FACULTE DES SCIENCES DE SFAX

LES SYSTÈMES DE RECHERCHE D'INFORMATION : DIFFÉRENTS ACTEURS

Chapitre I

Master de Recherche en Informatique (MRI 2)

2020/2021

Hager KAMMOUN BOUZAIENE

Plan

- Historique
- Introduction
- Document
- Requête
- Evaluation
- Reformulation de requête

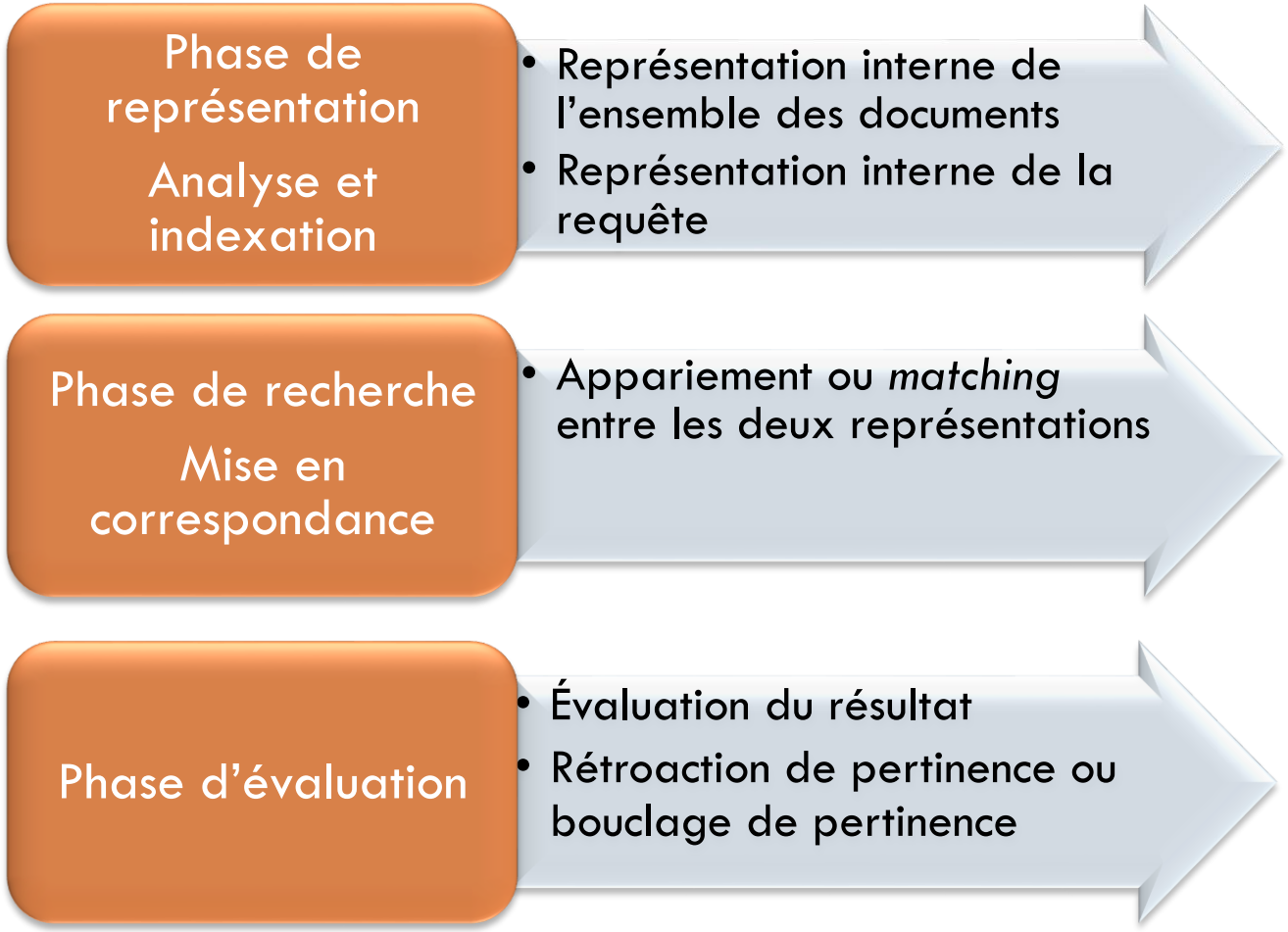
Historique

- Le domaine de la RI remonte au **début des années 1950** peu après l'invention des ordinateurs
- Les pionniers de l'époque étaient enthousiastes à utiliser l'ordinateur pour **automatiser la recherche des informations qui dépassaient la capacité humaine**
  - Explosion d'information après la guerre mondiale
- Recherche d'information ou « Information Retrieval » fut donné par **Calvin N. Mooers en 1948** pour la première fois dans le cadre de son mémoire de maîtrise
- La première conférence dédiée à ce thème **« International Conference on Scientific Information » s'est tenue en 1958 à Washington**
  - Les pionniers du domaine : Cyril Cleverdon, Brian Campbell Vickery, Peter Luhn, etc

