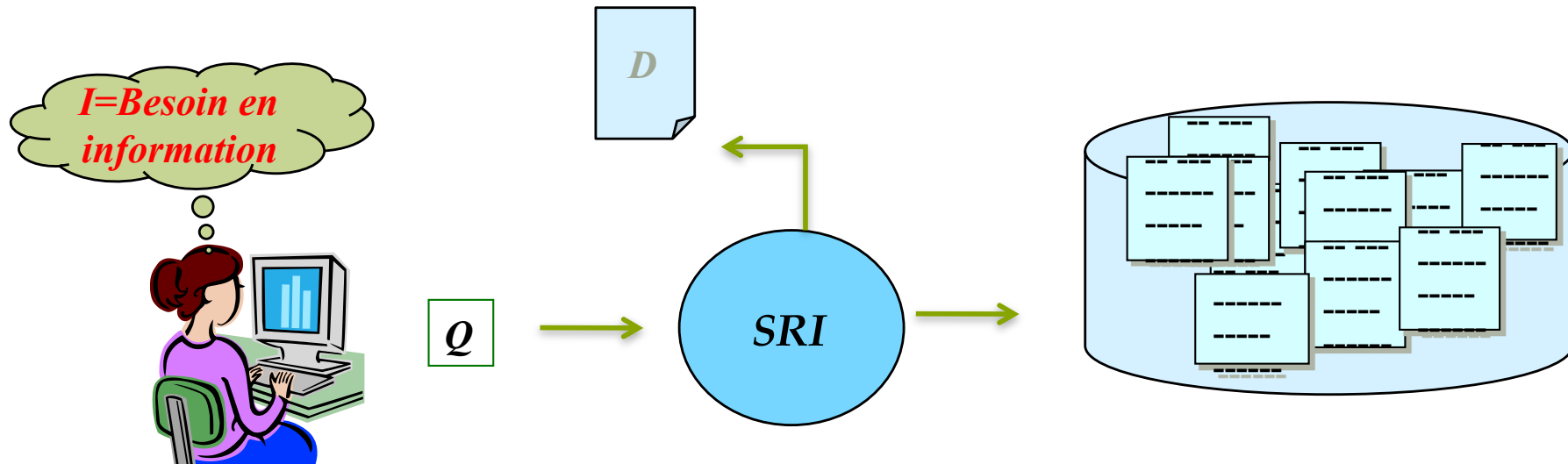
- Classification / Catégorisation
 - Classer les documents dans des catégories prédéfinies
- Filtrage d'information/ recommandation (filtering/ recommendation)
 - Recommandation
 - Dissémination sélective d'information
 - Système d'alerte
 - Dissémination sélective d'information
 - Push
 - Profilage (profiling)

Tâches de RI

- Résumé (Summarization)
 - Construire un résumé concis à partir d'un ou plusieurs documents
- Question-réponses (*Query answering*)
 - Chercher des réponses à des questions
 - par exemple
 - « Qui est averroes ? »
 - « Quelle la hauteur du Mont Blanc ? »
- Analyse thématique (Topic Analysis)
 - Extraire et analyser les thèmes des documents
- Partitionnement (Clustering)
 - Partitionner (regrouper) les documents (phrases, textes) en fonction de critères qu'ils partagent

Plan

- Fondements de la Recherche d'information (RI)
 - Introduction : définition, contours de la RI
 - Problématique de la RI
 - Ouvrir la boîte d'un système de RI
- Panorama RI – scénarios et applications
 - Thématiques de recherche en RI
- Conclusion



- Sélectionner dans une collection
 - les informations (items, documents, ..)
 - ... pertinentes répondant aux
 - ... besoins en information des utilisateurs

Information – document

- Formes

- Texte, images, sons, vidéo, graphiques, etc.
- Exemples texte : web pages, email, livres, journaux, publications, blog, Word™, Powerpoint™, PDF, forum postings, brevets, etc.

- Hétérogénéité

- langage (multilingues)
- media (multimédia)



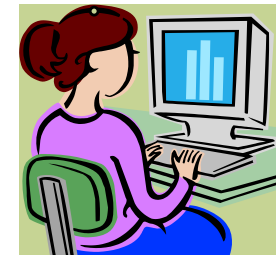
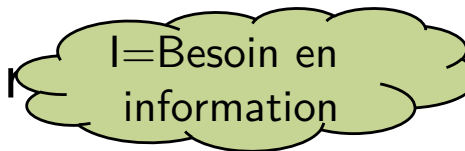
Question

- Information, document, unité/granule/passage
- Comment représenter le contenu du document (sens des mots) → Ambiguïté du langage naturel (polysémie, synonymie, ...)



