* 1. Introduction

La Recherche d’Information (RI) peut être définie comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défit est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l’attente de l’utilisateur. L’opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d’Information (SRI), ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l’utilisateur (requête) avec une représentation du contenu des documents (fiche ou enregistrement) au moyen d’une fonction de comparaison (ou de correspondance). L’essor du web a remis la RI face à de nouveaux défis d’accès à l’information, il s’agit cette fois de retrouver une information pertinente dans un espace diversifié et de taille considérable.

Ce chapitre a pour but de présenter le domaine de la RI. Dans la première partie, nous présentons les concepts de base de la RI. En particulier, nous décrivons les notions de document, de requête et de pertinence ; les processus d’indexation, de recherche et de reformulation de requêtes ; ainsi que, les modèles de RI. Dans la seconde partie est présentée l’évaluation des processus de systèmes de recherche d’information.

* 1. La recherche d’information

La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d’établir des représentations des documents dans le but d'en récupérer des informations à travers la construction d’index. L’informatique a permis le développement d’outils pour traiter l’information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l’information. On peut aujourd'hui dire que la recherche d'information est un champ transdisciplinaire qui peut être étudié par plusieurs disciplines utilisant des approches qui devraient permettre de trouver des solutions pour améliorer son efficacité.

* + 1. Définitions
* *Définition 1* : La recherche d’information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations **[1].**
* *Définition 2* : La recherche d’information est une branche de l’informatique qui s’intéresse à l’acquisition, l’organisation, le stockage, la recherche et la sélection d’information **[2].**
* *Définition 3* : La recherche d’information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l’accès à l’information pertinente pour un utilisateur ayant un besoin en information **[3].**

Toutes ces définition partagent l’idée que la RI à pour objet d’extraire d’un document ou d’un ensemble de documents les informations pertinentes qui reflètent un besoin d’information.

* + 1. Concepts de base de la RI

Le rôle d’un Système de Recherche d’Information (SRI) est de mettre en œuvre des techniques et des moyens permettant de retourner les documents pertinents d’une collection en réponse à un besoin en information d’un utilisateur, exprimée par un langage de requêtes qui peut être le langage naturel, une liste de mots clés ou un langage booléen .

Afin d’atteindre cet objectif, un processus d’indexation des documents de la collection est effectué. Il permet de construire une représentation synthétique des documents, appelée index. Lorsque l’utilisateur formule sa requête un processus similaire est effectué sur la requête. Il consiste à analyser la requête et établir une représentation interne. Puis, le système établit une correspondance entre la représentation de la requête et la représentation des documents (index) pour sélectionner et présenter les documents qui répondent le mieux au besoin de l’utilisateur (les documents pertinents). Le SRI s'appuie sur des modèles de RI pour établir cette correspondance entre les documents et la requête. L’architecture générale d’un SRI illustrée par la Figure I.1 fait ressortir des éléments constitutifs tels que : le document, le besoin en information, la requête et la pertinence, ainsi que trois principales fonctionnalités : l’indexation, la recherche et la reformulation de la requête. Dans ce qui suit, nous détaillons ces éléments et ces processus.

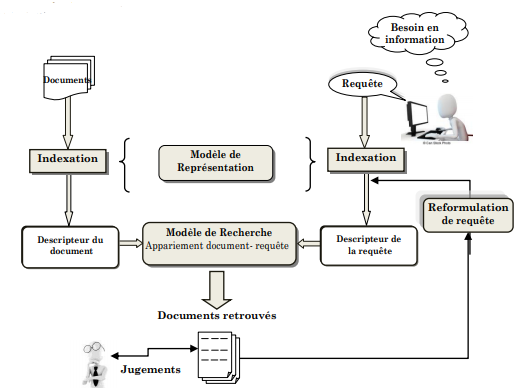


Figure I‑1 : Architecture générale d’un SRI.

**Collection de documents** : la collection de documents (ou fond documentaire) constitue l'ensemble des informations exploitables et accessibles. Elle est constituée d'un ensemble de documents. Dans le cas général et pour un souci d'optimalité, la base constitue des représentations simplifiées mais suffisantes pour ces documents. Ces représentations sont étudiées de telles sortes que la gestion (ajout suppression d'un document) ou l'interrogation (recherche) de la base se font dans les meilleures conditions de coût **[4].**

**Document** : Un document est un élément essentiel dans un SRI. Dans son acceptation courante, l’une des définitions possibles du terme document est de le considérer comme un support physique de l’information, qui peut être du texte, une page web, une image, une séquence vidéo, etc. Dans le cas d’un document texte on peut le représenter selon trois vues :

* **La vue sémantique (ou contenu)** : elle se concentre sur l’information véhiculée dans le document.
* **La vue logique** : elle définit la structure logique du document (structuration en chapitres, sections).
* **La vue présentation** : elle consiste en la présentation sur un médium à deux dimensions (alignement de paragraphes, indentation, en-têtes et pieds de pages, etc.). **[5]**

**Besoin d’informations**: la notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. Trois types de besoin utilisateur ont été définis :

* **Besoin vérificatif** : l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un autre exemple serait de chercher la date de publication d'un ouvrage dont la référence est connue. Un besoin de type vérificatif est dit stable, c'est-à dire qu'il ne change pas au cours de la recherche.
* **Besoin thématique connu** : l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et un domaine connu. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble. C'est ce qu'on appelle dans la littérature le label.
* **Besoin thématique inconnu** : cette fois, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations en dehors des sujets ou des domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.