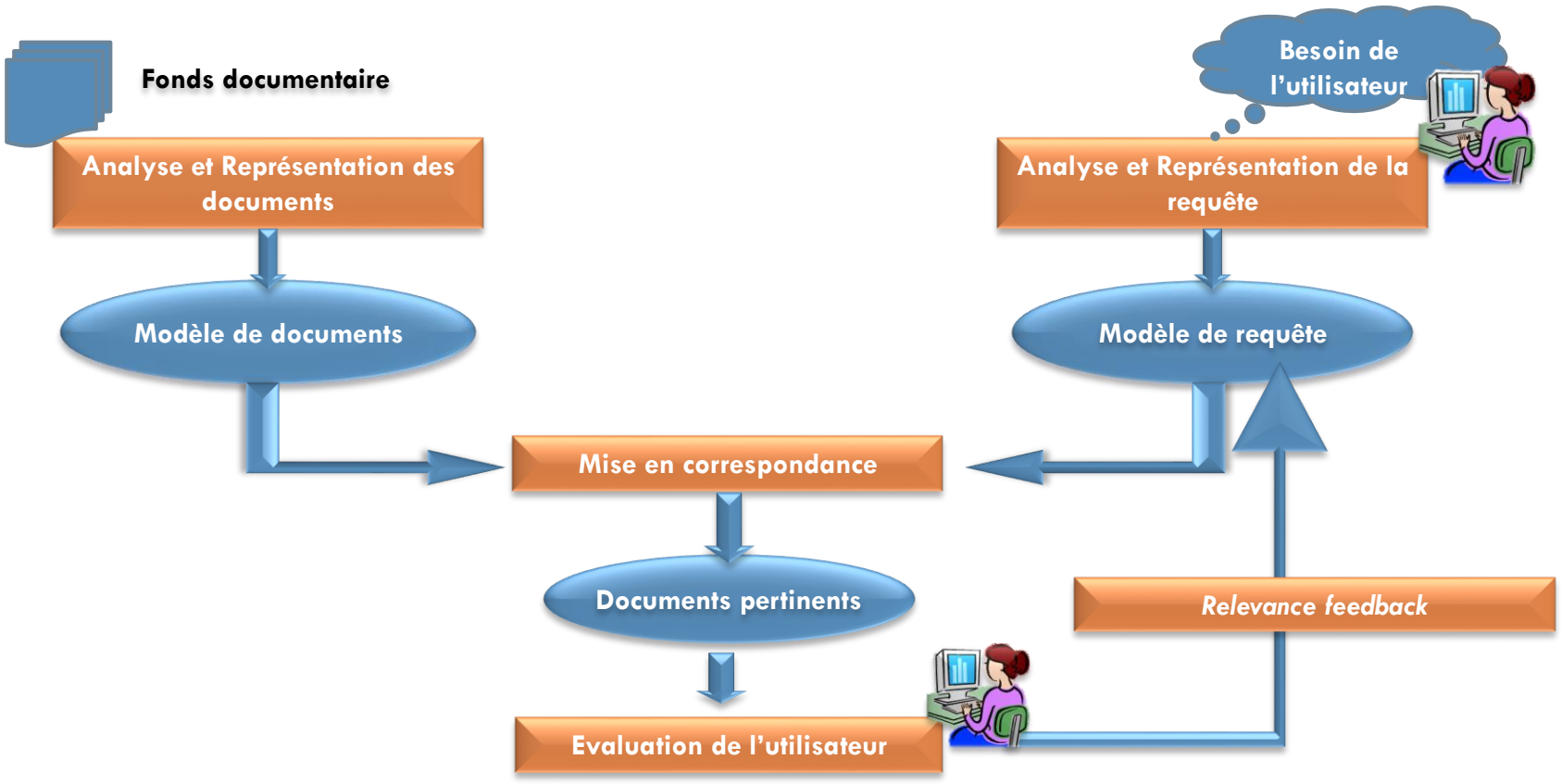
Introduction 1/6

- *Un SRI ne renseigne pas ses utilisateurs sur l'objet de leurs interrogations (ne fait pas évoluer leurs connaissances) mais indique simplement quels sont les documents (s'ils existent) en rapport avec leurs interrogations.* *(Lancaster 68)*
  - Restrictive mais correspond aux premiers modèles typiques des systèmes proposés
- Un SRI est censé retrouver les **documents pertinents** par rapport au **besoin d'information** d'un utilisateur exprimé à travers une **requête**
  - Compléter son état de connaissance par l'acquisition d'informations contenues dans des documents pertinents

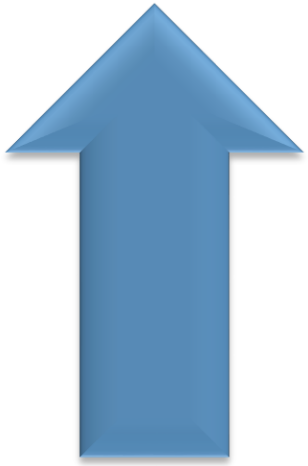
Introduction 2/6



Introduction 3/6



Introduction 4/6



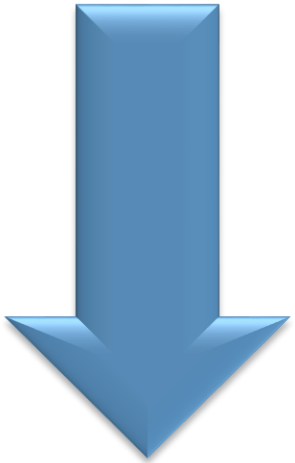
SGBD retrouver des faits précis en réponse à une requête ***Data Retrieval***

Données Structurées

Les sémantiques et les relations entre les objets sont connues

Langage de requêtes complexe

Le public est relativement ciblé



SRI interpréter un ensemble de connaissances pour déterminer les documents qui contiennent des informations pertinentes ***Information Retrieval***

Données partiellement structurées

La sémantique est mal définie

Usuellement le langage de requête est simple

Le public est plus large

Introduction 5/6

ISBN: 0-201-12227-8

Author: Salton, Gerard

Title: Automatic text processing: the transformation, analysis, and retrieval of information by computer

Editor: Addison-Wesley

Date: 1989

Content: <Text>

- Une partie structurée
  - Attributs externes

Recherche relativement simple

- Une partie (contenu) non structurée
  - Contenu

Recherche par le contenu pose beaucoup de problèmes (IR)

Intègre la recherche via les attributs externes

Introduction 6/6

- **Document**
  - Un document peut être un texte, un morceau de texte, une page Web, une image, une bande vidéo, …
  - Tout document qui peut constituer une réponse à une requête d'utilisateur

des documents textuels

- **Requête**
  - Exprime le besoin d'information d'un utilisateur (*Information need*)
- **Pertinence**
  - Le but de la RI est de trouver seulement les documents pertinents
  - Notion très complexe