Besoin en information (information needs)

- Besoin en information est une expression mentale d'un utilisateur
- Requête
 - Ensemble de mots-clés
 - → Une représentation possible du besoin en information



Requête



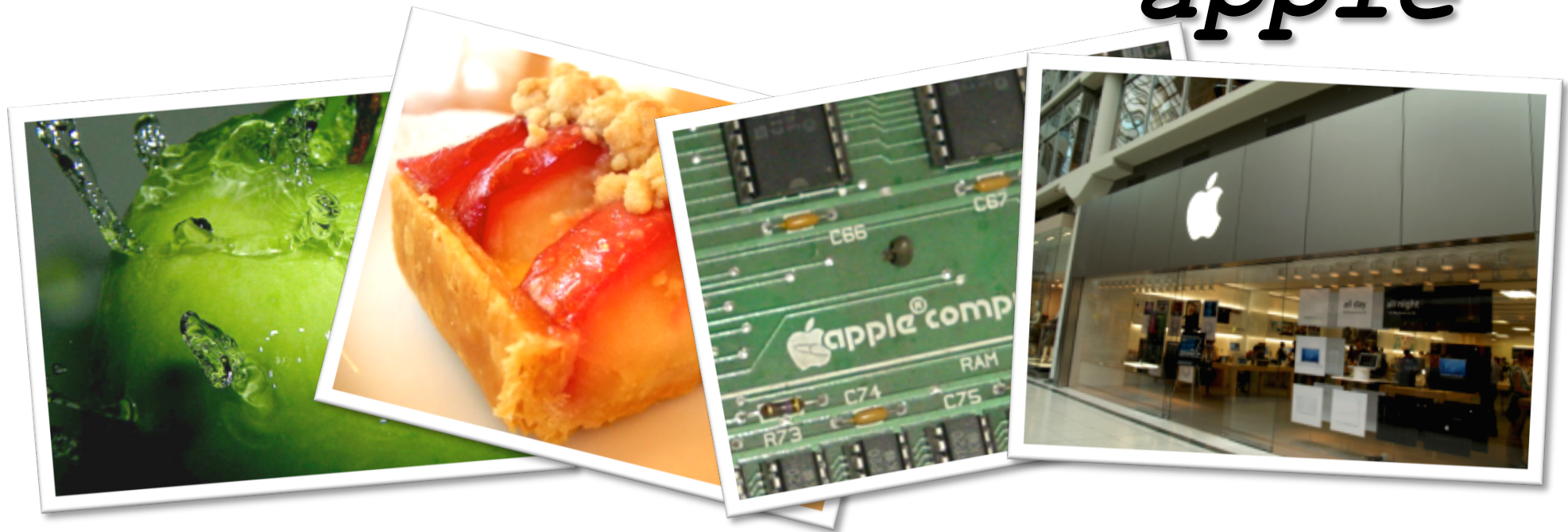
Question

- Comment capturer le besoin de l'utilisateur

Besoin en information-Intention de la requête (query intent)

What's in a query?

apple



© David J. Brenes, Daniel Gayo Avello, Kilian Pérez-González

- Au cœur de tout système de RI
 - Relation entre le document et ... la requête ou le besoin de l'utilisateur ?
- Plusieurs facteurs influencent la décision de l'utilisateur, tâche, le contexte, nouveauté, style, compréhension, temps, ...
- Pertinence par document

Goffman, 1969: '...the relevance of the information from one document depends upon what is already known about the subject, and in turn affects the relevance of other documents subsequently examined.'

Pertinence - relevance

Type of relevance(survey) (Saracevic 2007)

- Plusieurs pertinences
 - **Thématique** (topical): relation entre le sujet exprimé dans la requête et le sujet couvert dans le document.
 - **Contextuelle** (Situation) : relation entre la tâche, le problème posé par l'utilisateur, la situation de l'utilisateur et l'information retrouvée.
 - **Cognitive** : relation entre l'état de la connaissance de l'utilisateur et l'information sélectionnée



Question

- Processus subjectif (humain), dépend de plusieurs facteurs
→ difficile à automatiser

