

1 Introduction

La Recherche d'Information (RI) a pour but de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défi est de pouvoir trouver ceux qui correspondent aux besoins de l'utilisateur parmi le volume important de documents disponibles. La démarche de la RI est fait par des Systèmes de Recherche d'Information (SRI).

Le but de ces systèmes est de répondre au besoin en information de l'utilisateur. [24] C'est-à-dire mettre en correspondance une requête de l'utilisateur avec un contenu des documents en utilisant une fonction d'appariement.

Dans ce chapitre on va donner un aperçu général sur la recherche d'information et leur processus.

2 La recherche d'information

La RI est un domaine vaste qui se situe dans le cadre de plusieurs disciplines on cite : Classification /catégorisation (clustering), Question-réponses (Query answering), Filtrage d'information (filtering/recommendation), Résumé automatique (Summarization) et Fouille de textes (Text mining).

Définition 1

« Information retrieval (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) [6] ».

Définition2

« 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 [24] ».

Ces définitions partagent la même idée que la RI a pour objectif d'extraire les informations pertinentes dans un document ou d'un ensemble de documents qui reflètent un besoin d'un utilisateur (requête).

3 Processus de recherche d'information

En se basant sur une requête, le SRI exécute une série des méthodes qui permettent de trouver une liste des documents pertinents. Le processus de recherche d'information comprend plusieurs concepts : (le Figure 1.1)

- La collection de documents ou corpus.
- Le besoin en information.

- L'indexation.
- La fonction d'appariement requête-document.
- La fonction de modification de requête qui se traduit généralement par un mécanisme de reformulation des requêtes.

