الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة غرداية
Université de Ghardaïa

Faculté des Sciences et de la technologie
Département des Mathématiques et d'Informatique

**MEMOIRE DE FIN D'ETUDES**

En vue de l'obtention du

Diplôme de Master en Informatique

Option : Systèmes Intelligents pour l'Extraction des Connaissances

# L'apprentissage automatique et la comparaison entre le modèle VSM et Probabiliste

**Réalisé par:**

**MAMI Abdelkader**

**MATALLAHHouda**

**Encadrépar :**

**Mr. CHABBI Samir**

**Membres dejurys :**

**KARRACHAbd El Aziz Président**

**ADJILA Abd RahmanExaminateur**

**MAHDJOUB YoucefExaminateur**

**Session: juin 2018**

**Dédicace**

*Nous dédions cette mémoire*

*à nos parents*

*à nos familles*

*à nos collègues*

*à tous ceux qui nous ont toujours encouragé*

***MATALLAH***

***Et MAMI***

**Remerciements**

*Nous tenons tout d'abord à remercier* ***Dieu*** *le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce Modeste travail.*

*En second lieu, nous tenons à remercier notre encadreur Mr :* ***CHABBI Samir****, ces précieux conseils et orientations et son aide durant toute la période du travail.*

*Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail Et de l'enrichir par leurs propositions.*

*Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail*

# Table des matières

# Table des figures

# Liste des tableaux

**Résumé**

*La recherche d'information (RI) est le domaine qui étudie la manière de retrouver des informations dans un corpus, en fonction d'un besoin d'information.*

*La RI développe des modèles pour interpréter les documents, le besoin, et des techniques pour calculer des réponses rapidement même en présence de collections très volumineuses.*

*Alors que la recherche d'information est basée sur plusieurs modèles (booléen ; vectoriel ; probabiliste), ce projet se spécialise dans la comparaison de la pertinence des documentsselon les deux modèlesvectoriel et probabilistepourdéterminer le meilleurepour trouver l'information prévue par l'utilisateur. Aussion a proposé un modèle qui résoudre le problème des requêtes malexprimées en utilisant l'expansion de requête.*

*Mots clé*

*Recherche d'information, modèle de recherche d'information,expansion de requête.*

*ملخص*

استرجاع المعلومات هو المجال الذي يدرس كيفيه العثور علي المعلومات في مجموعة من الوثائقوطور في هذا المجال نماذج لتفسير الوثائق ، وحاجة المستخدم (الاستعلام) , والتوفيق بينها,وتقنيات لحساب الإجابات بسرعة حتى في وجود مجموعات كبيرة جدا.

وفي حين أن البحث عن المعلومات يستند إلى عده نماذج (منطقي , شعاعي,احتمالي) , فان هذا المشروع متخصص في مقارنه أهميه الوثائق وفقا لكل من النموذجين الشعاعي و الاحتماليلتحديد أيهما أفضللعثور علي المعلومات التي يريدها المستخدم. كما تم اقتراح نموذج لحل مشكله عدم التعبير الجيد عن حاجة المستخدم و استعلاماته باستخدام توسيع الاستعلام.

*الكلمات المفتاحية*

*البحث عن المعلومات ,نموذج استرجاع المعلومات, توسيع الاستعلام .*

***Abstract***

*Information retrieval (IR) is the field that studies how to find information in a corpus, depending on a need for information.IR develops models to interpret documents, need, and reconcile them, techniques for calculating responses quickly even in the presence of very large collections.*

*While the search for information is based on several models (Boolean; vector; probabilistic), this project specializes in comparing the relevance of documents according to both Vector and probabilistic models to determine the best to find the information provided by the user. Also a model has been proposed that solve the problem of misexpressed queries using query expansion*

**Key words**

*Information retrieval, information retrieval model, query expansion.*

# Introduction Générale

*La Recherche d'Information (RI) est un domaine de découverte de l'information, comporte plusieurs phases : de l'expression d'un besoin d'information jusqu'à un résultat qui satisfait plus ou moins ce besoin. Le but ultime est de trouver parmi le volume important de documents disponibles, ceux qui correspondent au mieux au besoin d'utilisateur. L'opérationde la RI est réalisée Par outils des appelés Systèmes de Recherche d'Information (SRI).*

*Chaque système de recherched'information est basé sur un modèle de recherche d'information. Alors comment définir le meilleur modèlede recherche parmi les modèles classiques ? Et comment injecter l'apprentissage automatique dans ces modèles ?*

*Ce projet est organisé en quatre chapitres, le premier chapitre contient les notions et les concepts de base de la recherche d'information, le deuxième chapitre est consacré au domaine d'apprentissage automatique car il est très important pour la recherche d'information, donc le troisième chapitre parle de l'intégration de l'apprentissage automatiquedans la recherche d'information.*

*Le dernier chapitre contient la comparaison entre les deux modèles vectoriel et probabiliste aussi dans ce chapitre nous avons touché la partie de l'expansion de requête. Enfin une conclusion générale.*

# Chapitre I

## Système de recherche d'information : Modèles et techniques

# 1. Introduction:

La recherche d'information est un domaine historiquement lié aux sciences de l'information. Elle traite l'information dans la manière de l'organiser et de la façon de la sélectionner, elle peut être définie comme une activité qui dans le but de répondre à une question vise à localiser et à traiter une ou plusieurs informations au sein d'un environnement documentaire complexe.

Dans ce premier chapitre, nous allons définir les concepts de base de la recherche d'information et les systèmes de recherche d'information (SRI).

## 1 .1. Définitions et concepts

Dans cette section, nous allons présenter quelques définitions et concepts liés au domaine de recherche d'information.

### 1.1.1. Recherche d'information

La recherche d'information est undomaine qui consiste à récupérer l'informationà travers la représentation des documents. L'informatique a permis de développer des outils pour traiter l'information et représenter des documents au moment de leur indexation, aussi pour : Rechercher, sélectionner, organiser et stockerl'information. Aujourd'hui, on peut considérer la recherche d'information comme étant, un domaine transdisciplinaire qui peut être étudié par plusieurs disciplines afin de trouver des solutions pour améliorer son efficacité[1].

- **Autres définitions**

Il y a plusieurs définitions de la recherche d'information, nous citons quelques suivantes **:**

- ✓ **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. [2]
- ✓ **Définition 2**
- ✓ La Recherche d'Information (RI) consiste à sélectionner dans une collection de documents ceux susceptibles d'être pertinents vis à vis d'un besoin en information d'un utilisateur. [3]

- ✓ **Définition 3**

- ✓ La recherche d'informations (RI) est un processus très essentiel pour exploiterles données. Elle traitela représentation, le stockage,l'organisation etl'accès oul'extractiondes informations. [4]

### 1.1.2. Collection de documents

La collection de documents (ou fond documentaire) constitue l'ensemble des informations exploitables, compréhensible etaccessibles. 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 fasse dans les meilleures conditions de coût.

### 1.1.3.Document

Le document constitue l'information élémentaire d'une collection de documents.

L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document.

### 1.1.4. 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.

## 1.2. Objectifs des systèmes de recherche d'information

L'objectif général d'un système de recherche d'informations est de minimiser les frais généraux d'un utilisateur qui localise les informations nécessaires. La surcharge peut être exprimée comme le temps qu'un utilisateur passe dans toutes les étapes menant à la lecture d'un élément contenant les informations nécessaires (par exemple, la génération de requêtes, l'exécution de la requête, l'analyse des résultats de la requête pour sélectionner les éléments à lire, la lecture des éléments non pertinents). Le succès d'un système d'information est très subjectif, basé sur les informations nécessaires et la volonté d'un utilisateur d'accepter les frais généraux. Dans certaines circonstances, il faut les informations peuvent être définies comme toutes les informations qui sont dans le système qui se rapportent aux besoins d'un utilisateur.