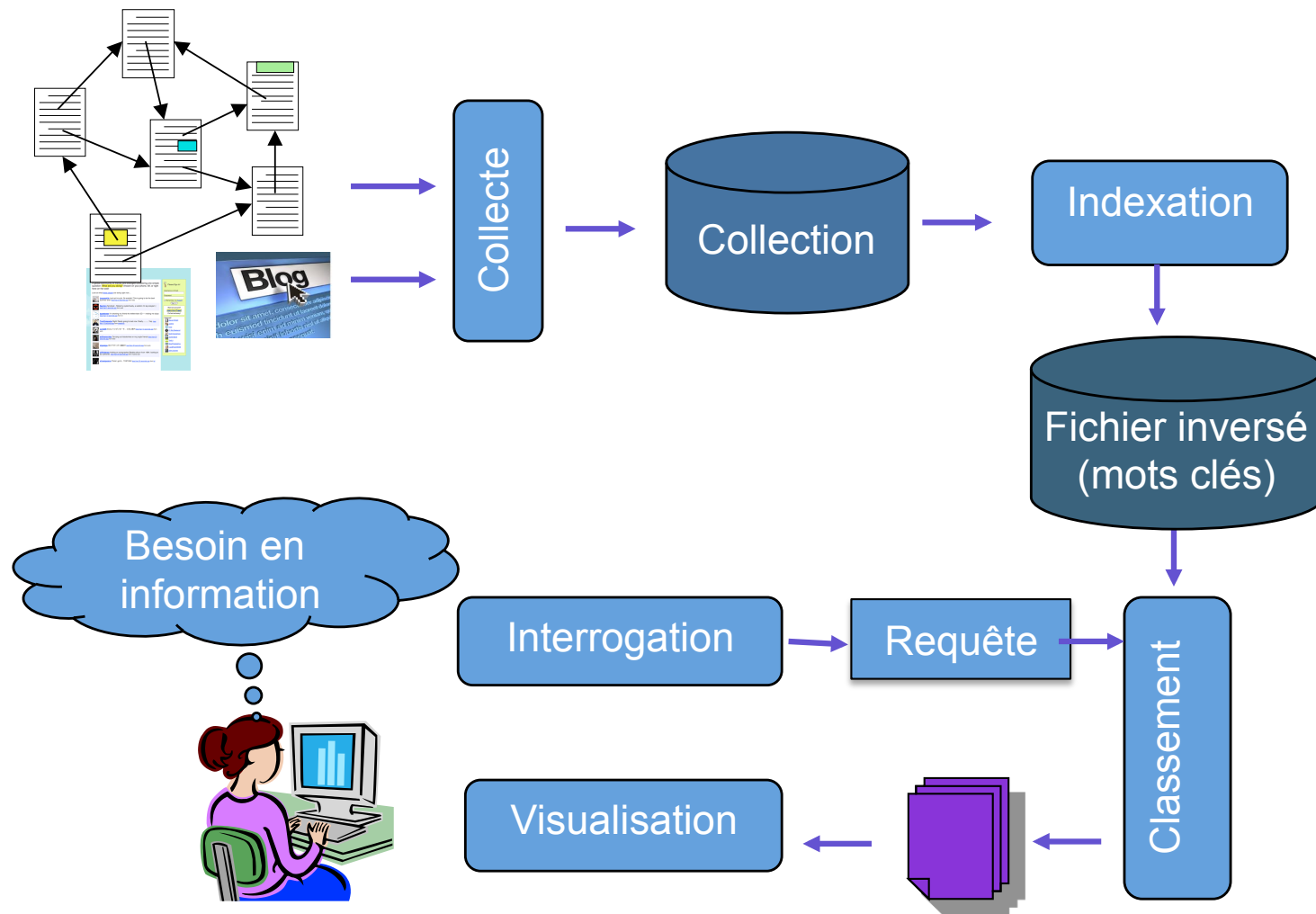
Hypothèses de base

- Besoin = requête
 - Besoin confondu avec la requête utilisateur (une liste de mots clés)
- Document et requête
 - Représentés par des termes (mots simples, groupes de mots, ...) → Sac de mots
- Pertinence
 - Traduite par la similarité de vocabulaire (mots) entre la requête et le document → thématique

Démarche RI

- Interpréter le texte au lieu de le comprendre
- Représenter le contenu comme une liste de mots simples (sac de mots « Bag of words »)
- Exploiter des indicateurs statistiques (comptage de mots) du texte pour évaluer sa pertinence

Architecture d'un SRI

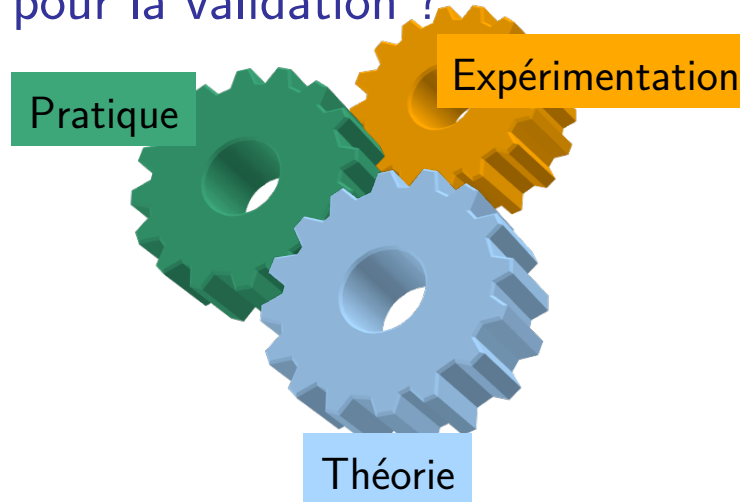


Recap.