**Requête** : la requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l’interface entre le SRI et l’utilisateur. Divers types de langages d’interrogation sont proposés dans la littérature. Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique **[4].**

**Pertinence** : La pertinence est une notion fondamentale et cruciale dans le domaine de la RI. Cependant, la définition de cette notion complexe n’est pas simple, car elle fait intervenir plusieurs notions. Basiquement, elle peut être définie comme la correspondance entre un document et une requête ou encore comme une mesure d’informativité du document à la requête. Essentiellement, deux types de pertinence sont définis : la pertinence système et la pertinence utilisateur **[5].**

* **La pertinence Système** est souvent présentée par un score attribué par le SRI afin dévaluer l’adéquation du contenu des documents vis-à-vis de celui de la requête. Ce type de pertinence est objectif et déterministe.
* **Pertinence utilisateur** quant à elle, se traduit par les jugements de pertinence de l’utilisateur sur les documents fournis par le SRI en réponse à une requête. La pertinence utilisateurest subjective, car pour un même document retourné en réponse à une même requête, il peut être jugé différemment par deux utilisateurs distincts (qui ont des centres d’intérêt différents). De plus, cette pertinence est évolutive, un document jugé non pertinent à l’instant t pour une requête peut être jugé pertinent à l’instant t+1, car la connaissance de l’utilisateur sur le sujet a évolué **[5].**
  1. Processus d’indexation

Pour que la recherche d’information se réalise avec des coûts acceptables, il convient d’effectuer une opération fondamentale sur les documents de la collection. Cette opération est nommée indexation **[6][7]** Elle consiste à associer à chaque document une liste de mots clés appelée aussi descripteur, susceptible de représenter au mieux le contenu sémantique des documents. La finalité de l’indexation est donc de produire une représentation synthétique des documents, formé de termes, ces termes peuvent être extraits de trois manières :

* ***Manuelle*** : chaque document de la collection est analysé par un spécialiste du domaine ou un documentaliste. L’indexation manuelle assure une meilleure précision dans les documents restitués par le SRI en réponse aux requêtes des utilisateurs **[8]**.Néanmoins, cette indexation présente un certain nombre d’inconvénients liés notamment à l’effort et le prix qu’elle exige (en temps et en nombres de personnes). De plus, cette indexation est subjective, qui est liée au facteur humain, différents spécialistes peuvent indexer un document avec des termes différents. Il se peut même arriver qu’un spécialiste indexe différemment un document, à différents moments.
* ***Semi-automatique*** : la tâche d’indexation est réalisée ici conjointement par un programme informatique et un spécialiste du domaine **[9]**. Le choix final des descripteurs revient à l’indexeur humain. Dans ce type d’indexation un langage d’indexation contrôlé est généralement utilisé.
* ***Automatique*** : dans ce cas, l’indexation est entièrement automatisée. Elle est réalisée par un programme informatique et elle passe par un ensemble d’étapes pour créer d’une façon automatique l’index. Ces étapes sont : l’analyse lexicale, l’élimination des mots vides, la normalisation (lemmatisation ou radicalisation), la sélection des descripteurs, le calcul de statistiques sur les descripteurs et les documents (fréquence d’apparition d’un descripteur dans un document et dans la collection, la taille de chaque document, etc.) et enfin la création de l’index et éventuellement sa compression. Nous détaillons ces différentes étapes ci-dessous.

Bien que l'indexation se base sur des techniques relativement établies, il peut y avoir plusieurs indexations différentes d'un même texte, aussi valables les unes que les autres, en fonction de l'usage qui doit en être fait et du public auquel elles s'adressent.

Les différentes étapes de l’indexation sont schématisées comme suit :

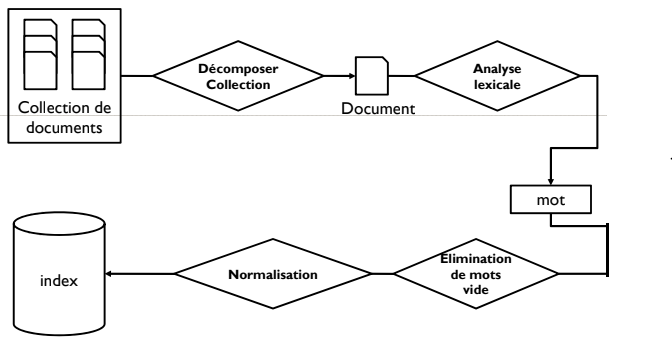


Figure I‑2 : Indexation d'un document.

* + 1. L’analyse lexicale

Elle permet de convertir un texte de document en une liste de termes. Un terme est un groupe de caractères constituant un mot significatif **[10]**. L’analyse lexicale permet de reconnaître les espaces de séparation des mots, les chiffres, les ponctuations, etc.

