



Introduction à la RI

Mohand Boughanem

bougha@irit.fr

<http://www.irit.fr/~Mohand.Boughanem>

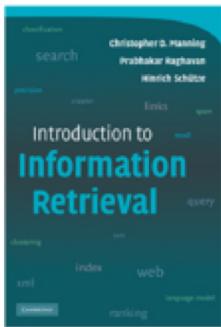
Université Paul Sabatier de Toulouse

Laboratoire IRIT , UMR5055

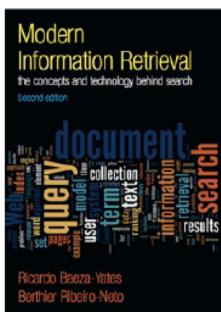
- Recherche d'information (RI) :

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





IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).



Information retrieval deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organization of the information items should be such as to provide the users with easy access to information of their interest.



- 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.



- 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)

- 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.

- 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

• Recherche adhoc (search)

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

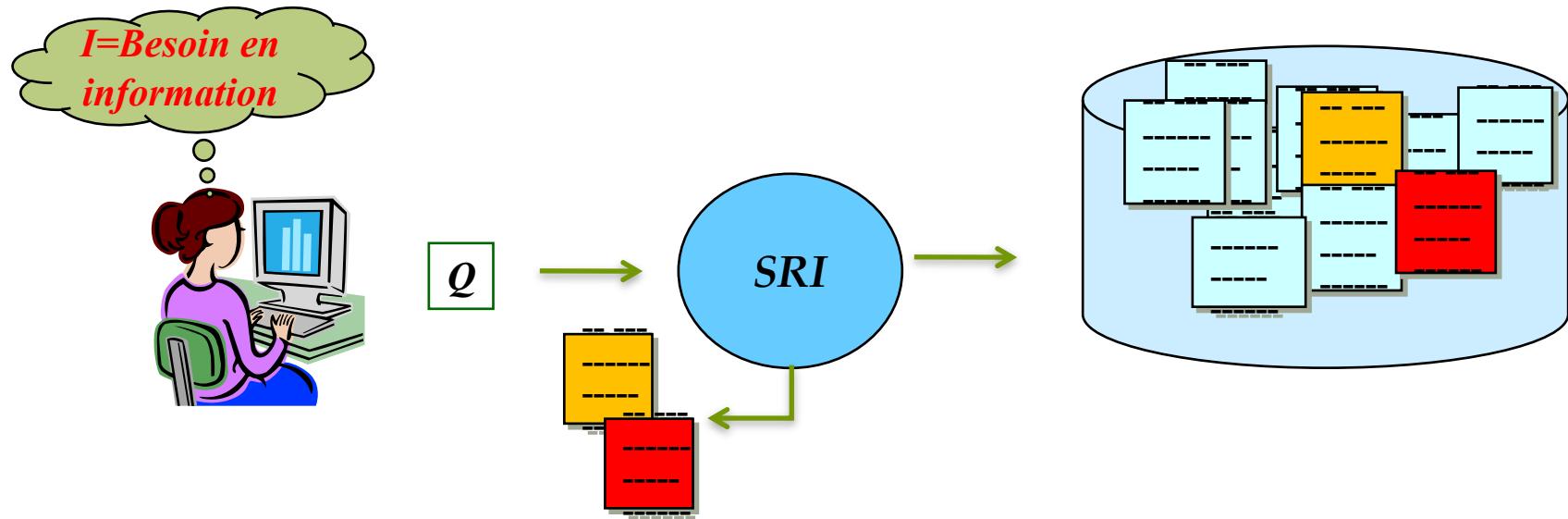
- Question-réponses (*Query answering*)
 - Chercher des réponses à des questions
 - par exemple
 - « où se trouve l'N7? »
 - « quel est l'âge de Macron ?»
- Catégorisation(classification)
 - Classer identifier un document dans une des catégories prédéfinies
- Clustering (partitionnement)
 - Regrouper les objets (textuels) en fonction de caractéristiques communes

- Filtrage d'information/ recommandation
(filtering/recommendation)
 - Recommandation
 - Dissémination sélective d' information
 - Système d' alerte
 - Dissémination sélective d' information
 - Push
 - Profilage (profiling)
- Résumé automatique
 - document summarization
 - multidocument summarization

Plan

- Introduction
- Chapitre 1 : Concepts de base de la RI
- Chapitre 2 : Représentation de l'information
- Chapitre 3 : Modèles de RI : modèle vectoriel, modèles probabilistes
- Chapitre 4 : Evaluation des performances en RI
- Chapitre 5 : Représentation distribuée des termes
- Chapitre 6 : Apprentissage automatique et RI
- Chapitre 7: RI et WEB
- Conclusion

Chapitre 1 : concepts de base de la RI



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

• 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



- Comprendre le contenu vs. l'interpréter
 - Problème : Ambiguïté du langage naturel (polysémie, synonymie, ...)
 - Information, document, unité/granule/passage

- 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 ?