Document

Analyse et indexation

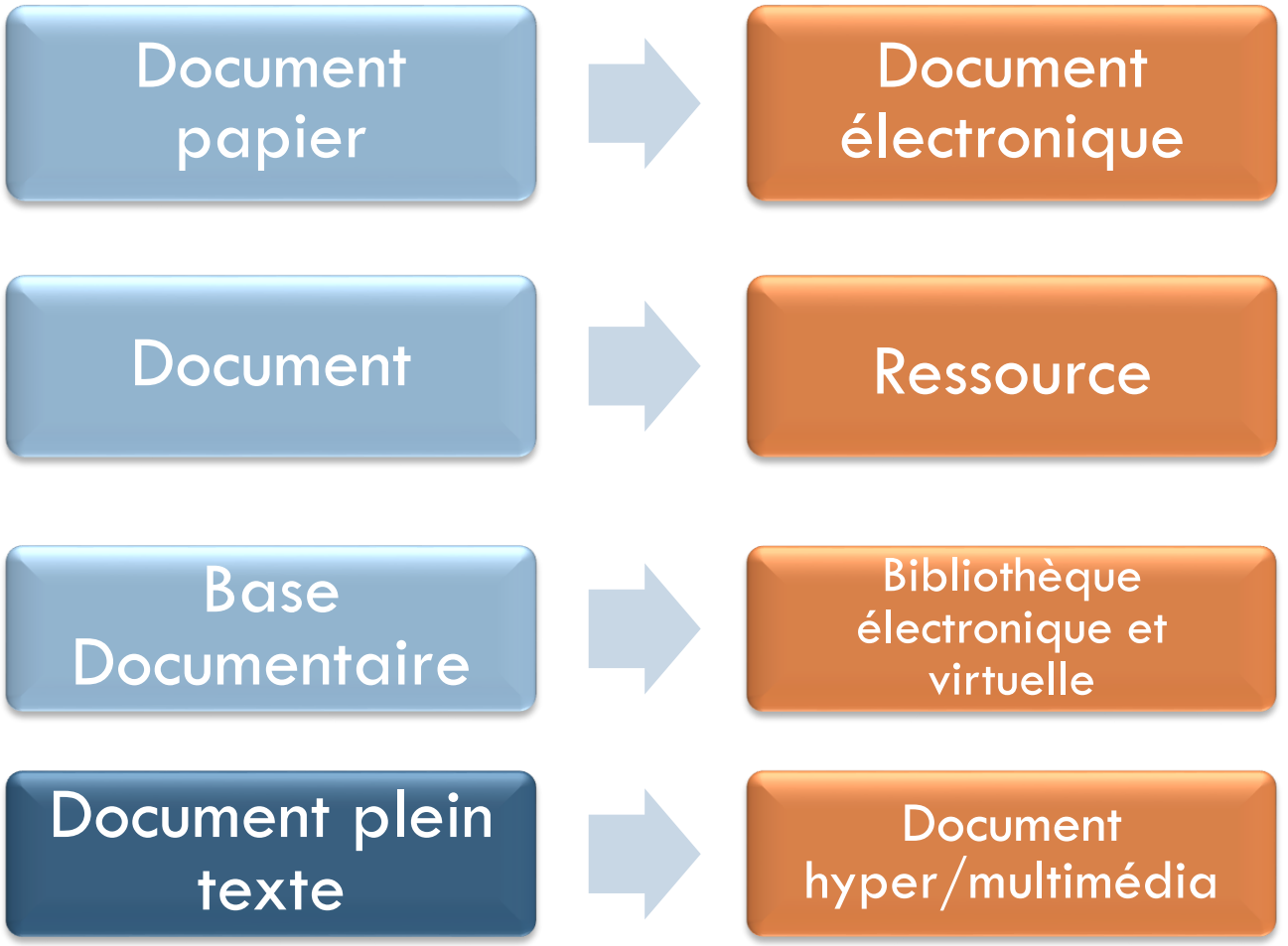
Approche linguistique

Approche statistique

Document 1/3

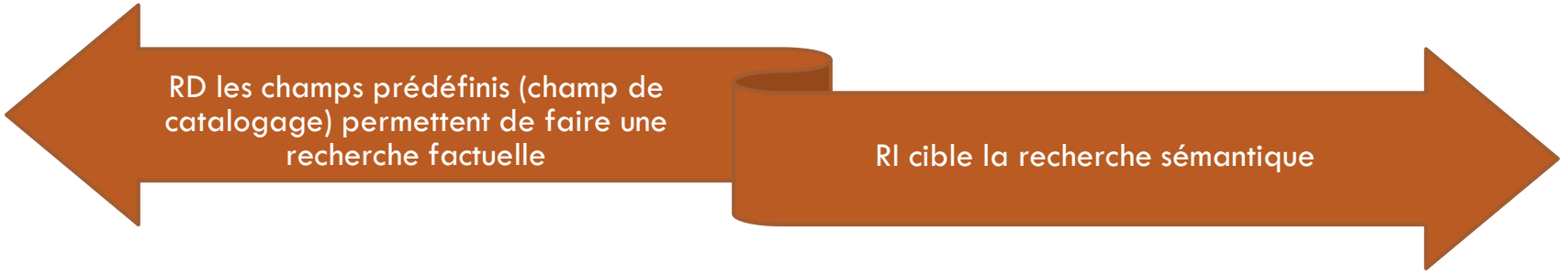
- Le document est l'élément central d'un SRI
- Le terme document est porteur du sens ***transmission de connaissances, de savoir et d'information***
  - *L'information contenant est le support de la connaissance contenue. Ainsi informer est une activité dans laquelle la connaissance intervient. Savoir est le résultat d'avoir été informé : de ce point de vue, il ressort que toute information représente de la connaissance (Blouch93)*

Document 2/3


Document papier
Document électronique
Document
Ressource
Base Documentaire
Bibliothèque électronique et virtuelle
Document plein texte
Document hyper/multimédia

Document 3/3

- Informations structurelles
  - Structure logique
    - pour synthétiser le contenu
    - identifier des concepts importants : titre, auteurs, résumé, …
- Informations factuelles
  - Liées aux valeurs données à certains attributs
    - date de publication, numéro de volume, …
- Informations sémantiques
  - En rapport directe avec les mots contenus dans un document
  - Contenu informationnel du document