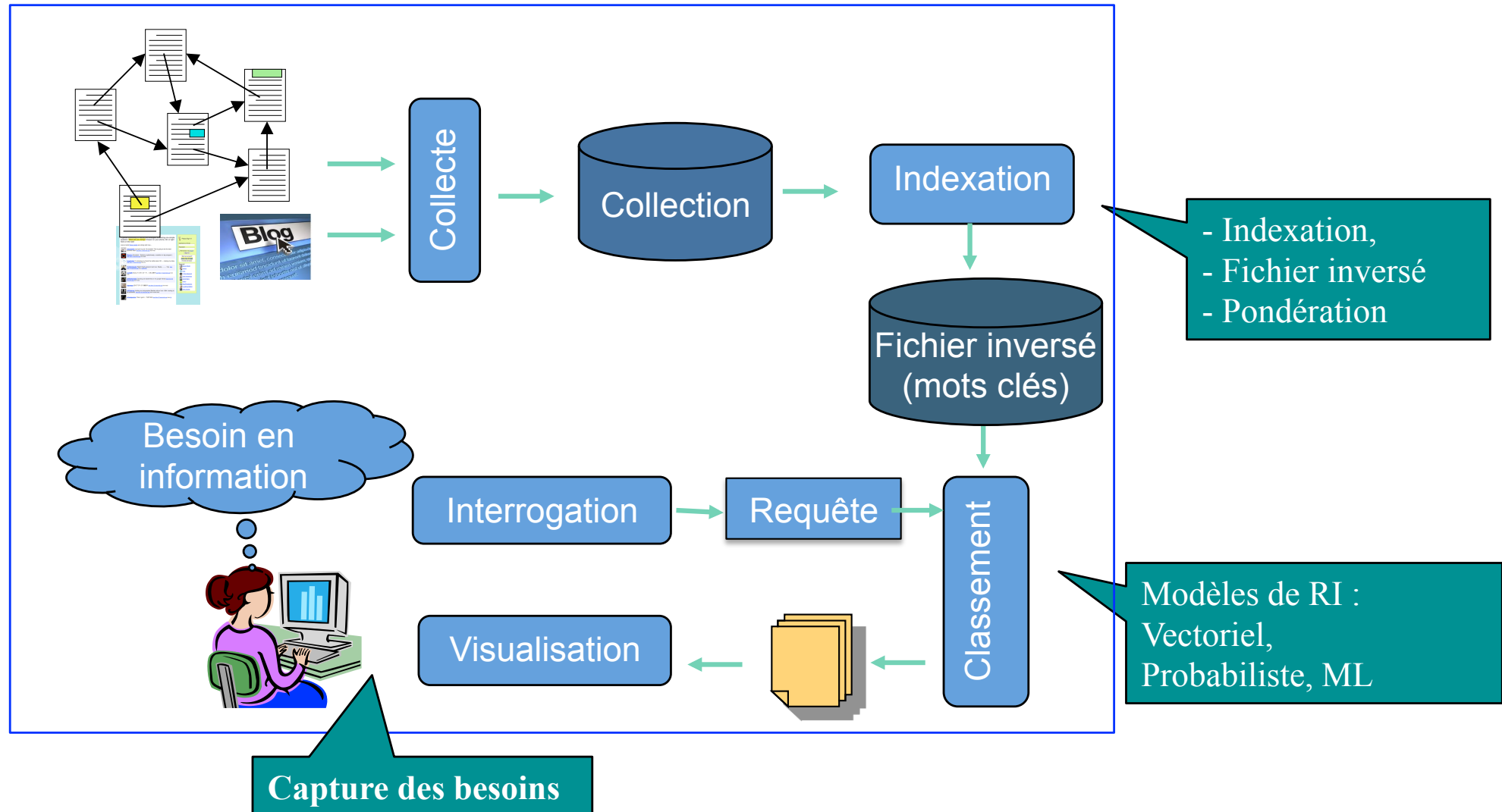
- Représentation (indexation) du document
 - Comment construire une représentation à partir du document ?
 - Quelle organisation physique pour ces index?
- Représentation des besoins
 - Comment capturer le besoin de l'utilisateur ?
 - Comment exprimer le besoin (langage de requêtes) ?
- Comparaison/ranking document-requête (des représentations)
 - Comment mesurer (décider) la pertinence d'un document ?
- Évaluation des performances
 - Comment comparer les SRI ?
 - Quelle démarche (empirique/ analytique) ?
 - Quelles métriques ?

Recap.