Un système qui prend en charge la récupération raisonnable nécessite moins de fonctionnalités que celle nécessitant une récupération complète. Dans de nombreux cas, la récupération complète est une fonction négative, car elle surcharge l'utilisateur avec plus d'informations que nécessaire. Il est donc plus difficile pour l'utilisateur de filtrer les informations pertinentes utiles à partir des éléments critiques. [5]

### 1.3. Historique

La longue histoire de la recherche d'informations ne commence pas par l'Internet. Ce n'est que dans la dernière décennie et demieque les moteurs de recherche Web sont devenus omniprésents et la recherche est devenue intégrée dans le tissu de bureau et les systèmes d'exploitation mobiles. Avant la grande utilisation quotidienne des moteurs de recherche, les systèmes de recherche d'information (RI) ont été retrouvés dans des applications commerciales et de renseignement depuis les années 1960. Les premiers systèmes de recherche informatisés ont été construits à la fin des années 1940 et ont été inspirés par l'innovation pionnière dans la première moitié du XXe siècle. Comme avec de nombreuses technologies informatiques, les capacités des systèmes de recherche a augmenté avec des augmentations de la vitesse du processeur et la capacité de stockage. Le développement de ces systèmes reflète également une progression rapide loin des approches manuelles basées sur la bibliothèque pour l'acquisition, l'indexation et la recherche d'informations vers des méthodes de plus en plus automatisées. [6]

## 2. Indexation pour la recherche d'information

Au cours de cette section, on va présenter quelques définitions liées à l'indexation des documents et des requêtes, les techniques d'indexation, le langage documentaire et le processus d'indexation.

### 2.1. Définitions

#### 2.1.1. Pertinence

Dans la recherche d'informations, le terme «pertinent» est utilisé pour représenter un élément contenant les informations nécessaires. En réalité, la définition de la pertinence n'est pas une classification binaire, mais une fonction continue. Du point de vue de l'utilisateur «pertinent» et «nécessaire» sont synonymes. D'un point de vue du système, l'information pourrait être pertinente à une instruction de recherche (c.-à-d : correspondant aux critères de l'instruction de recherche) même si elle n'est pas nécessaire/pertinente à l'utilisateur. [5]

### 2.1.2. Indexation

L'indexation est une étape très importante dans le processus de RI. Elle consiste à déterminer et extraire les mots les plus discriminants d'un document ou d'une requête qui couvrent au mieux leur contenu sémantique. La qualité de la recherche dépend en grande partie de la qualité de l'indexation.

Le résultat de l'indexation constitue, ce que l'on appeléle descripteur du document ou de requête. Ce dernier est souvent une liste de termes ou groupe de termes significatifs pour le texte correspondant, généralement assortis de poids représentant leur degré de représentativité du contenu sémantique de l'unité qu'ils décrivent.

Les descripteurs) sont rangés dans une structure appelée dictionnaire constituant le langage d'indexation.

On a 3 phases qui composent l'indexation :

- L'extraction des mots( termes)du document.
- La sélection des termes descriptifspour un document.
- La pondération des termes.

### 2.1.3. Système de recherche d'information

**Définition 1**

Un système de recherche d'informations est un système capable de stocker, de récupérer et de maintenir des informations. Les informations contenues dans ce contexte peuvent être composées de texte (y compris les données numériques et de date), d'images, d'audio, de vidéo et d'autres objets multimédia. Bien que la forme d'un objet dans un système de recherche d'informations est diversifiée, l'aspect texte a été le seul type de données qui se prêtait à un traitement fonctionnel complet. Les autres types de données ont été traités comme des sources très instructives, mais sont principalement liés à la recherche en fonction du texte.[5]

**Définition 2**

Un Système de Recherche d'Informations (SRI) est un système informatique qui permet de retourner à partir d'une collection de documents, ceux dont le contenu qui répondau besoin en informations d'un utilisateur, exprimé par une requête.

### 2.2. Techniques d'indexation

L'indexation peut se faire de 3 manières différentes : manuellement (faite par un humain), de manière semi-automatique , ou de manière automatique (créée par un programme informatique). [1]

#### 2.2.1. Indexation manuelle

Une personnedésigné les termes d'indexation : les descripteurs associés à chaque texte. Elle est basée sur un vocabulaire contrôlé (lexique, liste hiérarchiques, thésaurus, ontologie).

- Lexique : liste des mots clés.
- Liste hiérarchiques : des concepts et de la notation.
- Thésaurus : liste des mots clés + relations sémantiques entre ces mots clés.
- Ontologie : liste des concepts + relations entre ces concepts. [7]

#### 2.2.2. Indexation semi-automatique (ou supervisée)

Un programme détecte, pour chaque document, des descripteurs qui sont proposés à un utilisateur (qui peut valider, supprimer ou, parfois, ajouter des descripteurs)

#### 2.2.3. Indexation automatique (non supervisée)

Le programme fonctionne sans intervention humaine. L'indexation automatique a été crée afin de remédier aux problèmes liés aux approches précédentes.

Dans cette approche, le contenue des textes déterminé selon deux grandes méthodes d'analyse :

- Analyse linguistique : elle est basée sur la reconnaissance des mots.
- Analyse statistique : elle est basée sur la fréquence des mots. [7]

### 2.3. Langage documentaire

Un langage documentaire comporte les étapes suivantes qui sont nécessaires à l'indexation:

#### 2.3.1. Analyse documentaire

Cette étapepermet d'observer, d'identifier et de comprendre un textepourle rendre utilisable. Elle est basée surla reconnaissancedes concepts essentiels d'un document : **[**8]

Et qui se résument en :

- De quoi parle le document ?
- Comment peut-il être interrogé ?

### 2.3.2. Sélection des concepts les plus pertinents

Il faut développer les « concepts » retenus dans le document analysé pour pouvoir faire l'objet d'une recherche ultérieure. **[**8]

### 2.3.3. Formalisation de ces concepts en langage naturel

Il faut prendre en compte :

- L'ambiguïté liée aux « particularités »
  - Homonymie (une prononciation, plusieurs sens : un seau et un sot)
  - Synonymie
  - Polysémie
- La « variabilité » des formes lexicales :
  - Substantifs
  - Adjectifs
  - Verbes
- La structure des unités lexicales :
  - Unitermes
  - Mots composés

### 2.3.4. Réduction de ces concepts en langage documentaire

« Si l'indexation des documents est une opération complexe et incertaine, les langages documentaires ont été crées pour la simplifier, la guider et la normaliser » [9]

La fonction des langages documentaires est l'élimination des problèmes liés à la nature (langage naturel). C'est pourquoi, les langages documentaires sont désignés sous le terme de «vocabulaire contrôlé ». **[**8]. L'opération consiste alors à choisir le ou les descripteurs, qui permettent d'obtenir le meilleure « transcription » du concept retenu.

En représentant le contenu d'un document primaire sous la forme d'une liste de descripteurs, le documentaliste produit un deuxième document, plus facile à consulter, donc une utilité de document secondaire visible pour l'aide à la sélection. [8]