Document : Analyse et indexation 1/5

- Analyse ou indexation
  - Extraire l'information utile
  - Ressortir les sujets spécifiques, les représentants ou termes descripteurs
- Vu par les spécialistes de l'informations et de la documentation
  - Indice de classification issu d'un langage classificatoire
    - Classification de la *Library of Congres* ou CDU
  - Une liste de descripteurs issu d'un langage documentaire (thesaurus : Mesh *Medical Subject Headings*)
    - Relation de hiérarchies spécifique générique
    - Relation de synonymie
    - Relation de voisinage

Un outil forgé pour la représentation de la base, il constitue une passerelle entre l'univers des documents et l'univers des utilisateurs de la base : langage pivot (Dachelet 90)

  - Une liste de mots-clés (langage libre ou langage naturel)
  - Un résumé de plusieurs lignes

Document : Analyse et indexation 2/5

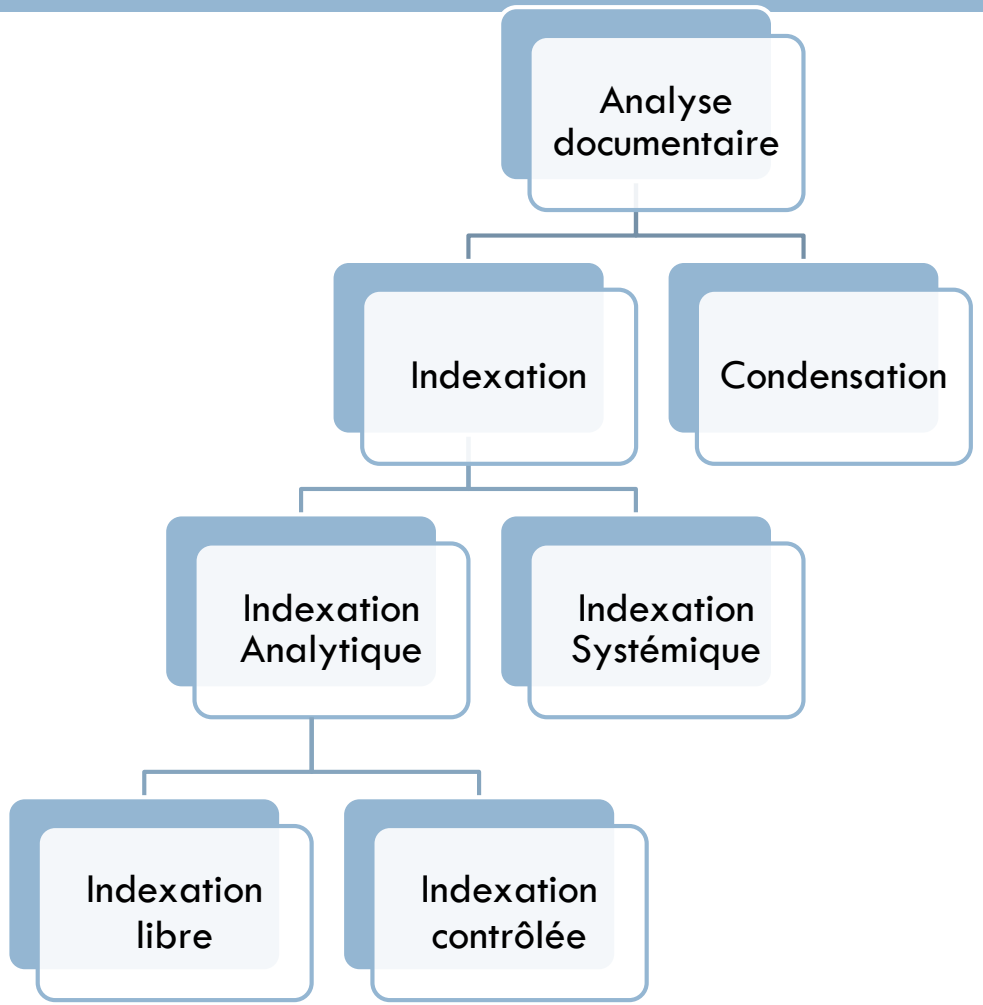
- Deux grands types d'analyse
  - Condensation : génération de résumé
  - Indexation
    - Par indices de classification : **systémique ou synthétique**
    - Par descripteurs, mots clés : **analytique**

Manuelle par un intermédiaire humain

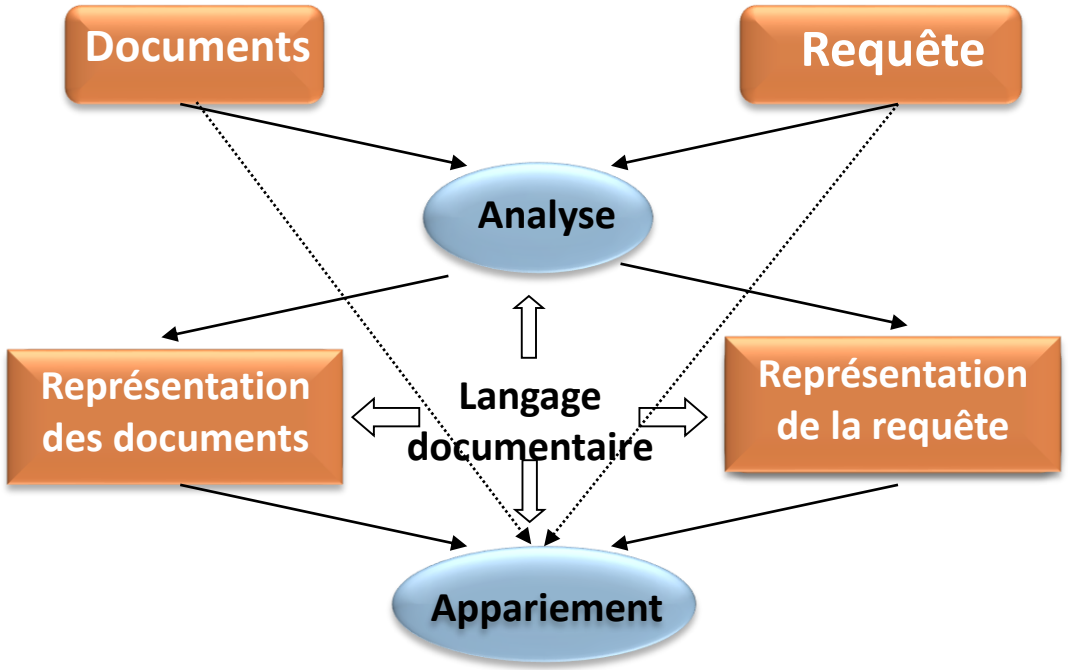
- Inconsistance : deux personnes peuvent indexer différemment
- Moins régulières
- Lente