* + 1. L’élimination 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 (aussi appelée anti-dictionnaire ou stop-list),
* 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. Cependant, elle peut réduire le taux de rappel, en réponse à des requêtes bien spécifiques (par exemple, la requête be or not be).

* + 1. La normalisation

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 **[11].**

L’inconvénient majeur de cette opération est qu’elle supprime dans certains cas la sémantique des termes originaux, c’est le cas par exemple des termes derivate/derive, activate/active, normalisés par l’algorithme de Porter.

* + 1. Le choix de descripteurs

Elle consiste à déterminer le type d’unités élémentaires pour représenter les documents. On parle aussi de descripteur. L’objectif est d’avoir une représentation des documents permettant une moindre perte d’information sémantique possible. On distingue plusieurs types de descripteurs **[12].**

* **Les mots simples** : les mots simples du texte de document en éliminant les mots vides,
* **Les lemmes** ou les racines des mots extraits.
* **Les N-grammes** : qui sont une représentation originale d'un texte en séquence de N caractères consécutifs. On trouve des utilisations de bi-grammes et trigrammes dans la recherche d’information.
* **Les mots composés** : groupes de mots ou expression (phrase en anglais) sont souvent plus riches sémantiquement que les mots qui les composent pris séparément. Par exemple, le mot composé "imprimante laser" est plus précis que "imprimante" et "laser" pris isolément. Cet argument a conduit à leur large utilisation en RI.
* **Les concepts** : qui sont des expressions pris généralement d’une structure conceptuelle, tels que les thésaurus ou les ontologies.
  + 1. La création de l’index

Au terme du processus d’indexation, un ensemble de structure de données sont crées. Ces dernières permettent un accès efficace à la représentation des documents. Le fichier inverse est la structure de données la plus utilisée **[6][13]** il enregistre pour chaque descripteur les identificateurs des documents qui le contiennent et sa fréquence dans chacun de ces documents.

Généralement, les structures de données sont compressées avant d’être enregistrées sur le disque, ce qui permet de réduire la taille de l’index. Parmi les méthodes de compression utilisées on peut citer la méthode Elias Gamma **[14]** qui opère au niveau bit requérant ainsi beaucoup d’opérations pour la compression et la décompression **[15].**

* 1. Appariement document requête

La fonction d’appariement document-requête permet de mesurer la valeur de pertinence d’un document vis-à-vis d’une requête. Afin de réaliser cela, le SRI représente le document et la requête avec un même formalisme, puis le SRI compare les deux représentations. Le résultat de cette comparaison se traduit par un score qui détermine la probabilité de pertinence (degré de similarité ou degré de ressemblance) du document vis-à-vis de la requête. Cette fonction d’appariement est notée *RSV* (*d,q*) (Retrieval Statut Value), où *d* représente un document de la collection et *q* la requête. Cette valeur permet ensuite au SRI d’ordonner les documents renvoyés à l’utilisateur.

* 1. Les modèles de recherche d’information

L’étape d’appariement, décrite précédemment, repose sur des modèles de RI dont l’objectif est de fournir une formalisation du processus de RI et un cadre théorique pour la modélisation de la mesure de pertinence. Il existe un grand nombre de modèles de RI textuelle développé dans la littérature. Ils reposent sur l’utilisation de la logique, l’algèbre, la théorie de la probabilité et les statistiques. Comme le montre la figure I.3, on peut distinguer trois grandes classes de modèles, regroupés selon les fondements mathématiques sur lesquels ils se basent **[16].**

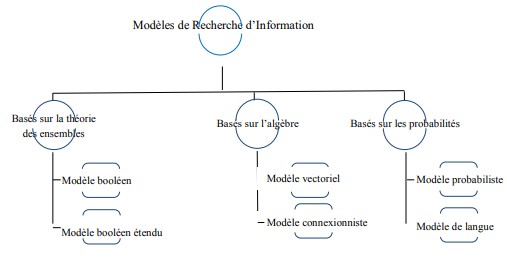


Figure I‑3 : Taxonomie des modèles de RI [Baeza-Yates et al.1999].

Ces modèles ont en commun le vocabulaire d’indexation basé sur le formalisme mots clés et diffèrent principalement par le modèle d’appariement requête-document. Le vocabulaire d’indexation V = {ti}, i ∈ {1, ..., n} est constitué de n mots ou racines de mots qui apparaissent dans les documents. Le modèle de RI est défini par un quadruplet (D, Q, F, R(q,d)) : où

* D est l’ensemble de documents
* Q est l’ensemble de requêtes
* F est le schéma du modèle théorique de représentation des documents et des requêtes
* R(q,d) est la fonction de pertinence du document d à la requête q.

Nous présentons dans la suite les principaux modèles de RI : le modèle booléen, le modèle vectoriel et le modèle probabiliste.

* + 1. Le modèle booléen

Le modèle booléen est basé sur la théorie des ensembles. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Chaque document est représenté par une conjonction logique des termes non pondérés qui constitue l’index du document. Un exemple de représentation d’un document est comme suit : d = t1 ∧ t2 ∧ t3... ∧ tn.