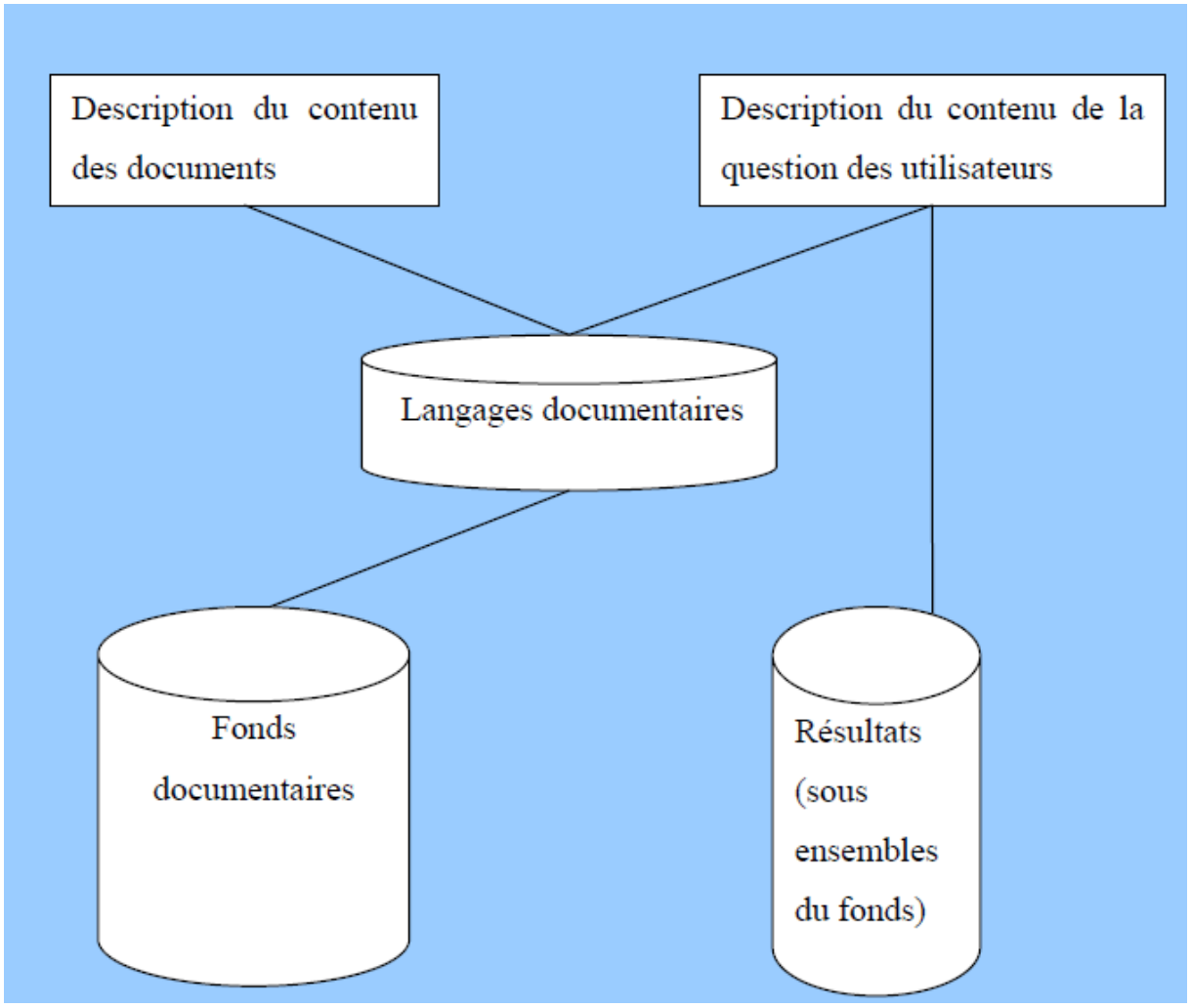


**Fig. 1.Les langages documentaires.[8]**

### 2.4. Processus de normalisation

Ce processus peut contenir plusieurs étapes, les plus importantes sont les suivantes:

#### 2.4.1. Elimination des mots vides

Les mots vides sont des mots qui ne portent aucun sens (les articles, les conjonctions de coordination, les verbes auxiliaires, etc. Ils ne peuvent pas être indexés, il faut donc les éliminer. [10]

### 2.4.2. Racinisation

Elle consiste à tronquer un mot de telle sorte qu'à partir du mot résultant on peut reconstruire ses différentes variantes morphologiques. Il existe des algorithmes qui permettent de réaliser cette opération, comme l'algorithme de Porter. Le problème dans cette étape est que le sens peut-être perdu dans le cas ou la racine extraite peut être commune à des mots se rapportant à des concepts différents. On peut voire ceci dans l'exemple des mots : ***port***, ***port***es et ***port***era qui ont la même racine ***port*** mais qui se rapportent à trois concepts différents. [10]

### 2.4.3. Lemmatisation

Les mots peuvent être classés en deux catégories : Lemmes et formes canoniques. (Infinitif pour les verbes, singulier pour les noms, etc.)

Les mots obtenus par flexion de ces lemmes : conjugaison d'un verbe, changement de genre ou de nombre, etc. Par exemple, le mot « ***devrait*** » est obtenu par flexion du verbe « ***devoir*** ». La lemmatisation consiste à remplacer un mot par son lemme, les mots ***étudiant*s** et **étudiez** seront remplacés par leurs lemmes : ***étudient*** où ***étudier*** selon le contexte et *porter***.** Cette opération est plus couteuse que la racinisation parce qu'elle besoin d'une analyse morphologique et syntaxique des phrases. [10]

## 2.5. Processus d'indexation

Le processus de l'indexation permet de transformer l'information contenue dans un document vers un autre espace de représentation traitable par un système de recherche d'information , le processus nous renvois une liste d'index a partir d'une ensemble de documents , on utilise cette liste résultat pour les systèmes de recherche et aussi dans différentes applications comme :comparaison et classification des documents, proposition des mots-clés, faire une synthèse automatique de documents …..[11].

Le processus d'indexation peut être représenté par le schéma suivant :

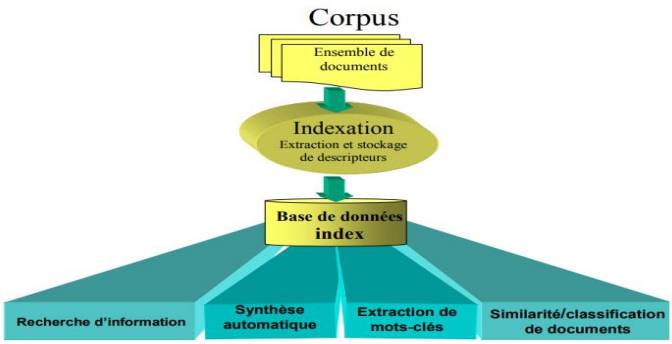


**Fig . 2. Processus d'indexation [11]**

### 2.5.1. Etapes de processus d'indexation

Le processus d'indexation se compose de plusieurs étapes, on peut les schématiser comme suit :

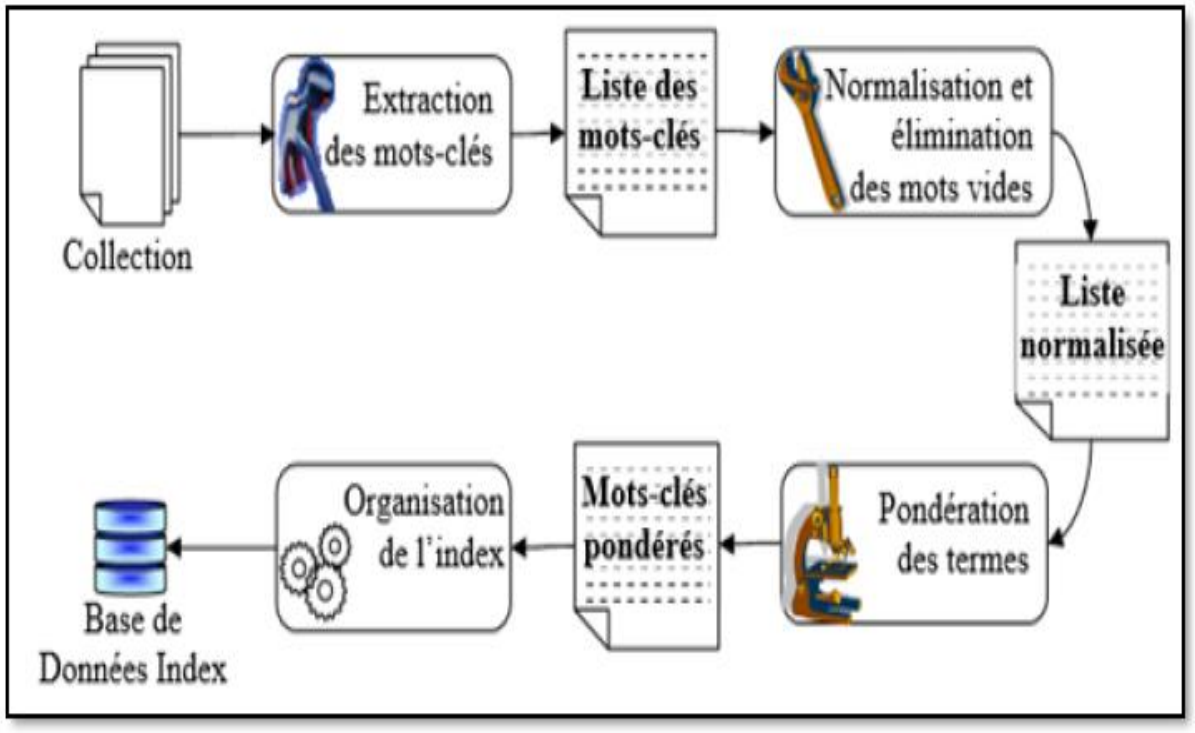


**Fig .3.Les étapes de processus d'indexation [12]**

**Etape 1**

**Extraction des mots clés appelée Tokenization en anglais**

Cette phase représente la segmentation des documents en unités, qui est basée généralement sur la ponctuation et sur une liste de séparateurs, le résultat de cette étape est un ensemble de mots. [10]

**Etape 2**

**La normalisation des mots-clés du document** ( elle contient plusieurs étapes , qu'on a expliquées dans " le processus de normalisation" ) .