Automatique

Document: Analyse et indexation 3/5



Document : Analyse et indexation 4/5



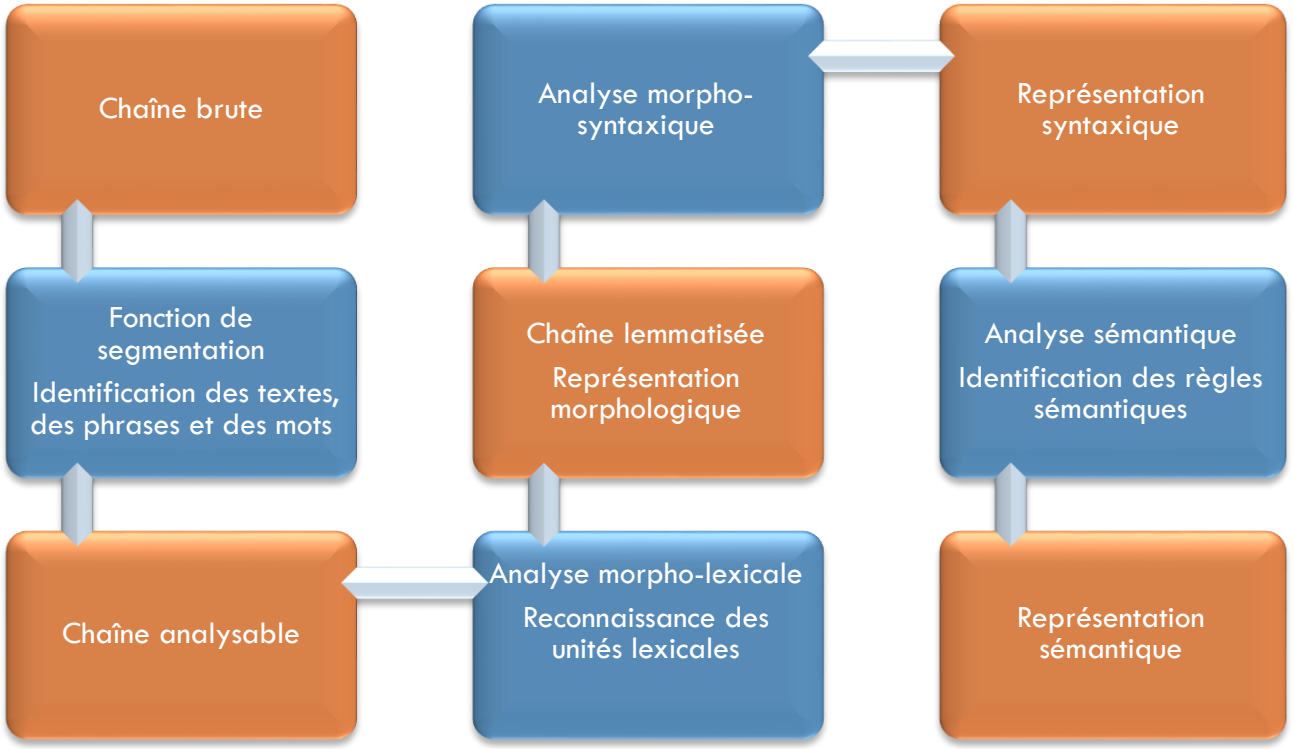
La plupart des SRI utilise une approche basée sur une phase d'analyse

Le but est de réduire le temps de rechecrhe

Document : Analyse et indexation 5/5

- Deux techniques d'indexation
  - Approche linguistique (TALN)
    - Nature linguistique du contenu à analyser
    - Le contenu est considéré comme un ensemble de mots qui entretiennent différents rapports entre eux
    - Permet d'affiner la recherche dans le cas des systèmes QR ou QA
  - Approche statistique
    - La fréquence d'apparition d'un terme dans un document est intimement liée à son importance
    - Le mot est une unité linguistique porteuse de sens

Document : techniques linguistiques



Les techniques de TALN font appel aux techniques de l'intelligence artificielle

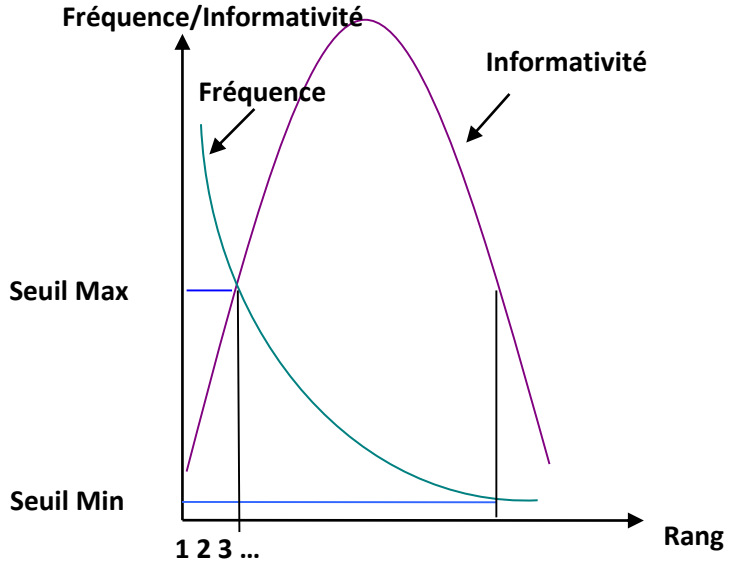
- **pour l'analyse des concepts**
- **pour la représentation et modélisation des connaissances : frames, réseaux sémantiques, scripts, etc.**