Une requête est une expression booléenne dont les termes sont reliés par des opérateurs logiques (OR, AND, NOT) permettant d’effectuer des opérations d’union, d’intersection et de différence entre les ensembles de résultats associés à chaque terme. Un exemple de représentation d’une requête est comme suit : q = (t1∧ t2) ∨ (t3∧ t4). La fonction de correspondance est basée sur l’hypothèse de présence/absence des termes de la requête dans le document et vérifie si l’index de chaque document dj implique l’expression logique de la requête q. Le résultat de cette fonction est donc binaire est décrit comme suit : *RSV (q, d) = {0,1}*

* + 1. Le modèle vectoriel

Le modèle vectoriel de base a été introduit par Salton **[17]** concrétisé dans le cadre du système SMART. Ce modèle se base sur une formalisation géométrique. En effet, les documents et les requêtes sont représentés dans un même espace, défini par un ensemble de dimensions, chaque dimension représente un terme d’indexation. Les requêtes et les documents sont alors représentés par des vecteurs, dont les composantes représentent le poids du terme d’indexation considéré dans le document (la requête). Formellement, si on a un espace *T* de termes d’indexation de dimension *n, T = {t1,t2,….,tj,….,tn}.* Un document d*i*est représenté par un vecteur *di (wi1,wi2,….,wij,….,win).* Une requête *q* par un vecteur q(*wq1,wq2,….,wqj,….,wqn).*Où *Wij(resp.Wqj)* représente le poids du terme tj dans le document *di* (respectivement dans la requête q).

Le modèle vectoriel offre des moyens pour la prise en compte du poids de terme dans le document. Dans la littérature, plusieurs schémas de pondération ont été proposés. La majorité de ces schémas prennent en compte la pondération locale et la pondération globale. La pondération locale permet de mesurer l’importance du terme dans le document. Elle prend en compte les informations locales du terme qui ne dépendent que du document. Elle correspond en général à une fonction de la fréquence d’occurrence du terme dans le document (noté *tf* pour term frequency), exprimée ainsi :

*Tfij = 1 + log(f(ti,dj))* (I.1)

Où *. f(ti,dj)* est la fréquence du terme ti dans le document dj.

Quant à la pondération globale, elle prend en compte les informations concernant le terme dans la collection. Un poids plus important doit être assigné aux termes qui apparaissent moins fréquemment dans la collection. Car les termes qui apparaissent dans de nombreux documents de la collection ne permettent pas de distinguer les documents pertinents des documents non pertinents (i.e. peu utile pour la discrimination). Un facteur de pondération globale est alors introduit. Ce facteur nommé *idf* (inverted document frequency), dépend d’une manière inverse de la fréquence en document du terme et exprimé comme suit :

*Idf = log(N/ni)* (I.2)

Où ***ni*** est la fréquence en document du terme considéré, et ***N*** est le nombre total de documents dans la collection.

Les fonctions de pondération combinant la pondération locale et globale sont référencées sous le nom de la mesure *tf × idf.* Cette mesure donne une bonne approximation de l’importance du terme dans les collections de documents de taille homogène. Cependant, un facteur important est ignoré, la taille du document. En effet, la mesure (*tf × idf)* ainsi définie favorise les documents longs, car ils ont tendance à répéter le même terme, ce qui accroit leur fréquence, par conséquent augmentent la similarité de ces documents vis-à-vis de la requête. Pour remédier à ce problème, des travaux ont proposé d’intégrer la taille du document dans les formules de pondération, comme facteur de normalisation **[18] [19].**

L’appariement document-requête dans le modèle vectoriel, consiste à trouver les vecteurs documents qui s’approchent le plus de vecteur de la requête. Cet appariement est obtenu par l’évaluation de la distance entre les deux vecteurs. Plusieurs mesures de similarité ont été définies**[20]** dont les plus courantes sont décrites dans **le tableau 1** ci-dessous.

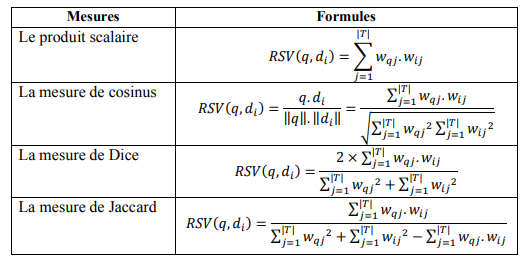
**

Tableau I‑1 :Les mesures de similarités utilisées dans le modèle vectoriel

Le modèle vectoriel caractérisé par sa prise en compte du poids des termes dans les documents, permet de retrouver des documents qui répondent partiellement à une requête. De plus, ce modèle offre un moyen facile pour classer les résultats d’une recherche, qui est basée sur la similarité potentielle entre documents et requête. L’inconvénient majeur de modèle vectoriel est qu’il repose sur l’hypothèse de l’indépendance des termes d’indexation, or ces termes dans les documents sont souvent sémantiquement liés. Plusieurs variantes du modèle vectoriel ont été proposées, pour remédier à cette limitation, c’est-à-dire prendre en compte la dépendance entre termes d’indexation. Parmi elles, on trouve, le modèle vectoriel généralisé, le modèle LSI (Latent Semantic Indexing) et le modèle connexionniste **[5].**

* + 1. Les modèles probabilistes
       1. Le modèle probabiliste de base

Le modèle probabiliste est fondé sur la théorie des probabilités. Il trie les documents selon leur probabilité de pertinence vis-à-vis d’une requête. La fonction de classement (tri) de ce modèle est exprimée ainsi :

RSV(q ,d)= (I.3)

L’idée de base de cette fonction est de sélectionner les documents ayant à la fois une forte probabilité d’être pertinents et une faible probabilité d’être non pertinents à la requête.