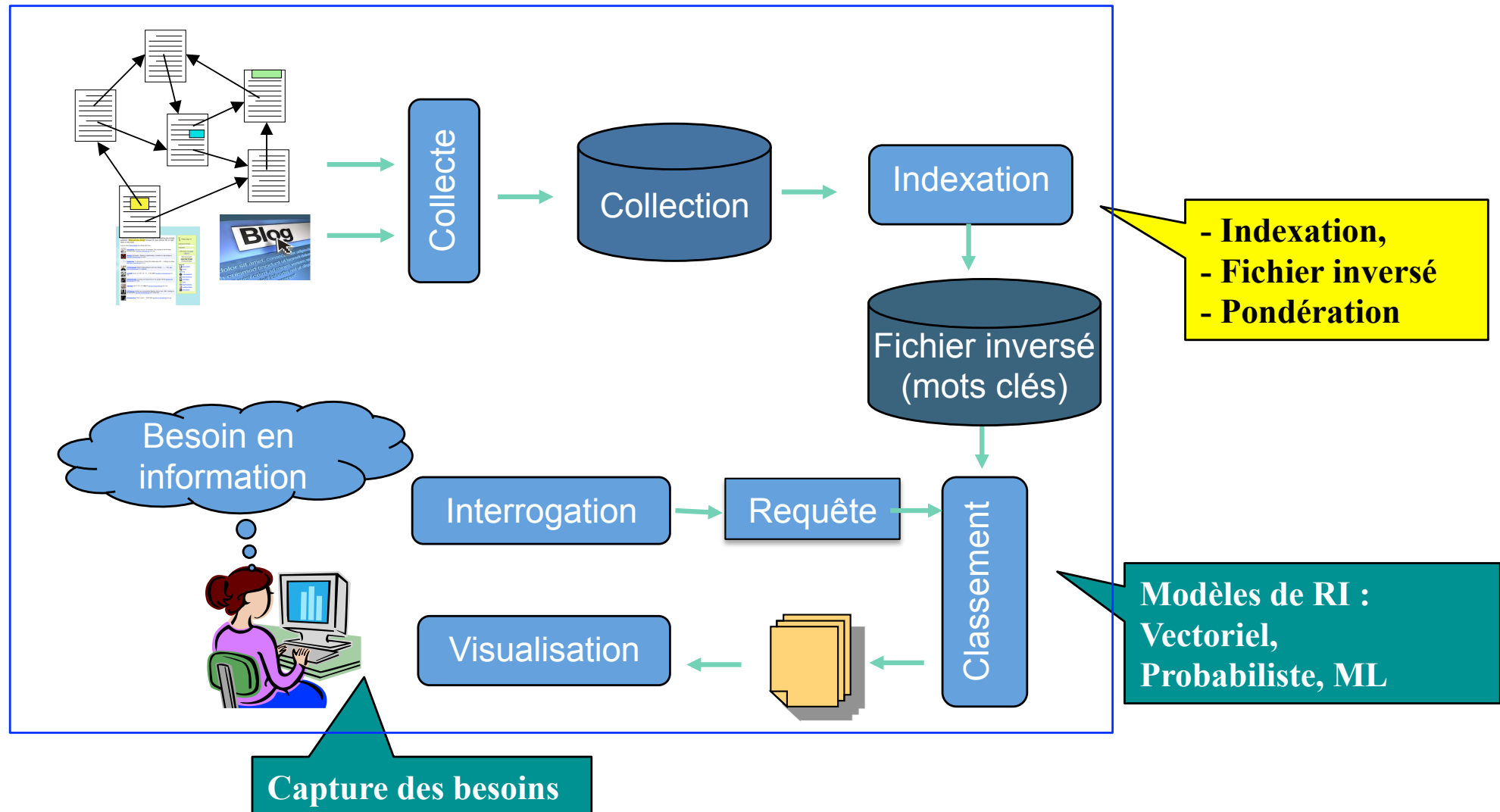
- Proposer des solutions
 - modèles, techniques, approches, outils pour répondre à ces problèmes
- ...avec 2 soucis majeurs
 - Quels supports théoriques ?
 - Souvent basés sur des théories mathématiques : Probabilités, statistiques, ensembles, algèbre, logique floue, analyse de données, ...
 - Quel(s) processus pour la validation ?



Ouvrir la “boîte d’un système de RI”



Ouvrir la “boîte d’un SRI”



Définition

- Processus permettant de construire un ensemble d’éléments « clés » permettant de caractériser le contenu d’un document / retrouver ce document en réponse à une requête
- Approches
 - Guidée par un vocabulaire contrôlé vs. Libre
 - Statistique (distribution des mots) et/ou TALN (compréhension du texte)
 - Approche courante est plutôt statistique avec des hypothèses simples
 - Redondance d’un mot marque son importance
 - Cooccurrence des mots marque le sujet d’un document
- Résultat d’une indexation textuelle
 - Chaque document est représenté par une liste d’index (de mots clés)
 - L’index permet d’accéder (sélectionner) aux documents