Document : techniques statistiques 1/6

- Calcul de fréquence
  - **Seuil minimum** : les termes dont la fréquence est supérieure à ce seuil sont considérés comme **pertinents**
  - **Seuil maximum** : les termes dont la fréquence est supérieure à ce seuil sont considérés comme **non pertinents**
  - **La loi de zipf (1949) stipule que le produit entre le rang d'un terme (son classement par ordre décroissant de fréquence) et sa fréquence est une constante**

- **L'informativité la plus importante est située entre les deux seuils**
- **L'informativité mesure la quantité de sens qu'un mot porte, n'est pas définie précisément dans la RI**
- **Pas de moyens théorique pour déterminer les deux fréquences pour lesquelles les termes sont éliminés**

Manière empirique



Document : techniques statistiques 2/6

- Calcul de tf*idf
  - Le tf *term frequency* mesure l'importance d'un terme pour un document

$$tf_{ik} = \frac{f_{ik}}{\sqrt{\sum_{j=1}^{n} (f_{ij})^2}}$$ Equ.1

$$tf_{ik} = 0.5 + 0.5\frac{f_{ik}}{f_i}$$ Equ.2

    - f_{ik} représente la fréquence d'occurrence d'un terme T_k dans le document D_i
    - tf_{ik} représente le poids de pertinence du terme T_k dans le document D_i
    - f_i représente le maximum des f_{ik} sur l'ensemble du document D_i
    - T représente l'univers des termes descripteurs des documents d'un corpus D et n = |T|
  - La somme du dénominateur (Equ.1) (facteur de normalisation) sert pour ne pas privilégier les documents longs par rapport à ceux qui sont courts mais qui sont pertinents par l'information qu'ils contiennent
  - La fréquence normalisée augmentée (Equ.2) permet de diminuer l'effet de la variation de la taille du document en attribuant les poids les plus importants aux descripteurs les plus fréquents dans le document

Document : techniques statistiques 3/6

- Le idf inverse *document frequency* mesure si le terme est discriminant c'est-à-dire non uniformément distribuée

$$idf_j = \log \frac{N}{N_j}$$

  - N correspond au nombre de documents dans tout le corpus
  - N_i correspond au nombre de documents indexés par le terme T_i
- Cette expression exprime le fait que l'importance d'un terme est inversement proportionnelle à sa distribution dans tout le corpus
  - Un descripteur T_k qui indexe tous les documents ne permet pas à un utilisateur de retrouver spécifiquement un document ce qui se traduit par une valeur nulle de son poids idf_k

Document : techniques statistiques 4/6

- Une formule tf*idf

$$w_{ik} = \frac{f_{ik} \log \frac{N}{N_k}}{\sqrt{\sum_{j=1}^{n} (f_{ij})^2 (\log \frac{N}{N_j})^2}} \qquad w_{ik} = (0.5 + 0.5 \frac{f_{ik}}{f_i}) \log \frac{N}{N_k}$$

- une valeur élevée de tf*idf implique que le terme est important dans le document et qu'il apparait peu dans les autres

Document : techniques statistiques 5/6

- L'utilisation de la formule tf*idf
  - donne de meilleurs résultats que l'utilisation de la fréquence d'occurrence seule
  - permet le filtrage et la pondération puisqu'il rend possible selon un certain seuil de choisir les termes
  - assure le compromis entre
    - Spécificité est proportionnelle au degré de précision des descripteurs à définir de manière spécifique certains documents
    - Exhaustivité est considérée comme proportionnelle au nombre de sujets couverts par les descripteurs des documents

Document : techniques statistiques 6/6

- Filtrage
  - Eliminer sur la base du calcul de fréquence
  - Eliminer certains mots qu'on qualifie de « mots vides » ou « *stop words* »
    - *Des prépositions, des prénoms, certains adverbes, …*
    - *stop list* ou anti-dictionnaire varie selon le domaine d'application
- Lemmatisation
  - Eliminer les différences non significatives pour réduire à une même forme
    - Des mots ayant des formes légèrement différents et un sens similaires : mots conjugués

•Une représentation du contenu du document « reflétant une partie du contenu » sous forme d'un ensemble de termes éventuellement pondérés
•Cette représentation dépend du modèle utilisé

Utilisateur, besoin d'information et requête

Utilisateur et besoin d'information

Requête

Visualisation de l'information

Utilisateur et besoin d'information 1/3

- L'utilisateur ne peut pas cerner son besoin par quelques mots normalisés dans une requête
  - Belkin (82) constate que l'utilisateur a recours à un SRI pour résoudre un problème ou satisfaire un intérêt particulier pour un sujet : un état anormal de connaissances (*Anomalous State of Knowledge ASK*)
    - Se faire aider par un professionnel
    - Consulter librement les bases de données bibliographiques
    - Naviguer dans les réseaux sans avoir une définition précise du problème
  - Effectue un processus ou cheminement pour exprimer sa recherche
  - Il est difficile sinon impossible de formuler une requête qui décrit complètement et précisément un besoin d'information

Utilisateur et besoin d'information 2/3

- Pour satisfaire au mieux l'utilisateur, il est essentiel de comprendre ses mécanismes cognitifs pour le modéliser dans un processus de recherche d'information
  - Usager pour dire utilisateur
  - Modèles quantitatifs qui modélisent le comportement externe de l'usager
  - Modèles analytiques qui modélisent le comportement interne de l'usager : connaissances, processus cognitif, etc.
    - Domaine de la psychologie cognitive
  - Modéliser l'utilisateur avec les paramètres suivants
    - USER statut de l'usager
    - UGAOL buts de l'usager (préférences ou stratégies de recherche)
    - KNOWN le niveau d'expertise ou niveau de connaissances de l'usager dans le domaine
    - IRS la familiarité de l'usager avec les système documentaires
    - BACK l'expérience de l'usager vis-à-vis du système concerné