**Etape 3**

Dans cette phase on utilise une approche de sélection des index et les pondérer.

# 3. Les modèles de recherche d'information

**Définition**

Un modèle de recherche représente le modèle du noyau d'un SRI. Il comprend la fonction de décision fondamentale qui permet de l'associer à une requête, l'ensemble des documents pertinents à restituer. Il est utilisé pour la recherche d'informations proprement dite et est étroitement lié au modèle de représentation des documents et des requêtes.

On a trois modèles classiques :

## 3.1. Modèle booléen

### 3.1.1 Présentation du modèle booléen

Le modèle booléen de recherche d'information est un modèle qui est basé sur la théorie classique d'ensemble. Les documents sont représentés comme un ensemble de termes qu'ils contiennent, (les mots ne doivent pas être utilisés tous), tandis que les requêtes sont représentées comme des expressions logiques. Les mots-clés dans la requête peuvent être liés entre eux avec les opérateurs booléens et, ou et non. Chaque terme peut avoir l'un des deux états logiques, il peut être présent (logique 1) ou absent (logique 0). [14]

La pertinence d'un document pour la requête d'un utilisateur est calculée en évaluant la valeur logique de la requête comme étant 1 ou 0. Une valeur de 1 est donnée à chaque terme dans la requête qui existe dans l'ensemble représentant le document, et 0 pour chaque terme qui n'existe pas dans la représentation du document. [14]

### 3.1.2Représentation des documents dans le modèle booléen

Dans la recherche d'information selon ce modèle, chaque document de la base de données doit être présenté comme un ensemble de termes. Afin de limiter la taille de chaque représentation, tous les mots ne doivent pas être stockés. Par contre un dictionnaire (ensemble) de mots intéressants est créé. Selon le but de la base de données le dictionnaire peut être réduit et ne contient que des mots pour un domaine spécifique.

Pendant le processus d'indexation, chaque document est comparé à l'ensemble des termes intéressants pour créer la représentation vectorielle. Si le dictionnaire de termes est créé en tant que vecteur contenant des mots, chaque document peut être représenté par un vecteur des uns et des zéros. La taille du vecteur doit être la même que celle du dictionnaire et pour chaque mot dans le dictionnaire, s'il est pertinent pour le document, alors le vecteur de

représentation de document contiendra 1 sur la même position que le mot sinon il contiendra 0 pour cette position. [14]

| | **Terme 1** | **Terme2** | **Terme 3** | **…** | **Terme M** |
|---|---|---|---|---|---|
| **Dictionnaire** | KPU | Université | Cours | … | Bibliothèque |
| **Doc. 1** | 1 | 0 | 0 | … | 1 |
| **Doc. 2** | 1 | 1 | 0 | … | 0 |
| **Doc. 3** | 1 | 1 | 0 | … | 1 |
| **…** | … | … | … | … | … |
| **Doc. N** | 0 | 0 | 0 | … | 1 |

**Fig.4. Exemple de représentation de documents pour le modèle booléen**

### 3.1.3. Détermination de la pertinence du document dans le modèle booléen

Afin de déterminer la pertinence d'un document pour la requête, la valeur logique de la requête doit être évaluée. Chaque terme de la requête a une valeur logique 1 si le mot existe dans le document (ou sa représentation) et la valeur logique 0 si elle n'existe pas . Une fois que tous les termes de la requête sont remplacés par des valeurs logiques, la requête peut être évaluée comme n'importe quelle phrase logique. Si le résultat d'évaluation est vrai, le document est considéré comme pertinent. [14]

## 3.2. Modèle vectoriel

### 3.2.1. présentation du modèle vectoriel

Dans ce modèle la représentation des document set les requêtes sont formés comme vecteurs dans un espace multidimensionnels c-à-d. :

Un document **d** est un vecteur représenté comme suite :

$$\vec{d} = \overrightarrow{w_{1j}}, \overrightarrow{w_{2j}}, \dots\dots\dots, \overrightarrow{w_{nj}} \quad \text{……(1)}$$

Où :

- $\vec{\mathbf{d}}$= est le document (ensemble des termes )
- $\overrightarrow{\boldsymbol{\omega}}$nj = est le poids du terme dans le document dj.

Une requête q est un vecteur représenté comme suite :

$$\vec{q} = \overrightarrow{w_{1q}}, \overrightarrow{w_{2q}}, \dots\dots\dots, \overrightarrow{w_{nq}} \quad \text{……(2)}$$

Où :

- **ω**qjest le poids du terme dans la requête qj.
- qj est la requête donnée

### 3.2.2.Fonction de pondération

La fonction de pondération la plus répandue est celle de Sparck Jones & Needham [13]

$$\vec{w} = \overrightarrow{tf_{ij}} \times \overrightarrow{idf_{ij}} \quad \text{.....(3)}$$

Où :

- tfji : Décrit le pouvoir descriptif du terme ti dans le document Dj
- idfi : Décrit le degré de généralité du terme ti dans la collection

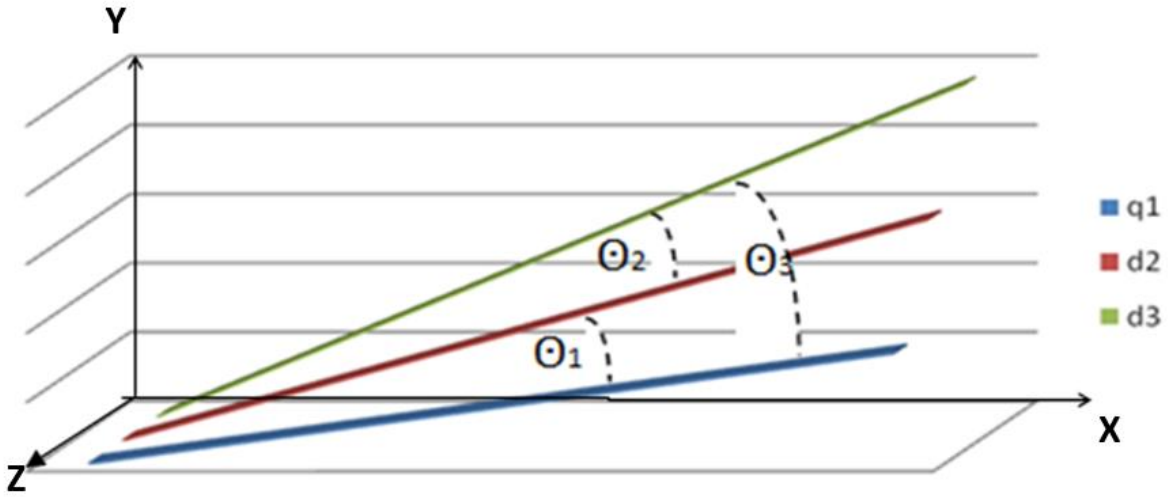


**Fig.5. Représentation d'espace vectoriel (Multidimensionnels)**

**Discussion figure 5**

''L'angle $\theta$ entre deux vecteurs est considéré comme une mesure de divergence entre les vecteurs, l'angle cosinus est utilisé pour calculer la similarité déterminée par l'angle entre le vecteur de document et le vecteur de requête qui sont représentés en V-dimensionnelle Espace euclidien.''[13]

### 3.2.3.Fonction de similarité

-Pour la mesure de similarité on définit l'angle dit "$\theta$"entre le vecteur de la requête **q** et les vecteurs des documents rapportés par le système, la similarité est mesurée comme le cosinus cos ($\theta$).