What's in a query?

apple



- 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.’

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



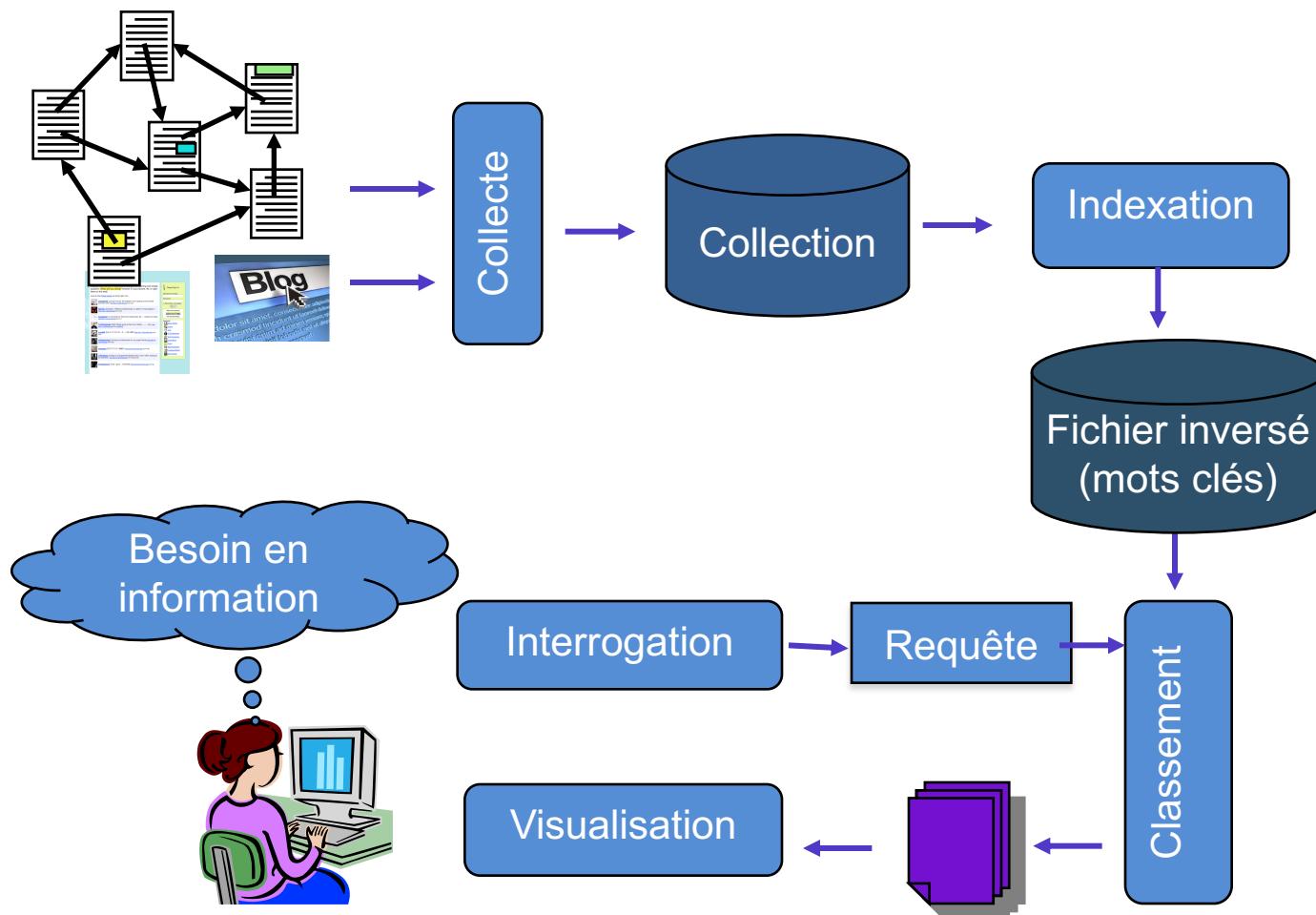
Question

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

- 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
- Exploiter les propriétés statistiques (comptage de mots) du texte plutôt que ses propriétés linguistiques



Concepts de base de la RI

Indexation, Fichier inverse

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.