Utilisateur et besoin d'information 3/3

- Quatre catégories ou stratégies de recherche ont été identifiées
  - Une demande précise : l'utilisateur sait exactement ce qu'il cherche
  - Une demande thématique : l'utilisateur cherche à explorer le corpus sur un thème particulier
  - Une demande connotative : l'utilisateur cherche une image par l'expression d'un visage ou par métaphore dans le contexte de la recherche textuelle
  - Une demande exploratoire : l'utilisateur veut se faire une idée sur le contenu du corpus et c'est après une consultation préalable que seront définies plus précisément ses besoins

Requête 1/2

- Une requête peut être exprimée de différentes manières
  - Sous forme de grille ou formulaire sur les champs de catalogages ou issus d'une structure logique
  - En langage naturel en utilisant des mots non contrôlés
  - En utilisant des phrases courtes en langage naturel
  - Sous forme de texte ou documents : requête par l'exemple QBE *query by example*
- Interrogation basée sur la visualisation globale de l'ensemble des documents intégrant des outils permettant d'exploiter l'ensemble en utilisant une approche classificatoire ou la navigation à travers une carte : classes et relations

Requête 2/2

- Requête booléenne
  - Exprimée à travers des termes connectés par des opérateurs booléens (et, ou, sauf)
- Requête vectorielle
  - Exprimée à travers des termes pondérés
- Possibilité de les sauvegarder pour de la diffusion sélective ou diffusion ciblée
  - Profil requête
    - Scruter systématiquement et en temps réel les informations nouvelles entrée dans la base

Visualisation de l'information

- La restitution des documents en réponse à une requête
  - Liste de titres ou de passages qui contiennent les termes de la requête
    - Il n'est pas envisageable de présenter le document dans son intégralité sauf s'il est suffisamment court
    - Il est possible de proposer des résumés automatiques
      - en attribuant une importance aux phrases qui contiennent les termes de la requête
  - Classement par ordre de pertinence décroissante par rapport à la requête
  - Représentation graphique individuelle
    - Représenter les documents et éventuellement les liens qui existent entre eux
    - Peu intéressante quand le corpus est de taille importante
  - Représentation graphique globale
    - Issue des méthodes de classification notamment des cartes auto-organisatrices de Kohonen

Evaluation

Paradigme système

Paradigme usager

Pertinence

Rappel et Précision

Corpus de test

Evaluation : paradigme système

- Paradigme système *matching paradigm*
  - Pertinence objective point de vue du système
  - Pertinence système ou *relevance*
    - les fonctions de traitement de l'information
    - les langages d'indexation pour représenter les documents
    - les langages d'interrogation pour formuler les requêtes
    - les algorithmes de mise en correspondance
  - Les éléments qui peuvent être analysés
    - la collection : son contenu et son étendu
    - sa description : la modélisation ou la représentation des documents
    - le type de recherche possible : les moyens d'accès, l'indexation, l'extraction des documents, l'analyse et la synthèse des informations du documents pour répondre aux questions
    - la présentation des résultats

Evaluation : paradigme usager

- Paradigme usager
  - Pertinence subjective point de vue de l'usager
  - Pertinence utilisateur ou Pertinence
    - Centrée sur l'usager, son besoin et son environnement
    - Le jugement de pertinence est lié à plusieurs paramètres : contexte de recherche, niveau d'expérience, niveau de connaissance, etc.

La pertinence est fonction de la qualité de l'information, elle est toujours liée à un utilisateur alors que la quantité d'information ne l'est pas

Evaluation : pertinence 1/2

- Le besoin d'information est transformé en une requête puis en une représentation de cette dernière
- Le document subit des transformations pour aboutir à une représentation de ce dernier
- Un bon système doit pouvoir effectuer une mise en correspondance entre les deux représentations qui reflètent bien la pertinence système qui à son tour correspond bien au jugement de pertinence de l'utilisateur
- Étant donné la différence entre le niveau utilisateur, le niveau système et le niveau interne du système il y a nécessairement une dégradation des représentations et donc du jugement

Evaluation : pertinence 2/2

Besoin d'information

Pertinence de l'utilisateur

Documents pertinents

requête

Pertinence du système

Corpus documentaire

Représentation de la requête

Correspondance

Représentation du document par rapport au corpus documentaire

- Pour déterminer si la représentation d'un document correspond à celle de la requête on doit développer un processus d'évaluation
- Les méthodes d'évaluation sont en relation avec la représentation de documents et de requête

Evaluation : rappel et précision 1/3

- Mesurer la qualité de service rendue par un SRI, revient à mesurer son efficacité au niveau de la recherche (quantifier la qualité)
- Mesurer la valeur de l'information contenue dans les documents réponses
  - Rappel (TR)
    - Mesure la capacité d'un système à sélectionner tous les documents pertinents de la collection
    - Correspond à la proportion des documents pertinents retrouvés par rapport à tous les documents pertinents du corpus
    - La proportion complémentaire est le silence qui correspond à la proportion des documents pertinents non retrouvés
  - Précision (TP)
    - Mesure la capacité d'un système à ne sélectionner que les documents pertinents
    - Correspond à la proportion des documents pertinents retrouvés par rapport à tous les documents retrouvés par le système
    - La proportion complémentaire est le bruit qui correspond à la proportion des documents retrouvés qui ne sont pas pertinents