Alors, la similarité est variée dans un intervalle de [0,1], et par conséquence le vecteur de la requête qest plus proche au vecteur d'un document **d** on dit qu'il est plus Similaire.

-La fonction de similarité entre les documents vecteurs Di et la requête Q est donnée par :

La mesure du Salton : [10]

$$RSV(Q_k, D_j) = \frac{\sum_{i=1}^{T} q_{ki}\, d_{ji}}{\left(\sum_{i=1}^{T} q_{ki}^{\,2}\right)^{1/2}\left(\sum_{i=1}^{T} d_{ji}^{\,2}\right)^{1/2}} \quad \text{.....(4)}$$

La mesure de Jaccard : [10]

$$RSV(Q_k, D_j) = \frac{\sum_{i=1}^{T} q_{ki}\, d_{ji}}{\sum_{i=1}^{T}(d_{ji})^2 + \sum_{i=1}^{T}(q_{ki})^2 + \sum_{i=1}^{T} q_{ki}\, d_{ji}} \quad \text{..... (5)}$$

La mesure de Singhal et al :[10]

$$RSV(Q_k, D_j) = \frac{\sum_{i=1}^{T} q_{ki}\, d_{ji}}{(1-s) + s * \frac{\sqrt{\sum_{k=1}^{N} d_{kj}^{\,2}}}{|D_j|}} \quad \text{..... (6)}$$

- Où : |Dj| : Longueur du document Dj
- s : Constante

- La procédure du modèle spatial peut être divisée en trois étapes.

- La première étape est l'indexation de document où les termes porteurs de contenu sont extraits du texte du document.
- La deuxième étape est la pondération des termes indexés pour améliorer la récupération des documents pertinents pour l'utilisateur.
- Dans la dernière étape, classer les documents par rapport à la requête selon des valeurs de similarité. [13]

### 3.3. Modèle probabiliste

Les résultats récupérés par les systèmes de recherche d'information probabiliste, dépendent des estimations et des probabilités. La première hypothèse est que les termes sont disperser différemment entre les documents pertinents et non pertinents. Un système probabiliste classe les documents et les trier en ordre décroissant de probabilité de pertinence pour l'information nécessaire une fois que la probabilité est calculée. Les résultats sont aussi précis que la probabilité calculée. [14]

### 3.3.1.Le modèle probabiliste de base

- Maron et Kuhun sont les deux personnes qui mettent ses bases, la similarité entre un document d et une requête **q** est calculée par une fonction de probabilité dit que le document **d** est pertinent ou non par rapport à une requête q écrite comme :

- si le d (document) est pertinent P=(d/q)
- sinon d est non pertinent P= $(\overline{d}/q)$[25]

La probabilité estimée par la probabilité conditionnelle ou le terme de la requête existe dans le document **d** ou non.

La formule donnée dans l'état général sous-dessous, confirme la mesure de l'existence d'un terme dans une requête q sachant que le document correspond est pertinent ou non. [3]

$$RSV(Q,d) = \frac{p(d/Q)}{p(\bar{d}/Q)} = \sum_{i=1}^{t} \log \frac{p(1-q)}{q(1-p)} \qquad \text{……(7)}$$

Où :p : est **P**(terme **ti** existe/**d** pertinent)

q : est P(terme ti présent/d non pertinent).

t : le nombre total de terme dans la requête

### 3.3.2Le modèle probabiliste BIR(Binary Independance Retrieval )

Dans ce modèle les documents doivent être représentés comme un vecteur d'ensembles des événements :

dE=(E1,E2,E3,……En)

Où : dE : est le document

En : l'ensemble des événements possibles

L'événement **E** ici dénote la présence ou l'absence de terme dans un document **d**, de manière indépendante.[3]

Par la définition précédente on suppose que le document est une liste de termes, alors :

$$RSV(p.q) = \frac{P(\frac{d}{R})}{P(\frac{d}{NR})} = \frac{P((\mathrm{E1},\mathrm{E2},\ldots\mathrm{En})/\mathrm{R})}{P(\frac{(E1,E2,\ldots,En)}{NR})} \qquad \text{……(8)}$$

# 4. Evaluation d'un modèle de recherche d'information

L'un des défis de la recherche d'information moderne est d'évaluer adéquatement le système de recherche d'information (SRI) afin d'estimer les performances futures dans un domaine d'application spécifique. On peut citerplusieurs méthodespour évaluer le système de recherche d'information

- ✓ Mesures de rappel / Précision
- ✓ Mesures combinées
- ✓ Collection TREC

## 4.1. Mesures de rappel / Précision

Les deux principales mesures associées aux systèmes d'information sont la précision et le rappel. Lorsqu'un utilisateur décide d'émettre une recherche sur un sujet, la base de données totale est logiquement divisée en quatre segments (Fig.6) les éléments pertinents sont les documents qui contiennent des informations qui aident le chercheur à répondre à sa question. Les éléments non pertinents sont les éléments qui ne fournissent aucune information directement utile. Il existe deux possibilités par rapport à chaque élément: il peut être rapporté ou non rapporté par la requête de l'utilisateur. La précision et le rappel sont définis comme:[5]

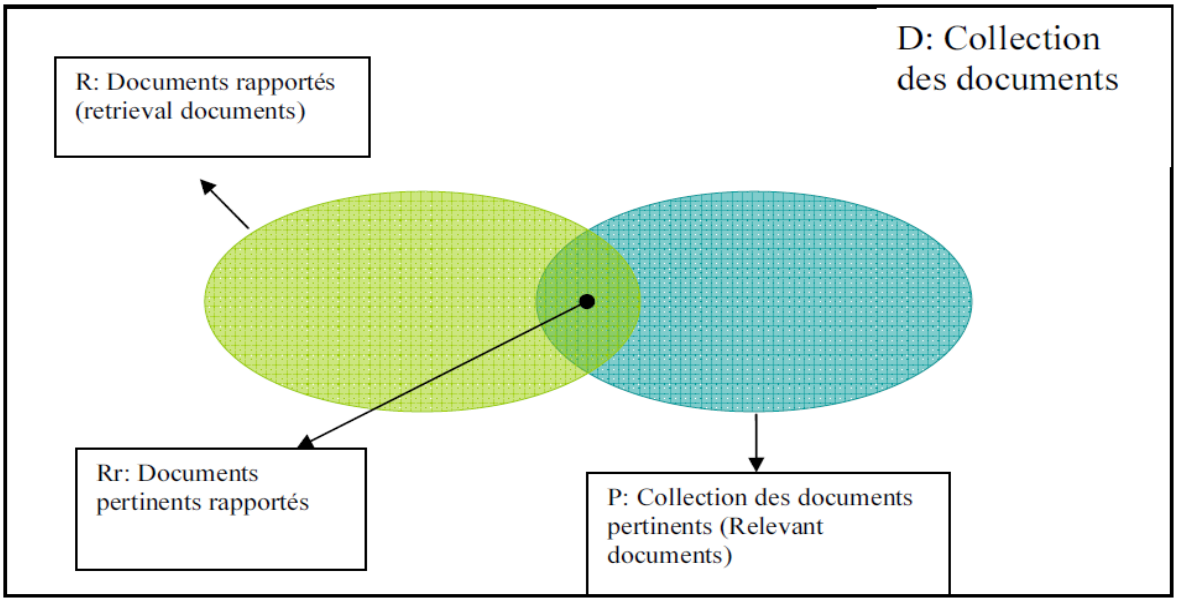


**Fig.6. segmentation des documents après une requête d'utilisateur**

Alors ,

$$\text{Précision} = \frac{\text{documents pertinents rapportés}}{\text{documents rapportés}}$$

$$\text{Rappel} = \frac{\text{documents pertinents rapportés}}{\text{documents pertinents dans la base}}$$