Indexation :
extraire les
mots,
traitement
statistique/li-
nguistique, ..

Term	N docs	Tot Freq	Ptr	Doc #	Freq
ambitious	1	1	1	2	2
be	1	1	2	1	1
brutus	2	2	3	2	2
capitol	1	1	5	1	1
caesar	2	3	6	2	2
did	1	1			2
enact	1	1			1
hath	1	1			1
I	1	2			2
i'	1	1			1
it	1	1			1
julius	1	1			2
killed	1	2			1
let	1	1			1
me	1	1			2
noble	1	1			1
so	1	1			2
the	2	2			2
told	1	1			1
you	1	1			2
was	2	2			2
with	1	1			2
				1	
				2	
				2	

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

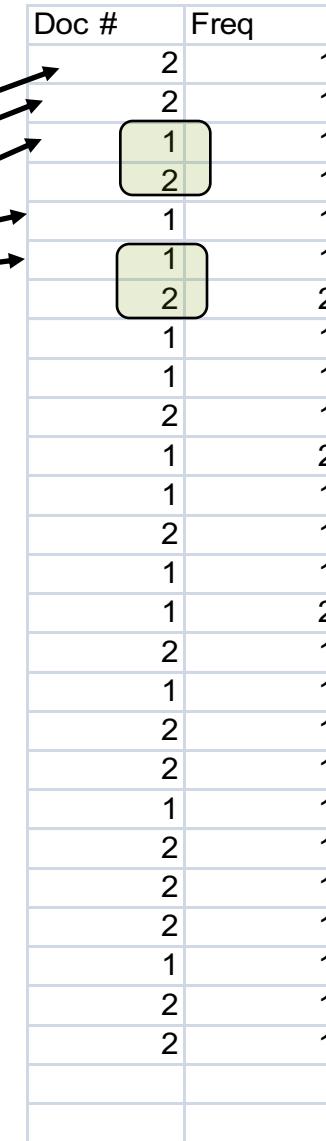
d2:
I did enact Julius
Caesar I was killed
 i' the Capitol;
Brutus killed me.

d3:
I did enact Julius
Caesar I was killed
I the^d capit^ol;
I did^d Julius
I did^d Julius
Caesar I was killed
I did^d Julius
Caesar I was killed



Caesar, brutus

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	



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.

- Facteurs utilisés par la majorité des modèles
 - Fréquence du terme dans le document (**tf**), sa fréquence dans la collection (**idf**), sa position dans le texte(p), taille du document (**dl**) ...

$$\text{Score}(q, d) = \text{fonction}(tf, idf, dl)$$

$$\text{Score}(q, d) = \left(\sum_{t \in q \cap d} f(t, q, d).a(d, q) \right)$$

Exemple

- Plusieurs modèles théoriques pour formaliser cette fonction
- Elle peut être apprise (apprentissage automatique, approche utilisée par la majorité des moteurs de recherche)

- ***PIV (vector space model)***

$$\sum_{w \in q \cap d} \frac{1 + \ln(1 + \ln(c(w, d)))}{(1 - s) + s \frac{|d|}{avdl}} \cdot c(w, q) \cdot \ln \frac{N + 1}{df(w)}$$

TF

- ***DIR (language modeling approach)***

$$\sum_{w \in q \cap d} c(w, q) \times \ln\left(1 + \frac{c(w, d)}{\mu \cdot p(w | C)}\right) + |q| \cdot \ln \frac{\mu}{\mu + |d|}$$

IDF**Length Norm.**

- ***BM25 (classic probabilistic model)***

$$\sum_{w \in q \cap d} \ln \frac{N - df(w) + 0.5}{df(w) + 0.5} \cdot \frac{(k_1 + 1) \times c(w, d)}{k_1((1 - b) + b \frac{|d|}{avdl}) + c(w, d)} \cdot \frac{(k_3 + 1) \times c(w, q)}{k_3 + c(w, q)}$$

- ***PL2 (divergence from randomness)***

$$\sum_{w \in q \cap d} c(w, q) \cdot \frac{tfn_w^d \cdot \log_2(tfn_w^d / \lambda_w) + \log_2 e \cdot \left(\frac{1}{\lambda_w} - \frac{tfn_w^d}{tfn_w^d + 1}\right)^2 + 0.5 \cdot \log_2(2\pi t^2 fn_w^d)}{tfn_w^d + 1}$$

Luhn's idea (1958): automatic indexing based on statistical analysis of text



Hans Peter Luhn
(IBM)