Approche générale

- Décomposer le texte (Parsing)
- Segmenter les séquences de caractères en mots (Tokenizing)
- Normaliser
 - Textuelle: ponctuation, dates, case
 - Linguistique : Racinisation (stemming)/ lemmatisation
- Supprimer les mots communs (stop word removal)
 - En fonction d'une "short list" "the", "and", "or", ou mots fréquents
- Regrouper les mots

<Title>: Information retrieval
(Corps du texte> : Information retrieval (IR)
is the science of searching of documents.
N.I.S.T launched TREC

Information retrieval Information retrieval IR
is the science of searching of documents
N.I.S.T launched TREC

information retrieval information retrieval IR
is the science, of **search**, document **NIST**
launched TREC

information, retrieval, information, retrieval,
IR, science, **search**, document NIST **launch**
TREC

information 2, retrieval 2, IR 1, science 1,
search 1, document 1, NIST 1, launch 1,
TREC 1

Un sac de mots
(BOW)

Ségmentation (tokenization)

- Pas d’espaces en chinois et en japonais
 - Ne garantit pas l’extraction d’un terme de manière unique

Chinese tokenization

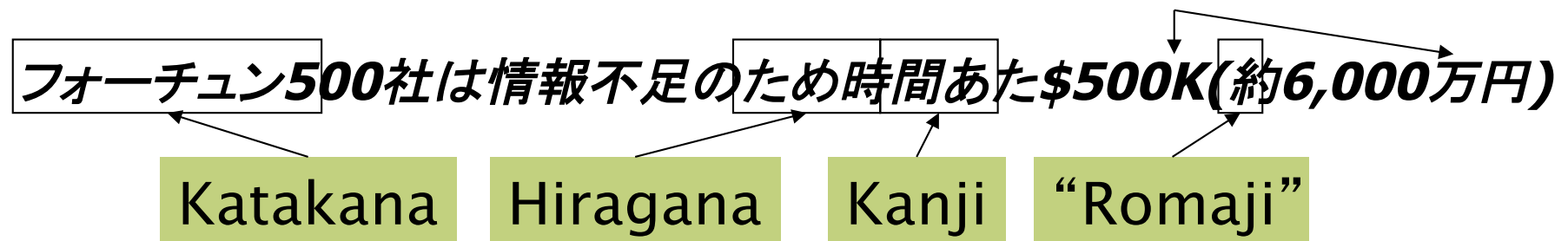
1. Original text

旱灾在中国造成的影响

(the impact of droughts in China)

Ségmentation (tokenization)

- Pas d’espaces en chinois et en japonais
 - Ne garantit pas l’extraction d’un terme de manière unique
- Japonais encore plus compliqué avec différents alphabets



L’ utilisateur peut exprimer sa requête entièrement en Hiragana

Fichier Inversé (Inverted File)

d1:
So let it be
with
Caesar. The
noble
Brutus hath
told you
Caesar was
ambitious

d2:
I did enact
Julius
Caesar I
was killed
i' the
Capitol;
Brutus
killed me.

Traitement
=
Indexation

Term	N docs	Tot Freq	Ptr
ambitious	1	1	1
be	1	1	2
brutus	2	2	3
capitol	1	1	5
caesar	2	3	6
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
1	1
2	1
2	1
1	1
2	1
2	1

d1:
So let it be with
Caesar. The noble
Brutus hath told you
Caesar was
ambitious

d2:
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

d3:
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

Exemple issu de C. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008

Fichier Inversé (Inverted File)