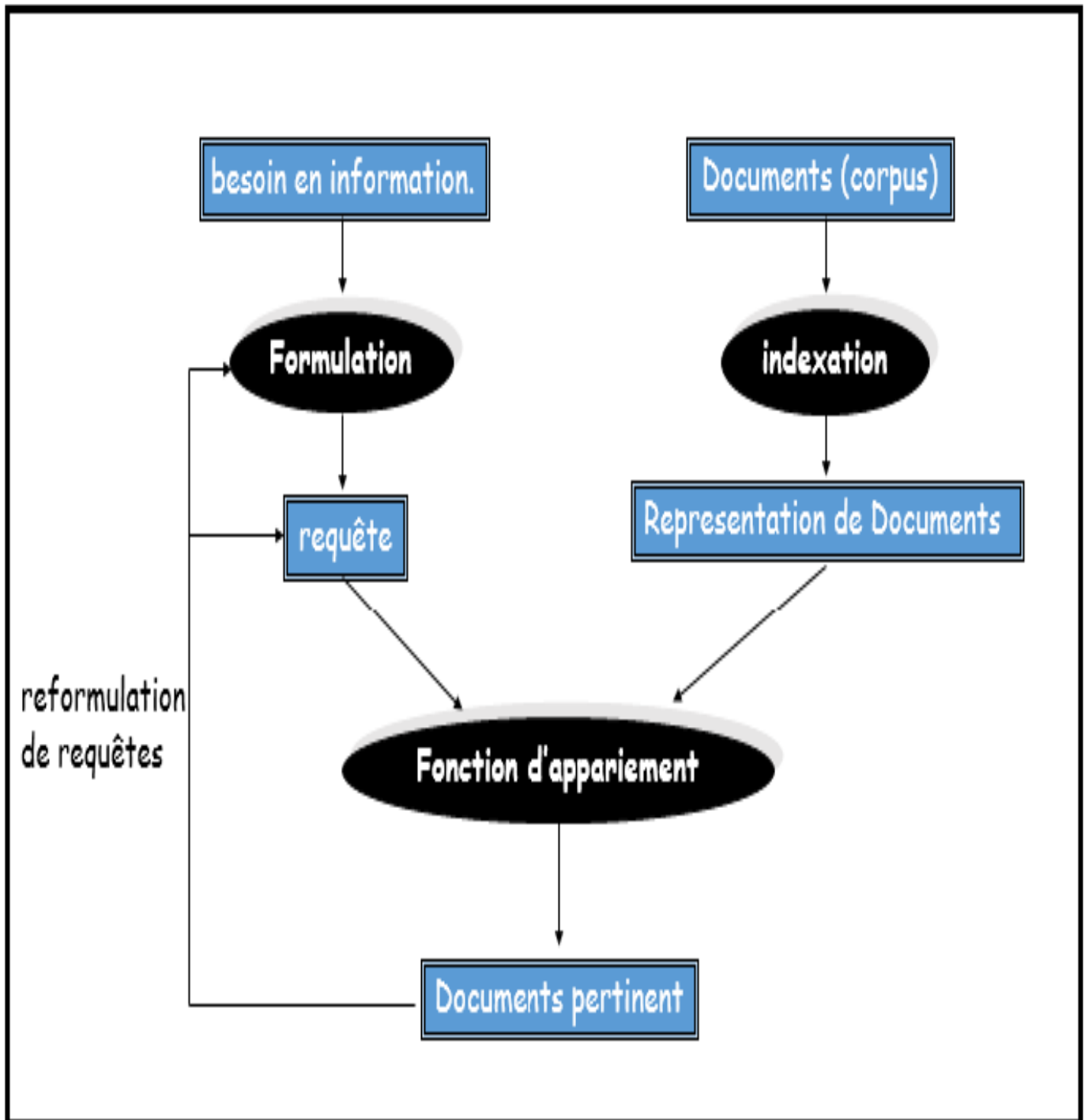


Figure 1.1 : Processus de Recherche d'Information. [18]

3.1 Collection de documents (corpus)

C'est une collection de documents constitue l'ensemble des informations exploitables et accessibles.

3.2 Besoin en information

Ce besoin est l'expression mentale de ce que l'utilisateur cherche. L'expression d'un besoin se fait par une requête qui permet l'interrogation d'un système de recherche d'information.

3.3 Indexation

L'étape d'indexation permet de réaliser le passage d'un document textuel (ou une requête) à une représentation exploitable par un modèle de RI par la construction de mots clés appelé langage d'indexation. [25]

3.3.1 Les Modes d'indexation

L'indexation peut être manuelle, automatique ou semi-automatique :

- Indexation manuelle : chaque document est analysé par un spécialiste du domaine correspondant ou par un documentaliste.
- Indexation automatique : chaque document est analysé à l'aide d'un processus entièrement automatisé.
- Indexation semi-automatique : le choix final reste au spécialiste du domaine correspondant ou documentaliste, qui intervient souvent pour établir des relations sémantiques entre les mots-clés et choisir les termes significatifs.

Dans les sections suivantes décrit les différentes étapes de l'indexation automatique.

3.3.2 Prétraitements

Une étape de prétraitements sur les documents et les requêtes est nécessaire.

3.3.2.1 Segmentation

Segmentation ou tokenisation est la tâche de séparer les mots (morphèmes) de texte. Elle fait référence à la division d'un mot dans des groupes de morphèmes. Nous pouvons utiliser des blancs (espace blanc) pour aider à cette tâche, mais il y a des cas difficiles, il est important de se rappeler qu'il n'y a pas de segmentation optimale seule. [3] On peut généralement trouver trois types de segmentation :

- La segmentation lexicale (*tokenisation*) qui est la segmentation d'un texte en segments lexicaux (*tokens*).

- La segmentation morphologique en cherchant à isoler les différents constituants des items lexicaux en unités distinctes, plus petites, qui sont les morphèmes.
- La segmentation syntaxique (*chunking*) qui consiste à isoler les différents constituants du texte en unités indépendantes, supérieures aux mots, comme les propositions, les syntagmes...etc.

-Token est une instance d'une séquence de caractères dans un document particulier sont regroupés en une unité sémantique utile pour le traitement.

-Types est la classe de tous les tokens contenant la même séquence de caractères.

Exemple : Je remue le ciel, le jour, la nuit

Liste des tokens : Je, remue, le, ciel, le, jour, la, nuit.

Liste des types : Je, remue, le, ciel, jour, la, nuit.

Généralement les langues appartenant à deux classes différentes : les langues «avec séparateurs» et les langues «sans séparateurs». [25]

Une langue dite « avec séparateur » présente un système d'écriture segmentée c'est-à-dire les mots sont nettement séparés par des délimiteurs (espace, signes de ponctuation, caractères spéciaux, ...), comme le français ou l'anglais.

Une langue dite « sans séparateur » présente un système d'écriture non segmentée où les mots ne sont pas séparés par des espaces c'est-à-dire les frontières des mots ne sont pas nettes, le japonais, le chinois et le thaï sont les représentants parfaits de cette deuxième famille de langues.

3.3.2.2 Normalisation

La normalisation est le processus de canonisant les tokens (même forme), la façon la plus classique de la normalisation est de créer des classes d'équivalence, qui sont nommés d'après un membre de l'ensemble.

Exemples :

- Accents, Diacritiques :

Élevé, élève, → eleve

Naïve → naive

- Conversion des caractères en majuscule ou minuscule.

USA →usa

General Motors→ general motors

- En arabe, suppression tachkil (kasra, fatha, tanwin...).