Evaluation : rappel et précision 2/3

- Rappel
- Précision

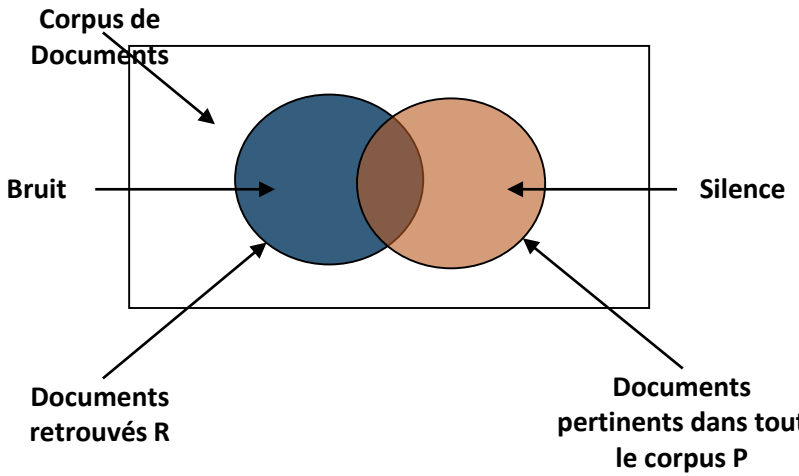
$$(TR) = \frac{|P \cap R|}{|P|}$$

$$(TP) = \frac{|P \cap R|}{|R|}$$

- P correspond à l'ensemble des documents pertinents par rapport à une requête
- R correspond à l'ensemble des documents retrouvés en réponse à une requête

- F-mesure constitue un compromis entre le rappel et la précision

$$F = \frac{2(TR \times TP)}{(TR + TP)}$$



Evaluation : rappel et précision 3/3

- Un système qui donne TP=1 et TR=1 signifie qu'il restitue tous les documents pertinents et rien que les documents pertinents
- Un SRI qui a tendance à restituer un grand nombre de documents
  - A une grande probabilité de retrouver les documents pertinents TR=1
  - A une grande probabilité de retrouver des documents non pertinents TP=0
- Un SRI qui a tendance à restituer un petit nombre de documents
  - A une faible probabilité de retrouver des documents non pertinents TP=1
  - A une faible probabilité de retrouver tous les documents pertinents TR=0
- Faire varier le nombre de documents restitués
  - Tracer la variation de TP en fonction de TR

Evaluation : corpus de test 1/2

- Pour pouvoir comparer des SRI entre eux des corpus textuels de test existent
  - Requêtes et documents pertinents correspondants sont connus a priori
  - Pour qu'un corpus de test soit significatif il faut qu'il possède un nombre de documents assez important
  - Plus le nombre de requêtes est grand plus l'évaluation des résultats est significative

Un système dont la courbe dépasse (se situe en haut à droite) celle d'un autre est considéré comme un meilleur système

Si les deux courbes se croisent on utilise la précision moyenne une mesure de performance

Evaluation : corpus de test 2/2

	Nombre de documents	Nombre de requêtes
CACM	3240	64
CISI	1460	112
CRAN	1400	225
MED	1033	30
TIME	425	83

- La collection CACM regroupe les titres et es résumés tirés du journal CACM.
- La collection Cranfiel traite des résumés du domaine "*aeronautical engineering*".
- La collection Medline est constituée d'articles tirés du journal "*medical journal*".
- La collection Time est constituée d'articles tirés du journal *Time*.

Les corpus plus récents (TREC : text retrieval conference) :

contiennent 100 000 (taille moyenne) voir des millions de documents (grande taille)

TREC est une rencontre scientifique qui fournit une base d'évaluation et de comparaison des moteurs de recherche du marché, elle permet aux participants de tester leurs systèmes sur des collections textuelles communes et offre un cadre et une méthodologie d'évalauation

Reformulation de requête

Rétroaction de pertinence

Expansion de requête

Reformulation de requête 1/4

- Une recherche donne rarement des résultats pertinents et complets du premier coup
  - Une recherche complète se présente comme un processus itératif ou incrémental mettant en œuvre plusieurs requêtes qui se succèdent et qui permettent d'affiner progressivement les réponses données par le système

- **Les techniques de reformulation consistent à modifier les requêtes pour ressembler davantage aux documents jugés pertinents et s'éloigner des documents non pertinents**
- **Plus la distance entre la requête initiale et la requête reformulée est grande plus il y a de nouveaux documents qui vont apparaître comme résultats de la nouvelle requête**
  - **Peuvent être assistées par l'utilisateur ou interactive : réinjection de requête rétroaction de pertinence ou *relevance feedback***
  - **Peuvent être automatiques : expansion de requête**

Reformulation de requête 2/4

- Rétroaction de pertinence
  - Repondération des termes ou par l'ajout ou le retrait des termes contenus dans les documents pertinents/non pertinents à la requête
  - Techniques qui permettent d'assister l'utilisateur à travers des vues synthétiques

$$Q^{new} = \alpha Q^{old} + \beta \frac{1}{|reldocs|} \sum_{reldocs} w_{t_i} - \gamma \frac{1}{|nonreldocs|} \sum_{nonreldocs} w_{t_i}$$

α permet de moduler l'importance de la requête précédente.
β permet de moduler le vecteur profil moyen des documents choisis.
γ permet de moduler le vecteur profil des documents rejetés.
α , β, γ sont des paramètres positifs $\in$[0,1]
Le paramètre α n'était pas initialement pris en compte dans la formule de Rocchio. Salton, l'a introduit ultérieurement et c'est la forme générale définie qui est souvent considérée.

Reformulation de requête 3/4

- Cette fonction (Rocchio) dérive de l'hypothèse qu'une requête idéale Q^{new} doit maximiser la différence de sa distance cosinus moyenne de ses documents pertinents et de sa distance cosinus moyenne de ses documents non pertinents

- Expansion de requête
  - Les termes peuvent être déduits d'un dictionnaire de synonymes, d'un thesaurus ou d'une étude statistique sur la co-occurrence de termes
    - Plus deux termes sont co-occurrents dans des documents plus ils sont fortement reliés
  - Les termes peuvent être déduits des documents jugés pertinents par le système (pertinence système)

Reformulation de requête 4/4

- Problèmes posés par la reformulation de requêtes
  - Peuvent dépondre du nombre de termes ajoutés, de leur sélection et de la manière avec laquelle ils sont ajoutés
  - Rétroaction de pertinence
    - d'un emploi lourd à l'utilisateur qui doit dialoguer avec le système
  - Expansion de requête
    - Les termes ajoutés ne sont pas toujours appropriés et peuvent par conséquent engendrer du bruit puisqu'on peut ajouter des termes qui ne sont pas en rapport le besoin de l'utilisateur