On a deux autres notions :

**Bruit =1 – Précision**

**Silence = 1- Rappel**

Tel que : le bruit est l'existence des documents pertinents non rapportés

le silence est l'existence des documents non pertinents rapporté par le système

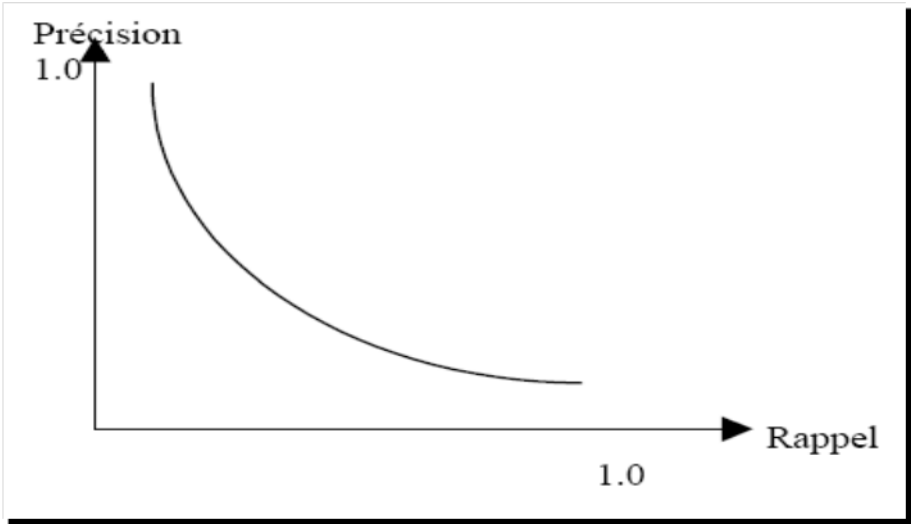


**Fig.7.Courbe générale de précision/rappel**

### 5.2. Mesures combinées

Cette idée est principalement motivée par [Korfhage, 1997]. Elle consiste à combiner les mesures standards de rappel/précision .Elle est motivée par les points suivants : [26]

- La difficulté de calcul du rappel maximal dans les collections volumineuses.

-l'incompatibilité de ces mesures dans le cas où la fonction d'appariement n'est pas une fonction d'ordre faible,

- La nécessité de combiner les deux aspects rappel/précision.

### 5.3. Collection TREC (Text REtrieval Conference )

Le programme TREC est un programme international créé au début des années 90 par le NIST (National Institute of Standards and Technology) et du DARPA (Defense Advanced Reserach Projet Agency). Ce programme aide d'évaluer des systèmes de recherche d'information. Il est devenu la référence en recherche d'information. Il a permis de définir les tâches et construire de larges collections de test dans la recherche d'information. [10]

Les différents éléments qui constituent le projet TREC sont :

1 **Tâches** : plusieurs tâches sont définies chaque année. L'objectif est de permettre l'évaluation d'approches spécifiques en recherche d'information concernant le filtrage, la recherche dans de très larges corpus (100 giga octet et plus) et les modèles d'interactions. [10]

2 **Les participants** : les participants dans ce projet sont 25 groupes qui ont participé à la première édition de TREC en 1992 et 66 groupes de 16 pays différents ont également participé à TREC8. [10]

3 **Source d'information :** Les documents sont tirés de la presse écrite en 1999 (Financial Time, Résumés de publication USDOE, SAN jose Mercury news, etc.). . [10]

**Structure et principe de construction de la collection** : un document TREC est soumis sous le format SGML. Il est identifié par un numéro et décrit par un auteur, une date de production et un contenu textuel. Une requête TREC est également identifiée par un numéro. Elle est décrite par un sujet, aussi une description sur les caractéristiques des documents pertinents associés à la requête. . [10]

## 6. Conclusion:

Dans ce chapitre nous avons présenté les principales notions et concepts de la recherche d'information et ses systèmes.

À travers les différentes sections que nous avons présentées, nous concluons que la recherche d'information, s'attache à définir des modèles et des systèmes afin de faciliter l'accès à un ensemble de documents se trouvant dans des bases documentaires ou encore sur le web. Le but est de permettre aux utilisateurs de retrouver les documents dont le contenu répond à leur besoin en information. Nous nous intéressons dans le cadre de cette mémoire à deux orientations en RI : d'abord, la comparaison de la pertinence de documents entre le modèle de recherche d'information vectoriel et le modèle probabiliste. Puis, l'intégration de l'apprentissage automatique dans ce domaine. Et c'est le prochain chapitre quitraiterales notions et les concepts de base de l'apprentissage automatique.

# Chapitre II

## Apprentissage automatique

# 1. Introduction

Le domaine de la recherche d'information est devenu indispensable pour la vie humaine. Pour améliorer le rendement et pour assurer l'efficacité des systèmes de recherches, il faut faire intervenir des nouvelles techniques, parmi lesquelles celles de l'apprentissage automatique.

## 1.1. Définition

L'apprentissage automatique (machine learning) est un champ de l'intelligence artificielle (IA). Il représente la discipline d'étudier la possibilité aux machines ou systèmes algorithmiques d'atteindre des connaissances par des expériences selon des lois et des techniques prédéfinits.


**Fig.8. schéma généralde l'apprentissage automatique [27]**

# 2. Historique

La 1ere année d'apparition de l'apprentissage automatique est 1935. Il a commencé avec les travaux du mathématicien Alan Turing, puis il a suivi le cycle suivant :

- Années 50
  - ✓ The Turing Test (1950)

Alan Turing, a créé le "test de Turing" pour déterminer si un ordinateur a une réelle intelligence. Pour réussir le test, un ordinateur doit être capable de tromper un humain en croyant qu'il est aussi humain.[15]

Selon ce type de test, un ordinateur est considéré comme ayant une intelligence artificielle si elle peut imiter les réponses humaines dans des conditions spécifiques.

- ✓ premier programme d'apprentissage d'ordinateur (1952) :

En 1952, Arthur Samuel a écrit le premier programme d'apprentissage. Le programme a été le jeu de dames, et l'ordinateur amélioré au jeu plus il a joué, l'étude qui se déplace a constitué des stratégies gagnantes dans un «mode d'apprentissage supervisé» et incorporant ces mouvements dans son programme. [15]

✓ The Perceptron ( 1957 )

Frank Rosenblatt a conçu le Perceptron qui est un type de réseau de neurone. Un réseau de neurone agit comme votre cerveau; le cerveau contient des milliards de cellules appelées neurones qui sont connectés ensemble dans un réseau. Le Perceptron relie un réseau de points où des décisions simples sont prises qui se réunissent dans le programme plus grand pour résoudre des problèmes plus complexes. [15 ]

- 1967 - Reconnaissance des motifs

L'algorithme "voisin le plus proche" a été écrit, permettant aux ordinateurs de commencer à utiliser la reconnaissance de modèle très basique. Lorsque le programme a reçu un nouvel objet, il l'a comparé avec les données existantes et l'a classé au voisin le plus proche, ce qui signifie l'objet le plus semblable en mémoire. [15]

- 1979- Le panier Stanford

Les étudiants de l'Université de Stanford ont inventé le "Stanford Cart" qui peut naviguer dans une pièce seule. Le Stanford Cart était un robot mobile télécommandé équipé d'une télévision. [15]

- 1981- Apprentissage basé sur l'explication

Gerald Dejong a introduit l'apprentissage fondé sur l'explication (EBL) dans un article publié en 1981. En EBL, la connaissance préalable du monde est fournie par des exemples de formation qui en font un type d'apprentissage supervisé. [15]

- Dans les années 1990

Les scientifiques ont commencé à appliquer l'apprentissage des machines dans l'exploration de données, les logiciels adaptatifs et les applications Web, l'apprentissage des textes et l'apprentissage des langues. Ils commencent à créer des programmes pour les ordinateurs pour analyser de grandes quantités de données et de tirer des conclusions-ou «apprendre»-à partir des résultats. [15]

## 3. Domaines d'apprentissage automatique

Aujourd'hui, l'apprentissage automatique est utilisé dans toutes nos applications quotidiennes. On le trouve sur TVs , sur téléphones portables etc.…

Ces domaines d'application son divers, on peut citer par exemple :

- Reconnaissance des vois, motifs et schémas
- Reconnaissance des langages naturels et des formes syntaxiques
- Moteurs de recherches.
- détection de fraudes.
- classification des séquences d'ADN.