“It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.” (Luhn 58)

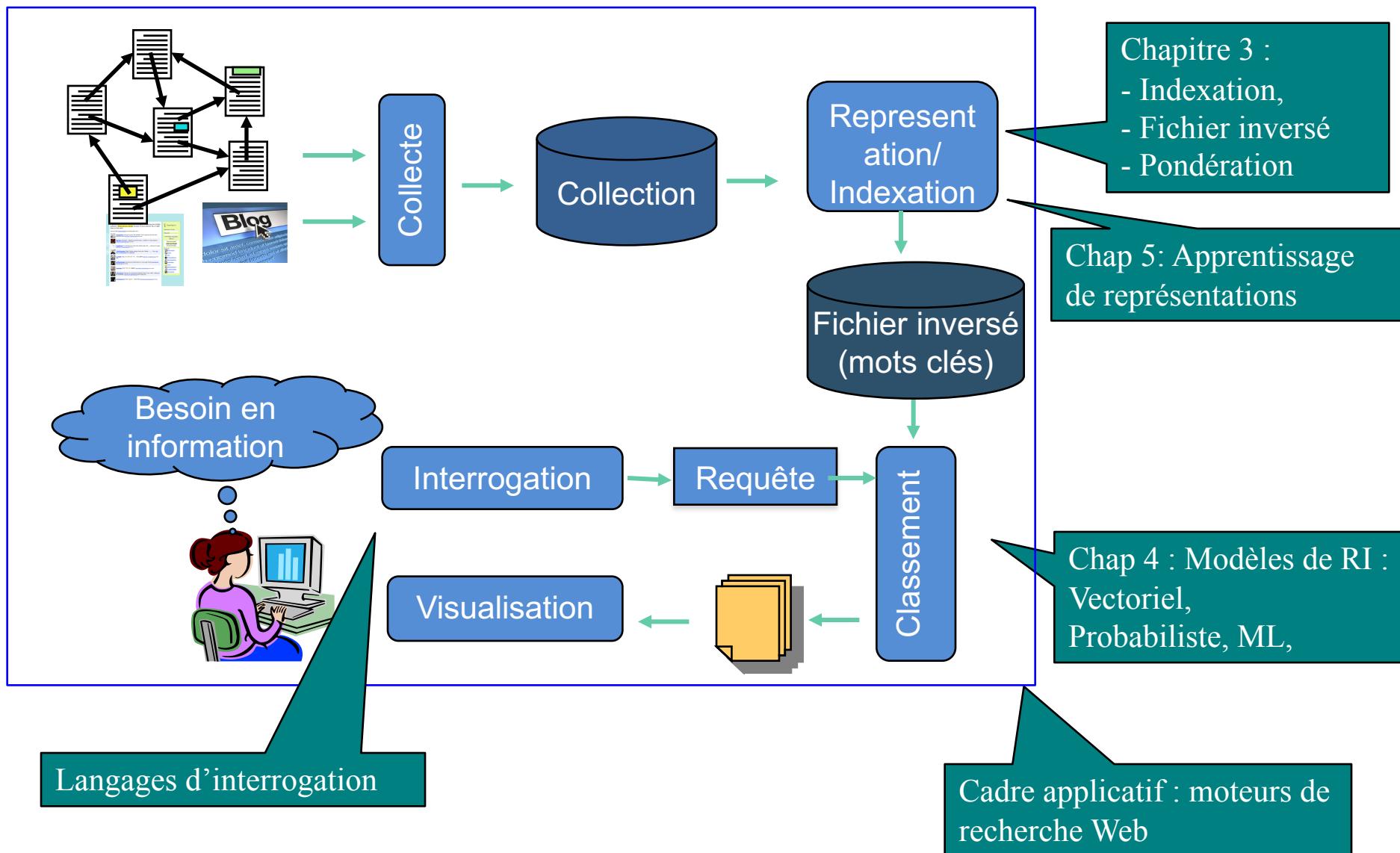
LUHN, H.P., 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development*, 1, 309-317 (1957).

LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2, 159-165 (1958).