كتب → كُتِبَ

3.3.2.3 *Etiquetage morphosyntaxique (optionnel)*

L'étiquetage morphosyntaxique comprend une analyse morphologique et une analyse syntaxique. Ces deux analyses sont précédées par certains prétraitements (traitement des ponctuations, majuscules, codages et formats). [5].

L'analyse syntaxique permet de segmenter les textes en propositions. Chaque proposition est formée de couple (entrée lexicale, catégorie). Les seules ambiguïtés qui demeurent sont internes à une catégorie. Les résultats de l'analyse des propositions sont des arborescences de structures syntaxiques attestées par la langue [4].

L'étiquetage morphosyntaxique joint à chaque mot d'une phrase sa catégorie morphologique et syntaxique.

Exemple : Soit la phrase « le chat est un animal domestique. », une analyse morphosyntaxique de cette phrase donne :

- « Le – article défini masculin singulier ».
- « Chat – nom commun masculin singulier ».
- « Est – verbe indicatif présent 3ème personne du singulier ».
- « Un – article indéfini masculin singulier ».
- « Animal – nom commun masculin singulier ».
- « Domestique – adjectif qualificatif ».
- « . – ponctuation forte (fin de phrase) ».

3.3.2.4 *Lemmatisation*

La lemmatisation consiste à fournir, pour une forme donnée, la représentation standardisée du mot correspondant, utilisée le plus souvent en entrée dans un dictionnaire de référence. [11]
La lemmatisation d'une forme d'un mot consiste à prendre sa forme canonique.

Exemple : Il existe beaucoup plus de formes du verbe avoir : ai, as, a, avons, ais, avons eu, ayez eu, eussions eu, aurions eu, etc. La forme canonique est **avoir**.

3.3.2.5 *Dés-suffixation « stemming »*

Dés-suffixation (appelé Stemming en anglais). Ils associent plusieurs mots ayant le même radical, c'est-à-dire enlever les affixes des mots pour ne conserver que la partie racine, en s'aidant de règles et de listes d'exceptions. Les algorithmes de « stemming » les plus connus sont ceux de Lovins et de Porter.

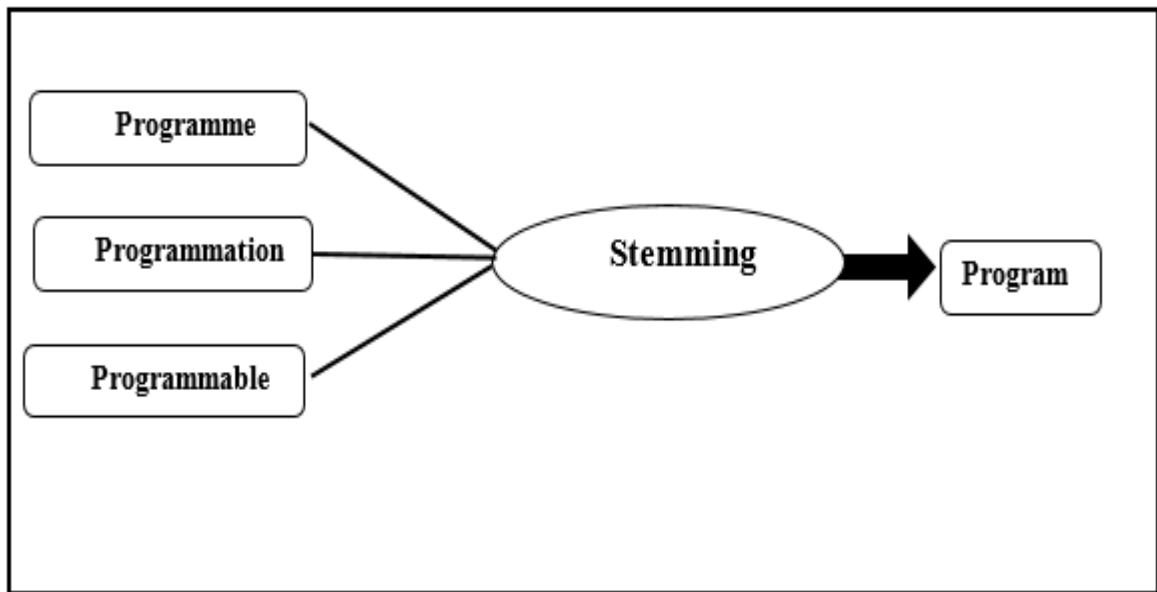


Figure 1.2 : Exemple d'un processus de stemming.