Dictionnaire

Mot	Nb Doc	FrqTotal	Ptr
Ambitious	2	6	1
Brutus	2	4	3
capitol	5	15	6

Posting simple

doc	Freq
doc1	3
doc2	2
doc1	1
doc3	7



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

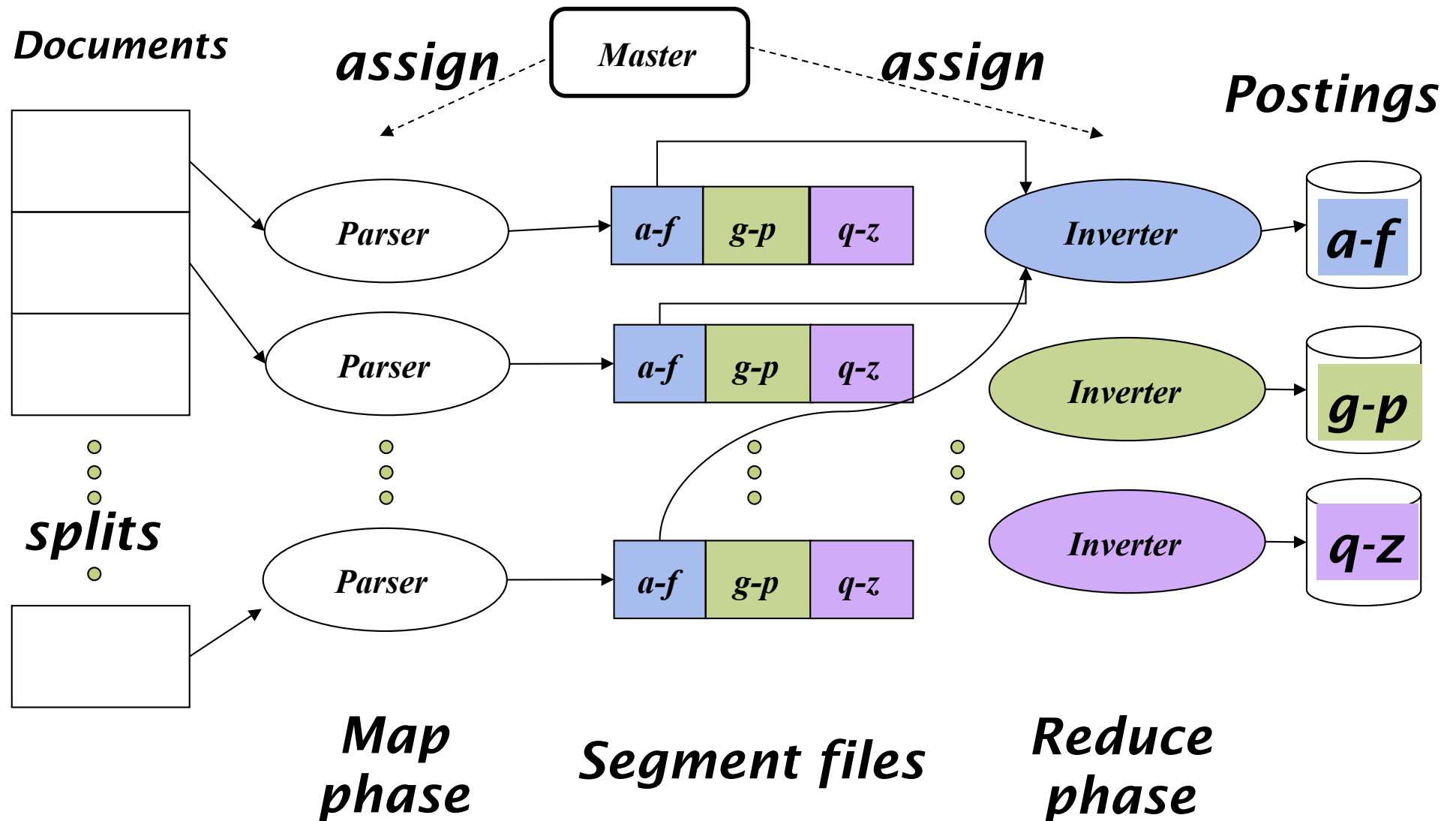
Posting riche

doc	Freq	position	balise
doc1	3	1, 4, 3	1, 5
doc2	2	1	
doc3	2	3	

Position du terme dans
le document
(important pour la
recherche
d'expressions)

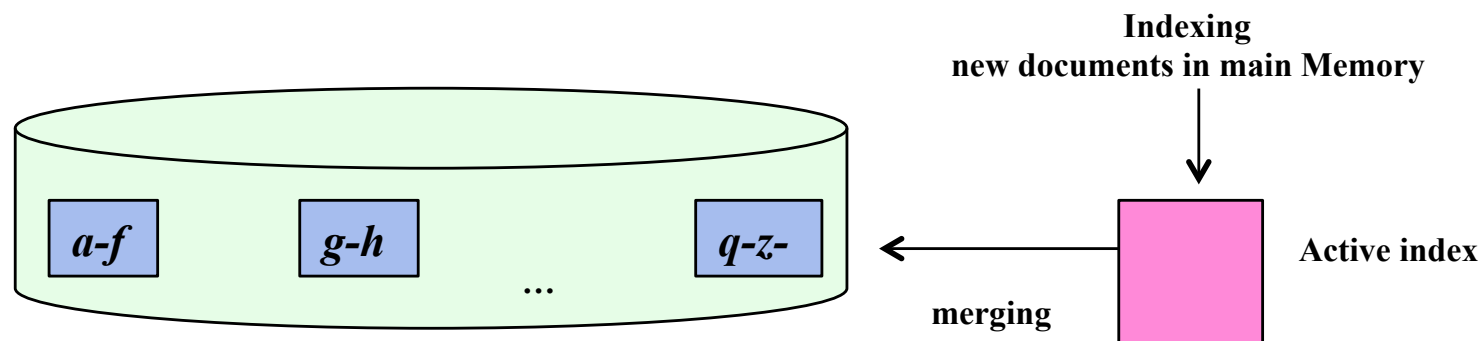
Balises (title,
body,
anchor, ..)