Où *P(Per |q,di)* et *P(NPer |q,di)*: la probabilité qu'un document d soit pertinent (*Per*) vis-à-vis de la requête q) respectivement non pertinent *(NPer)).*

En appliquant la formule de bayes pour les deux probabilités on obtient :

P(Per**|**q,di)= (I.4)

P(NPer**|**q,di)= (I.5)

Où :

*P(di)* est la probabilité de choisir le document di , on considère qu’elle est constante ;

*P(di|Per, q)* indique la probabilité que di fait partie des documents pertinents pour la requête q ;

*P(di|NPer, q)* indique la probabilité que di fait partie des documents non pertinents pour la requête *q* ;

*P(Per|q) et P(NPer|q)* indiquent respectivement la probabilité de pertinence et de non-pertinence d’un document quelconque (avec *(Per|q) + P(NPer|q) = 1 )* qui sont fixes. Après remplacement dans la fonction de tri, on aura la formule suivante :

Si on suppose que les termes d’indexation sont indépendants, alors on peut estimer les deux probabilités ainsi :

*P(di|Per, q) = ×*  *(I.6)*

*P(di|NPer, q) = × (I.7)*

Où *P(tj|Per, q)* indique la probabilité d’apparition du terme *tj* sachant que le document appartient à l’ensemble des des documents pertinents *et P(tj|NPer, q)* indique la probabilité d’apparition du terme *tj* sachant que le document appartient à l’ensemble des documents non pertinents.

En posant *Pi = P(tj|Per, q)* *, Qi = P(tj|NPer, q)* et *Pi = qi* pour les termes qui n’apparaissent pas dans la requête, et après simplification, le calcul du score de correspondance entre un document et une requête peut être exprimé ainsi :

*RSV (di, q) = (I.8)*

Afin de classer les documents avec cette formule, il faut estimer les valeurs des deux probabilités. En l’absence de collection (documents) d’apprentissage ; on peut attribuer la valeur fixe à *Pi* comme par exemple 0.5 ; comme elles peuvent être estimées à l’aide de l’avis de l’utilisateur sur les résultats d’une première recherche (réinjection de pertinence).

* + - 1. Le modèle de langue

Les modèles statistiques de langue sont exploités avec beaucoup de succès dans divers domaines : la reconnaissance de la parole **[21]**, la traduction automatique **[22][23]**, la recherche d’information **.**etc.

L’utilisation des modèles de langue en RI remonte à 1998. Le principe de ce modèle consiste à construire un modèle de langue pour chaque document, soit *Md* , puis de calculer la probabilité qu’une requête q puisse être générée par le modèle de langue du document, soit *P(q|Md).*

Le modèle de langue utilisé est souvent le modèle uni-gramme, la probabilité *P(q|Md)* est alors exprimée ainsi :

*P(q|Md)=* (I.9)

*P(t |Md)* peut être estimé en se basant sur l’estimation maximale de vraisemblance (maximum likelihood estimation). Elle est donnée par :

*P(t,Md)=*  (I.10)

Où *tf(t, d)* est la fréquence du terme *ti* dans le document *d*.

Pour remédier au problème posé par les mots de la requête absents dans le document, qui ont pour effet d’avoir la probabilité P(q|Md) nulle ; des techniques de lissage (smoothing) sont utilisées, dont le lissage de Laplace (ajouter-un), le lissage de Good-Turing, le lissage Backoff, le lissage par interpolation, etc. Leur principe consiste à assigner des probabilités non nulles aux termes, qui n’apparaissent pas dans les documents.

* 1. La reformulation de requêtes

Dans les SRI, la requête initiale seule est souvent insuffisante pour permettre la sélection de documents répondant au besoin de l'utilisateur. De ce fait, plusieurs techniques ont été proposées pour améliorer les performances des SRI. Ces méthodes apportent des solutions aux deux principales questions :

1. Comment peut-on retrouver plus de documents pertinents vis-à-vis d'une requête donnée?

2. Comment peut-on mieux exprimer la requête de l'utilisateur de manière à mieux répondre à son besoin?

La figure (I.4), présente les principales techniques d'amélioration des SRI par reformulation de la requête initiale en y ajoutant de nouveaux termes. La reformulation peut se faire par expansion automatique de la requête, par combinaison de différentes présentations de la requête ou par réinjection de pertinence. Nous présentons dans ce qui suit ces trois principales techniques **[4].**

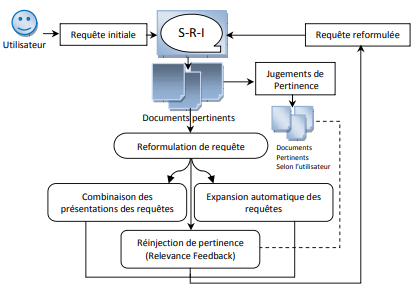


Figure I‑4 : Techniques d'améliorations des SRI par reformulation de requete.