- Théorie des ensembles :
 - Boolean model (± 1950)
- Algèbre
 - Vector space model (± 1970)
 - Spreading activation model (± 1989)
 - LSI (Latent semantic Indexing) (± 1994)
- Probabilité
 - Probabilistic model (± 1976)
 - Inference network model (± 1992)
 - Language model (± 1998)
 - DFR (Divergence from Randomness model) (± 2002)

- Evaluer l'utilité et l'intérêt d'un système (d'une technologie) de RI dans une application donnée
 - Mesurer son intérêt → satisfaction des ses utilisateurs
 - A travers des études utilisateur (user studies)
- Comparer les différents systèmes et méthodes (pour faire progresser l'état de l'art)
 - Effectué, généralement, à travers des collections de test (Benchmark) (jeu de test d'évaluation IR)

- Efficacité (Effectiveness) (Précision/Rappel) :
 - Mesure la capacité d'un système à classer les documents pertinents avant ceux qui sont non pertinents
- Efficience (Efficiency) (Temps et espace):
 - Rapidité avec laquelle l'utilisateur peut obtenir les résultats
 - Nombre de ressources de calcul nécessaires pour répondre à une requête?
 - Espace disque occupé
- Ergonomie
 - Facilité d'utilisation, de présentation des résultats
 - Faire des « user studies » (évaluation avec des utilisateurs réels)

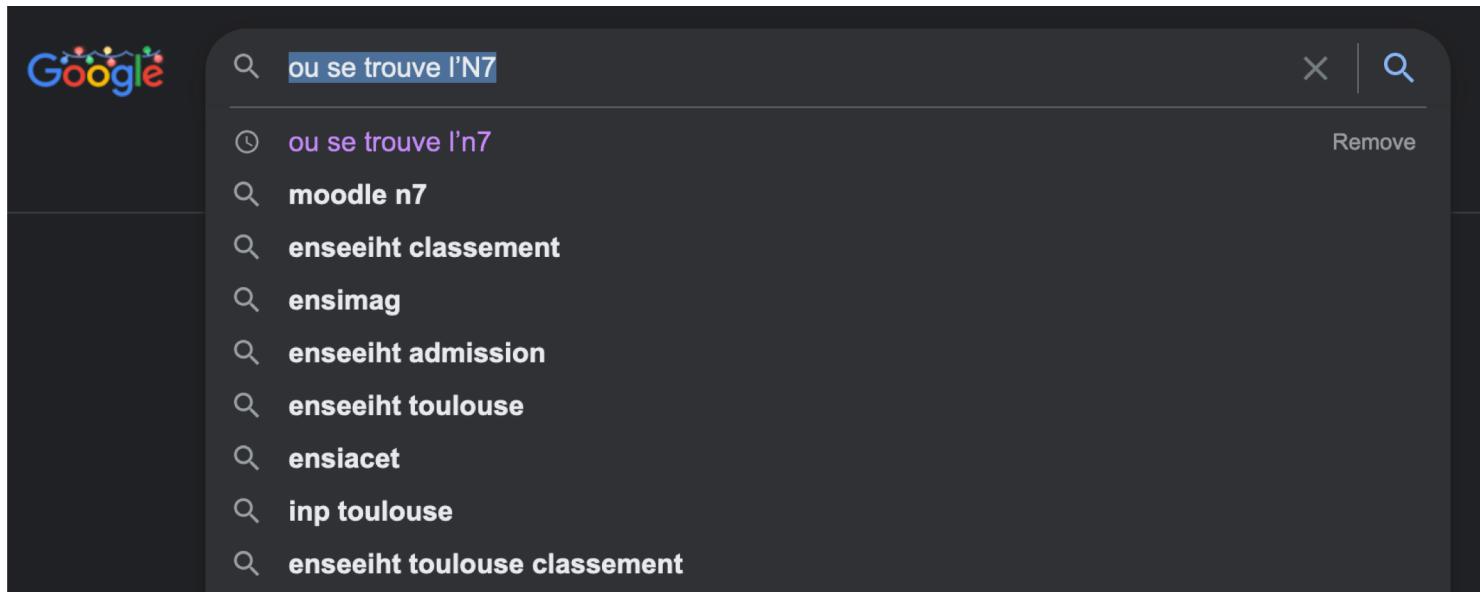


Langage d'interrogation

- Expression des besoins (Langage de requêtes)
 - Texte libre, Liste de mots clés
 - Avec / sans opérateurs booléens (AND, OR, NOT)
 - Images (...)
 - Aucun : navigation dans une liste de concepts (Yahoo,...)
- Requête
 - Liste de mots clés
- Ces deux notions sont souvent confondues



- Une requête «idéale» doit comporter tous les mots clés que l'utilisateur recherche → la similarité serait maximale
- Or, l' utilisateur recherche une information qu'il ne connaît pas à priori, il ne peut donc pas l' exprimer (décrire) de manière précise (idéale)
- Ce phénomène est qualifié par Belkin ASK “Anomalous State of Knowledge”



Fin