3.3.2.6. *Suppression des mots vides*

Les mots vides sont une division de langage naturelle. Le raison que les mots vides devraient être supprimés à partir d'un texte est qu'ils font le texte regardé plus lourd. La suppression de mots vides réduit la dimensionnalité de l'espace de terme. Les mots les plus courants dans les documents texte sont des articles, des propositions et pronoms, etc. qui ne donnent pas le sens des documents. Ces mots sont traités comme mot vides. [8]

3.3.3 Fonction de pondération

La pondération permet d'affecter à chaque terme d'indexation une valeur qui mesure son importance dans le document où il apparaît. [1]

Le poids d'un terme représente le degré de son importance dans le document. Il y a différents facteurs permettant la pondération des termes. *TF-IDF* est la méthode de pondération qui a été la plus étudiée en recherche documentaire. Le terme ***TF-IDF*** désigne un ensemble de pondérations et des sélections des termes.

- Term frequency *tf* : indique l'importance du terme dans le document, plus un terme est fréquent dans un document, plus il est important dans la description de ce document. Plusieurs formules de pondération locale ont été proposées, parmi lesquelles : la fonction brute (nombre d'occurrences), la fonction binaire et la fonction logarithmique. [10]

- Inverted document frequency *idf* : ce facteur mesure l'importance d'un terme dans toute la collection (pondération globale). La formule qui exprime l'importance d'un terme dans sa collection est : $\log(N/DF)$, où *DF* représente le nombre de documents contenant le terme et *N* représente le nombre total de documents de la base documentaire. [10]

Soit $tf(mi, dj)$ la fréquence du terme *mi* dans le document *dj* et $idf(mi)$ l'inverse du nombre de documents qui contiennent le terme *mi*, les poids **TF-IDF** sont simplement calculés comme suit : [22]

$$TF-IDF = tf(mi, dj) \times idf(mi) \quad (3)$$

3.3.4 Création de l'index

L'index, défini comme étant la structure de stockage utilisée pour mémoriser les informations sélectionnées pour la représentation du texte. Cette structure permet de sélectionner, pour n'importe quel terme, tous les documents où il apparaît.

3.4 L'appariement requête-document

Les SRI intègrent un processus de recherche/décision qui permet de sélectionner l'information pertinente pour l'utilisateur. A cet effet, une mesure de correspondance entre la requête et les documents est calculée. [1]

4.3.1 Mesure d'appariement

Les documents pertinents retournés par un SRI peuvent être définis comme les plus proches de la requête selon une certaine mesure d'appariement (distance). [23]

L'utilité de distance est de pouvoir comparer les similarités et les différences entre deux vecteurs. Une distance doit vérifier les axiomes suivants : [16]

- $d(A, B) = 0 \Leftrightarrow A = B$ (séparation).
- $d(A, B) = d(B, A)$ (symétrie).
- $d(A, C) \leq d(A, B) + d(B, C)$ (inégalité triangulaire).

3.4.1.1 la distance euclidienne [26.]

La méthode la plus utilisée. Pour calculer la distance entre les vecteurs est la distance euclidienne, cette mesure peut calcule comme suit :

$$\delta_{L2}(q, d) = \sqrt{\sum_{t \in V} |q_t - d_t|^2} \quad (4)$$

Tel que V est le vocabulaire de la collection, c'est-à-dire, l'ensemble des termes d'indexation de tous les documents).

3.4.1.2 la similarité cosine

La notion de similarité entre document et requête est liée au choix de la méthode de représentation des textes. La représentation la plus utilisée est la représentation vectorielle, dans le cadre dans lequel le document et la requête sont représentés par des vecteurs dans un espace vectoriel dont les dimensions sont associées à des unités linguistiques spécifiques (mot, *stems*, lemmes, etc.). La similarité entre document et requête est alors évaluée par une mesure de similarité définie sur cet espace vectoriel. [20]

La méthode le plus utilisé pour calculer la similarité entre de vecteurs est la similarité cosine, Cette mesure est calculée par la formule suivant : [17]

$$\cos(q, d) = \frac{\sum_{i=1}^n q_i * d_i}{\sqrt{\sum_{i=1}^n q_i^2 * \sum_{i=1}^n d_i^2}} \quad (5)$$

3.5 La fonction de modification de requête

La formulation du besoin en information d'un utilisateur en requête est difficile. Par conséquent, les documents trouvés par la requête initiale ne peuvent pas accomplir le besoin en information de l'utilisateur. Donc le système de recherche d'information fait appel à la fonction de modification des requêtes afin de corriger le chemin de la recherche.