* + 1. Expansion automatique des requêtes

L'expansion directe de la requête consiste à rajouter à la requête initiale des termes issus de ressources linguistiques existantes ou bien de ressources construites à partir des collections. Plus précisément, un niveau des ressources linguistiques, le but est d'utiliser un vocabulaire contrôlé issu de ressources externes. On peut alors utiliser des ontologies linguistiques (citons par exemple Wordnet). On peut également ajouter à la requête des variantes morphologiques des termes employés par l'utilisateur. Le but de ce mécanisme est d'assurer la restitution des documents indexés par des variantes des termes composant la requête. Les associations établies manuellement traduisent généralement des relations de synonymie et de hiérarchie. Les thésaurus construits manuellement sont un moyen efficace pour l'expansion de requête. Cependant, leur construction et la maintenance des informations sémantiques qu'ils contiennent sont coûteuses en temps et nécessitent le recours à des experts des domaines considérés. Pour cette raison, ils restent peu utilisés par les SRI. En ce qui concerne la seconde catégorie de ressources, elles sont construites en s'appuyant sur une analyse statistique des collections. Il s'agit de chercher des associations de termes afin d'ajouter des termes voisins à la requête. Il existe aussi d'autres méthodes entièrement automatiques telles que le calcul des liens contextuels entre termes et la classification automatique de documents. Les associations créées automatiquement sont généralement basées sur la cooccurrence des termes dans les documents. Les liens inter-termes renforcent la notion de pertinence des documents par rapport aux requêtes **[4].**

* + 1. Combinaison des présentations des requêtes

Plusieurs approches de RI utilisent une seule représentation de requête comparée à plusieurs représentations de document (algorithmes multiples de recherche). Il a été montré dans **[24]**qu'une recherche plus efficace peut être atteinte en exploitant des représentations multiples de requêtes ou des algorithmes de recherche différents ou encore en utilisant différentes techniques de réinjection. Une combinaison des représentations de requêtes peut augmenter le rappel d'une requête, tandis que la combinaison des algorithmes de recherche peut augmenter la précision. La base théorique de la combinaison des évidences a été présentée par Ingwersen **[25].** Il a en particulier montré que des représentations multiples d’un même objet, par exemple une requête, permettent une meilleure perception de l'objet qu'une seule bonne représentation. Cependant, il est important que chacune des sources d'évidences utilisées fournisse non seulement un point de vue différent sur l'objet, mais que ces points de vue aient différentes bases cognitives. Les représentations multiples d'une requête peuvent donner différentes interprétations du besoin en information. Une des approches de combinaison de multiples représentations de requêtes est proposée dans **[26].** Elle consiste à calculer les scores des documents directement depuis la fonction d'appariement document-requête en utilisant le même système de recherche mais différentes versions de la requête. Ensuite, les résultats obtenus par chacune des versions sont combinés pour avoir une seule liste finale. Ces versions sont issues soit des expressions d'une même requête par des chercheurs différents, soit des présentations d'une même requête dans des langages différents. Tamine et al, proposent dans **[27]** une technique de recherche d'information basée sur les algorithmes génétiques, plus précisément, elle propose d'utiliser une population de requêtes qui évolue à chaque étape de la recherche et tente de récupérer le maximum de documents pertinents **[4].**

* + 1. Réinjection de pertinence

Ces méthodes impliquent que l’utilisateur doit sélectionner les documents qu’il considère pertinents à partir des résultats issus de sa requête initiale. Ce jugement de pertinence de l’utilisateur est ensuite exploité pour reformuler la requête initiale en modifiant le poids des termes qu’elle contient et/ou en ajoutant de nouveaux termes considérés utiles pour retrouver des documents pertinents. La technique de réinjection de pertinence a été mise en place à l’origine dans le modèle vectoriel. Rocchio**[28]** a proposé le modèle de reformulation de requête suivant **:**

*QN = α.Q0 + β. -*  (I.11)

QN est le vecteur de la nouvelle requête (reformulée);

Q0 est le vecteur de la requête originale;

**R** est l’ensemble des vecteurs *r* des documents jugés pertinents par l’utilisateur;

est l’ensemble des vecteurs des documents jugés non pertinents par l’utilisateur;

sont les paramètres de la reformulation. On peut remarquer que cette formule permet d’obtenir une nouvelle requête dont le vecteur se rapproche des vecteurs des documents jugés pertinents et s’éloigne des vecteurs des documents jugés non pertinents. Dans le modèle probabiliste, la réinjection de pertinence est mise en place directement dans le modèle de mesure de pertinence. Elle consiste à revoir les poids des termes de la requête, comme suit :

*Wqj = log* (I.12)

Où : représente le nombre de documents pertinents ;

est le nombre de documents pertinents contenant le terme *qj*;