Construction fichier Inversé (MapReduce)

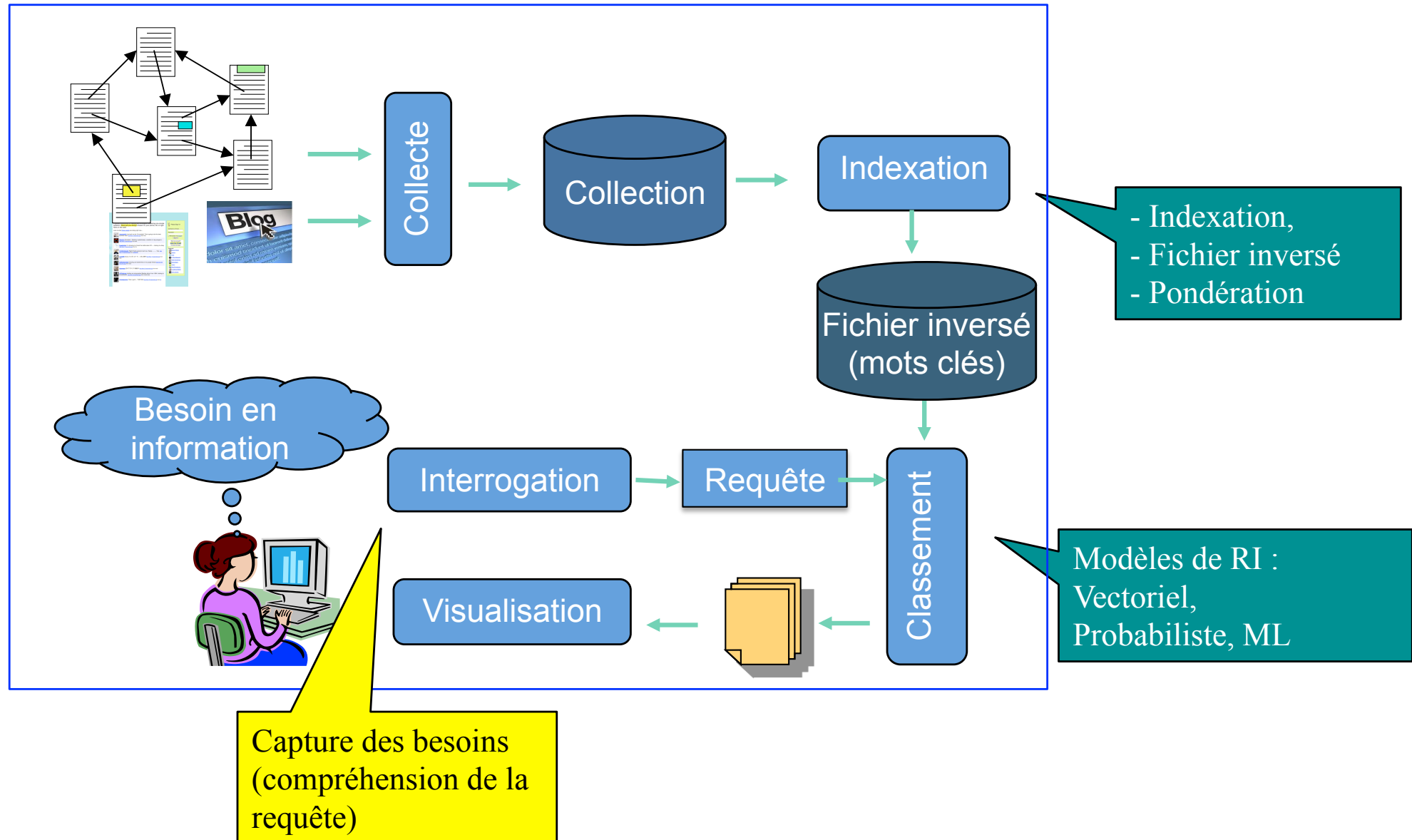


MAJ de l’index - Indexation dynamique

- Les collections sont souvent dynamiques (cas du Web)
 - → Ajout suppression et modification de documents
 - → Mise à jour du dictionnaire et du *posting*
- Solution (Simple mais inefficace) :
 - Reconstruire l'index à partir de zéro
- Une meilleure solution :
 - Maintenir un index en mémoire qui garde la trace de tous les changements
 - Dès que l’index est “plein” fusionner avec celui (ou ceux) du disque



Ouvrir la boîte d'un SRI



Capture des besoins-compréhension de la requête

- Requête idéale difficile à construire (l'utilisateur ne sait pas décrire ce qu'il recherche)
- Mieux capturer les besoins (l'intention)
 - Suggestion de requête, correction, ...
 - Identification de l'intention de la requête
 - Aller au delà d'une requête unique

Je reviens sur ce point plus loin