4 Représentation « les modèles de RI »

Un modèle de recherche d'information a pour rôle de fournir une formalisation du processus de recherche d'information. Dans cette section on va donner un aperçu rapide des principaux modèles de recherche d'information.

4.1 Le modèle booléen [25]

Dans Le modèle booléen, les documents et les requêtes sont représentés par des ensembles des mots clés. Chaque document est représenté par une conjonction logique des termes non pondérés qui constitue l'index du document. Un document peut représenter par exemple :

$$d = t_1 \wedge t_2 \wedge t_3 \dots \wedge t_n \quad (6)$$

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. Une requête peut représenter par exemple :

$$q = (t_1 \wedge t_2) \vee (t_3 \wedge t_4) \quad (7).$$

La fonction de correspondance entre un document et une requête est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifier si l'index de chaque document d_j implique l'expression logique de la requête q .

4.2 Le modèle probabiliste

Modèle particulier, développé spécifiquement pour la recherche documentaire, utilise des pondérations justifiées sur des bases probabilistes dépendantes des requêtes (une requête est formée d'un ensemble de mots clés pour effectuer une recherche documentaire). La représentation finale de chaque document est un vecteur dont chaque composante est une pondération associée à un terme. [2]

Le modèle probabiliste consiste à calculer la pertinence d'un document en fonction de pertinences connues pour d'autres documents.

4.3 Le modèle vectoriel

Dans le cadre des modèles vectoriels considérés, l'espace de représentation des documents est un espace vectoriel dont chaque dimension est associée à une unité linguistique particulière, appelée *terme d'indexation*. [21]

Définition : « Dans le modèle vectoriel standard, chaque document \mathbf{d} est représenté par un vecteur à \mathbf{n} dimensions $(\mathbf{w}_1, \dots, \mathbf{w}_n)$, où \mathbf{w}_i est le poids du terme \mathbf{t}_i dans le document \mathbf{d} . Un terme peut être un mot, un lemme ou un composant (plusieurs mots ou lemmes ou stems) ». [19]

Dans le domaine de la recherche documentaire, plusieurs méthodes sont classées sous le nom de « modèles vectoriels ». Les méthodes les plus connues sont fondées sur le principe de « sac de mots » ou « bag of words »

« Sac de mots : Soit W le dictionnaire, l'ensemble de tous les termes (mots) qui se produisent ou moins une fois dans une collection de documents D . La représentation sac-de-mots de document d_n est un vecteur de poids $(w_{1n}, \dots, w_{|W|n})$ tel que les poids $w_{in} \in \{0, 1\}$ et indiquer la présence ou l'absence d'un terme i particulier dans un document n »

La transformation d'un ensemble de documents D dans la représentation de BOW permettre à l'ensemble transformé pour être considéré comme une matrice, où les lignes représentent les vecteurs de document, et les colonnes sont des termes. Car les documents sont traités comme

vecteurs donc ils peuvent être comparés à l'aide de mesures de distance / similarité classiques. [15]

5 Critères d'évaluation des SRI

Les utilisateurs d'un SRI ont des besoins très variés et des critères assez différents pour juger si un document est pertinent. Donc il est essentiel de disposer de techniques d'évaluation permettent de juger l'efficacité des SRI à retrouver les documents pertinents. [25]

Les systèmes de RI sont évalués en fonction de la pertinence des documents retrouvés. Les deux principales mesures utilisées pour évaluer un SRI sont la précision et le rappel.

5.1 Rappel

Un système de RI aura beaucoup de rappel s'il retrouve la plupart des documents pertinents du corpus pour une requête.

$$\text{rappel} = \frac{\text{nombre total de documents pertinents retrouvés par le système}}{\text{nombre total de documents pertinents dans le corpus}} \quad (8)$$

5.2 Précision

Un système de RI sera très précis si presque tous les documents retrouvés sont pertinents.

$$\text{précision} = \frac{\text{nombre total de documents pertinents retrouvés par le système}}{\text{nombre total de documents retrouvés par le système}} \quad (9)$$

5.3 F-mesure

La F-mesure prend en considération la précision et le rappel simultanément. Elle est définie comme la combinaison pondérée du taux de rappel et du taux de précision.

$$F - \text{mesure} = \frac{2 * \text{précision} * \text{rappel}}{\text{précision} + \text{rappel}} \quad (10)$$

6 Conclusion

La recherche d'information a pour but de faciliter l'accès à un ensemble de documents se trouvant dans des bases documentaires et pour permettre aux utilisateurs de retrouver les documents dont le contenu répond à leur besoin en information. Dans ce chapitre nous avons présenté les principales notions et concepts de RI et SRI.