***dfj*** est le nombre de documents contenant le terme *qj*;

***n*** est le nombre total de documents dans la collection.

* 1. Evaluation des performances d’un système d’information

L’évaluation d’un SRI constitue une étape importante dans l’élaboration d’un modèle de RI, puisqu’elle permet de paramétrer le modèle, et de fournir des éléments de comparaison entre modèles. L’évaluation nécessite alors, la définition d’un ensemble de mesures et méthodes d’évaluation et collections de test, assurant l’objectivité de l’évaluation. L’évaluation d’un SRI, peut être appréhendée selon deux aspects : un aspect efficience et un aspect efficacité. L’aspect efficience dépend de l’évaluation cognitive de l’utilisateur, tels que la facilité d’utilisation du système, rapidité d’accès, temps de réponse à une requête, présentation des résultats, etc. L’aspect efficacité concerne la capacité du système à sélectionner le maximum de documents pertinents et un minimum de documents non pertinents**].**

* + 1. Les Collections de test

Une collection (ou corpus) de test constitue le moyen d’évaluation des SRI. Elle est généralement composée d'un ensemble de documents, d’un ensemble de requêtes et des jugements de pertinence associés à ces requêtes. L’évaluation d’un SRI consiste à comparer les résultats retournés par ce dernier par rapport aux jugements de pertinence. Des mesures d’évaluation, décrites dans la section suivante, sont utilisées pour effectuer cette comparaison. Les collections de test sont le résultat de projets d’évaluation qui se sont multipliés depuis les années 1970, on peut citer la collection CACM1, la collection CISI2, la campagne CLEF3 et la campagne TREC4.

La campagne TREC constitue à ce jour la campagne de référence dans le cadre de l’évaluation des systèmes de recherche d’information et cela depuis son lancement en 1992 **[29][30]** L'objectif de cette campagne est de proposer une plate-forme qui réunit des collections de test, des tâches spécifiques et des protocoles d'évaluation pour chaque tâche afin de mesurer les différentes stratégies de recherche**[31].**

Les tâches proposées se sont diversifiées d’une campagne à une autre (à raison d’une par an) ; parmi les tâches proposées dans TREC 2012 on peut citer : recherche d’information sur le web, recherche d’information médical, recherche d’information dans les micros blogs, recherche d’information contextuelle, etc. La taille des collections augmente au fil des années, passant de 2 Go dans TREC 1 à 25 To dans TREC 2011. Chaque collection est composée d'un certain nombre de documents, allant de quelques milliers à plusieurs millions. Les documents sont codés à l'aide de SGML dans un format spécifique TREC. La Figure I.5 illustre un exemple d’un document TREC.

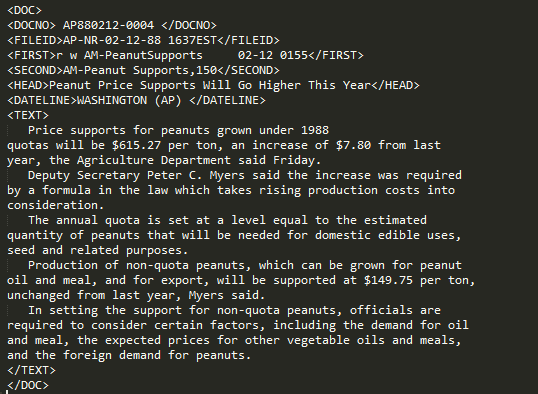
****

Figure I‑5 : Exemple d'un document TREC.

Chaque collection TREC a généralement 50 à 100 requêtes correspondantes. Une requête TREC est structurée comme suit: un identifiant de requête unique TREC, un titre, une description plus détaillée du besoin en information et une rubrique qui explique dans quelles circonstances un document doit être jugé pertinent ou non pertinent pour une requête. Un exemple d’une requête TREC est montré dans la figure I.6.

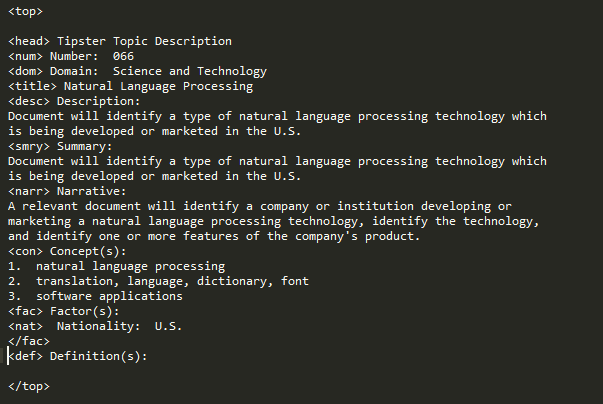
****

Figure I‑6 : Exemple d'une requête TREC.

* + 1. Mesures d'évaluation du SRI

Le principal objectif d’un système de recherche d’information est de restituer à l’utilisateur tous les documents pertinents et de rejeter tous les documents non pertinents. Cet objectif est évalué à l’aide de différentes mesures d’évaluation **[32].** On présente ci-dessous les plus utilisées.