- Imagerie et diagnostic médical
- bioinformatique
- interfaces cerveau-machine
- Analyse financière
- Jeux

## Exemples

❖ Un système d'apprentissage automatique peut permettre à un robot, ayant la capacité de bouger ses membres mais ne sachant initialement rien de la coordination des mouvements permettant la marche, peut lui permettre d'apprendre à marcher. Le robot commencera par effectuer des mouvements aléatoires, puis, en sélectionnant et privilégiant les mouvements lui permettant d'avancer, mettra peu à peu en place une marche de plus en plus efficace.

❖ La reconnaissance de caractères manuscrits est une tâche complexe car deux caractères similaires ne sont jamais exactement égaux. On peut concevoir un système d'apprentissage automatique qui apprend à reconnaître des caractères en observant des « exemples », c'est-à-dire des caractères connus.

## 4. Types d'apprentissage automatique

Plusieurs types d'apprentissage automatique sont apparus. Chaque type est caractérisé par son objectif et les besoins de son système réalisé. Dans cette section, on va parler de l'apprentissage supervisé, non supervisé, semi supervisé, par renforcement, par transfert et profond.

### 4.1. Apprentissage supervisé

L'apprentissage supervisé est assez fréquent dans les problèmes de classification parce que l'objectif est souvent d'amener l'ordinateur à apprendre un système de classification que nous avons créé. Plus généralement, l'apprentissage de la classification est approprié pour tout problème où la déduction d'une classification est utile et la la classification est facile à déterminer [16]

L'apprentissage supervisé laisse souvent la probabilité pour les entrées non définies. Ce modèle n'est pas nécessaire tant que les entrées sont disponibles, mais si certaines des valeurs d'entrée sont manquantes, il n'est pas possible de déduire quoi que ce soit sur les sorties .[16 ]

Dans le domaine de l'apprentissage supervisé qui traite beaucoup avec la classification. Ce sont les types d'algorithmes: [16]

- Classifieurs linéaires
  - ✓ régression logique
  - ✓ classifieur bayésien naïf

- ✓ Perceptron
- ✓ Support machine vecteur

- Classifieurs quadratiques
- K-means clustering
- Booster
- Arbre décisionnel
  - ✓ forêt aléatoire
- Réseaux de neurones
- Réseaux bayésiens

**4.2. Apprentissage non supervisé**

L'apprentissage non supervisé semble beaucoup plus difficile: le but est de faire en sorte que l'ordinateur apprenne à faire quelque chose que nous ne lui disons pas comment faire! Il y a en fait deux approches de l'apprentissage non supervisé. La première approche consiste à enseigner l'agent non pas en donnant des catégorisations explicites, mais en utilisant une sorte de système de récompense pour indiquer le succès. Notez que ce type de formation s'intégrera généralement dans le cadre du problème de décision parce que l'objectif est de ne pas produire une classification, mais de prendre des décisions qui maximisent les récompenses. Cette approche généralise bien au monde réel, où les agents pourraient être récompensés pour faire certaines actions et punis pour faire les autres. [16]

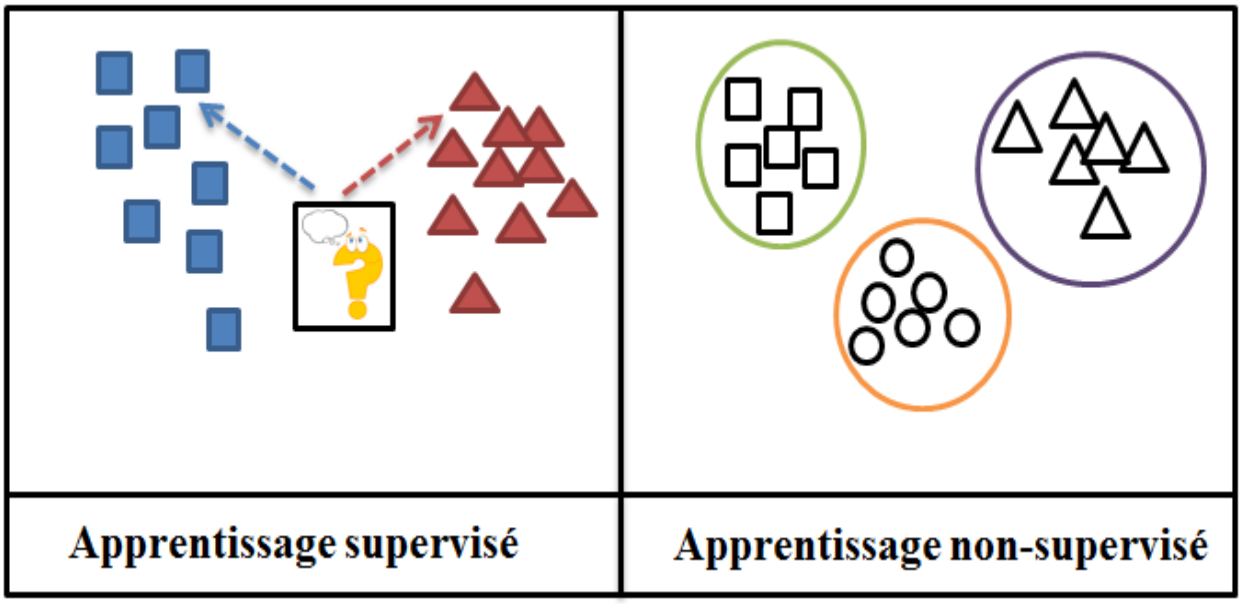


**Fig.9. apprentissage supervisé / apprentissage non supervisé[16]**

### 4.3. Apprentissage semi supervisé

Ce type a attiré l'attention de plus en plus récemment. Il est défini entre l'apprentissage supervisé et non supervisé, contient des données étiquetées et non marquées, et apprend conjointement les connaissances d'eux. [16]

### 4.4. Apprentissage par renforcement

L'apprentissage par renforcement est utilisé pour résoudre les problèmes de prise de décision (généralement une séquence de décisions), comme la perception et le mouvement des robots, le joueur d'échecs automatique et la conduite automatique des véhicules. Cette catégorie d'apprentissage ne sera pas discutée plus loin dans cette thèse, et les lecteurs pourraient se référer à [16]

## 5. Processus d'apprentissage automatique

Pour construire un modèle d'apprentissage bien défini et solide, il faut passer par des étapes de préparation suivantes :

1) Fixer le problème : Elle consiste à définir l'objectif à atteindre.

2) Collecte de données: les données peuvent être écrites sur papier, texte enregistré fichiers et feuilles de calcul ou stockés dans une base de données SQL. Données doivent être rassemblés dans un format électronique adapté Analyse.

3) Préparation des données :

   La qualité de tout projet d'apprentissage machine est basée sur la qualité des données qu'elle utilise. Il est suggérer que 80% de l'effort en apprentissage automatique soit consacré à la préparation des données. Cette étape exige une grande partie de l'homme

   Intervention. Cette étape consiste en:

   - Nettoyage de la qualité des données en éliminant les données non significatives car dans l'état réel les données peuvent être manquantes, non propres, et même erronées.
   - Désignation des données en fixant les données en sortie et les données en entrée.

4) Apprentissage : La tâche spécifique d'apprentissage de machine informera la sélection d'un algorithme approprié. Nous «alimentons» ensuite les données au modèle pendant cette phase et nous obtiendrons un apprenant. Un apprenant est un algorithme

d'apprentissage automatique qui a été formé sur certaines données et ajusté pour adapter les données le mieux possible.