* **La précision** : est le rapport du nombre de documents pertinents restitués par le système (*SP*) sur le nombre total de documents restitués *(R),* exprimée ainsi :

*Précision* = (I.13)

* **Le rappel** : est le rapport du nombre de documents pertinents restitués (*SP*) sur le nombre total de documents pertinents *(P),* exprimé ainsi :

Rappel = (I.14)

* **Le bruit** : la mesure d’évaluation bruit est une notion complémentaire à la précision, elle est définie par *B = 1 - P* où *P* est la précision du SRI.
* **Le silence** : la mesure d’évaluation silence est une notion complémentaire au rappel, elle est définie par *S = 1 - R* où *R* est le rappel du SRI.
* **La précision moyenne non interpolée (MAP) :**

La précision moyenne non interpolée (Mean Average Precision) est calculée en deux étapes. D’abord on calcule la précision moyenne pour une requête donnée () ainsi pour chaque document pertinent retrouvé on calcule sa précision () qui est égale au nombre de documents pertinents retrouvés sur le rang de ce document ; pour les documents retrouvés non pertinents leur précision est égale à zéro.

La précision moyenne pour une requête donnée est alors obtenue en calculant la  
moyenne des précisions des documents pertinents, exprimée ainsi :

= (I.15)

Avec

Si est retrouvé

Sinon

=

Où dénote le rang du document qui a été retrouvé et qui est pertinent pour la requête, le nombre de documents pertinents retrouvé au rang et N est le nombre total de documents pertinents pour la requête q.

Dans la seconde étape, on calcule la précision moyenne pour un ensemble de en effectuant la moyenne des précisions moyennes de chaque requête, elle est exprimée ainsi :

MAP (I.16)

Où dénote la précision moyenne pour la requête « j » et M représente le nombre de requêtes considérées.

### La précision à N document : C’est la proportion des documents les plus pertinents retournés DPR au rang N, alors la précision est exprimée ainsi :

(I.17)

* **Courbe de Rappel-Précision** : un système idéal devrait retourner tous les documents pertinents et que les documents pertinents ; c’est à dire un taux de précision et de rappel égal à 100%. Cette situation ne se produit pas dans un système réel car le taux de précision et de rappel sont antagonistes. En effet, Lorsque la précision augmente, le rappel diminue et inversement. Ainsi, pour mesurer les performances d’un système il faut utiliser les deux mesures conjointement. Cela est réalisé en calculant la paire des mesures (taux de rappel, taux de précision) à chaque document restitué. Nous considérons par exemple une requête pour laquelle il existe cinq (5) documents pertinents dans le corpus. Le Tableau I.2 illustre le calcul de la précision et de rappel pour les dix (10) premiers documents retournés par un SRI. La lettre (P) précise que le document est pertinent.

|  |  |  |  |
| --- | --- | --- | --- |
| Rang document | Pertinence | Rappel | Précision |
| Doc 1 | P | 1/10 | 1 |
| Doc 2 | P | 2/10 | 1 |
| Doc 3 |  | 2/10 | 2/3 |
| Doc 4 |  | 2/10 | 2/4 |
| Doc 5 | P | 3/10 | 3/5 |
| Doc 6 | P | 4/10 | 4/6 |
| Doc 7 |  | 4/10 | 4/7 |
| Doc8 | P | 5/10 | 5/8 |
| Doc 9 |  | 5/10 | 5/9 |
| Doc 10 |  | 5/10 | 5/10 |

Tableau I‑2 :Exemple de calcul de rappel et de précision pour une requete.

La figure I.7 illustre la courbe de rappel et précision correspondante aux résultats du Tableau I.2.

Pour rendre la courbe lisible on ne garde que la précision calculée à chaque point de rappel (c’est à dire à chaque document pertinent restitué).

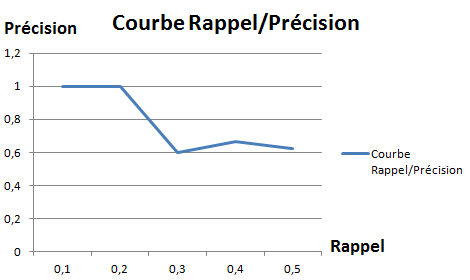


Figure I‑7 : Courbe de Rappel et Précision.

La courbe ci-dessus permet d’évaluer les performances du système pour la requête considérée. Afin d’évaluer le système pour un ensemble de requêtes, on calcul la moyenne des précisions à chaque niveau de rappel. Comme les niveaux de rappel ne sont pas unifiés pour l'ensemble des requêtes, on retient généralement 11 points de rappel standards de 0 à 1 avec un pas de 0.1. Les valeurs de précision sont calculées par une interpolation linéaire. Pour deux points de rappel, *i* et *j, i < j,* si la précision au point *i* est inférieure à celle au point *j*, on dit que la précision interpolée à *i* égale la précision à *j*. Cette interpolation est encore discutable, mais présente un intérêt dans l’évaluation de systèmes de recherche d’information. Elle permet entre autre de construire des courbes décroissantes plus simple à comparer **[6].**

* 1. Conclusion

Dans ce chapitre nous avons présenté les principales notions et concepts de la recherche d’information. Nous avons, particulièrement, introduit des notions de base, telles que le besoin en information, la requête, le document et la pertinence. Nous avons aussi décrit les processus de base de la RI, à savoir l’indexation, l’appariement requête-document. Ensuite, nous avons étudié les différents modèles de la RI. Enfin, la reformulation de requête. Le chapitre suivant portera sur la pertinence à priori de document ainsi qu’une méthode de combinaison des scores.