5) Évaluation de la performance du modèle:

parce que chaque apprenant dans une solution biaisée, Il est important d'évaluer la façon dont l'algorithme a appris de son expérience. Selon le modèle utilisé, nous pourrions être en mesure d'évaluer la précision de l'apprenant à l'aide d'un test dataset .

6) Amélioration des performances du modèle: si de meilleures performances sont nécessaires, il devient nécessaire d'utiliser des stratégies plus avancées pour améliorer les performances du modèle, ou passer à un modèle différent, compléter avec des données supplémentaires et effectuer une préparation supplémentaire travail sur les données .

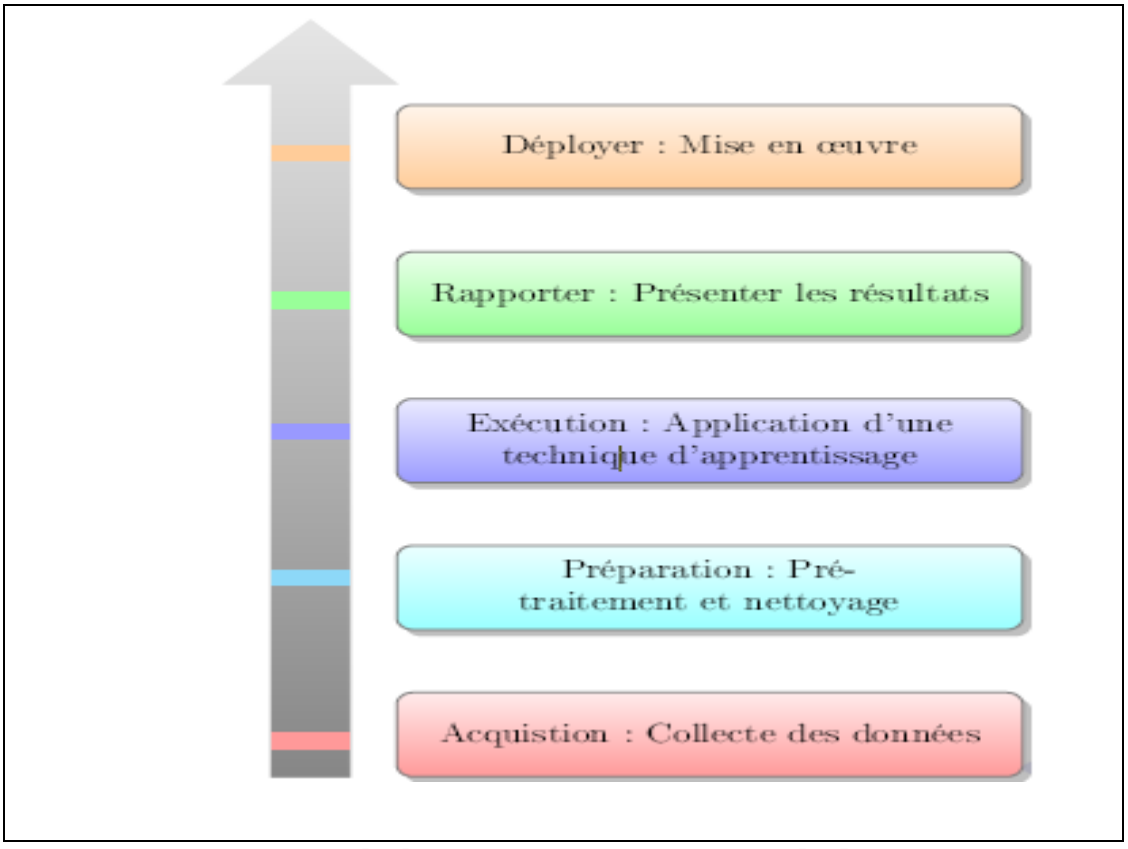


**Fig.10. Cycle d'apprentissage automatique [27]**

## 6. Conclusion

Dans un intervalle plus vaste, les méthodes et les techniques de l'apprentissage automatique fournissent des possibilités capables d'améliorer et d'augmenter l'efficacité et la performance des systèmes qui permettre de retourner l'information désirée par l'utilisateur.

# Chapitre III

## Intégration de l’apprentissage automatique dans la RI

# 1.Introduction

Dans les chapitres précédents, nous avons présenté les concepts et les notions de base de recherche d'information, et les concepts de base de l'apprentissage automatique, dans ce chapitre, on parlera de l'intégration de l'apprentissage automatique dans la recherche d'information.

# 2. Apprentissage dans la reformulation de requête

Lorsqu'un utilisateur trouve des difficultés pour exprimer son besoin exact en information, les systèmes de recherche d'information retournent certains documents qui ne satisfont pas son besoin.

Donc, pour résoudre ce problème et pour retourner à l'utilisateur des documents plus pertinents par rapport à sa requête une amélioration est utilisée qui s'appelle **la reformulation de requête** initiale.

**Définition**

"La reformulation de la requête est la procédure qui permet d'organiser les termes des requêtes dans une structure autre que le sac de mots, en utilisant des fonctions, comme la proximité et la pondération, fournies par le modèle de recherche."[23]

La reformulation de requête est faite par un processus.

## 2.1. Processus de reformulation de requête

C'est un processus évolutif et interactif. Il consiste à utiliser la requête initiale pour commencer la rechercher. Pour modifier les termes de la requête initiale en considérant la pertinence et/ou la non-pertinence de documents, le processus doit ré-pondérer les termes de la requête initiale pour y ajouter (ou supprimer) d'autres termes. La requête obtenue permet d'orienter la recherche pour le sens des documents pertinents.

On peut schématiser le processus de reformulation comme suite : [21]

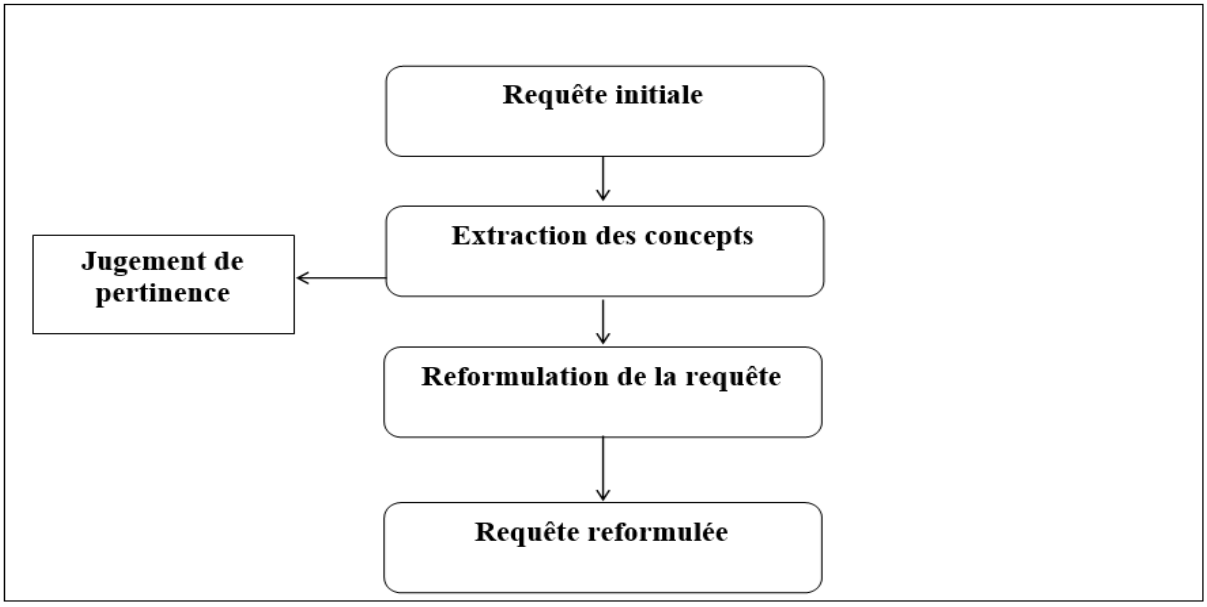


**Fig.11. Processus de reformulation de la requête utilisateur**

### 2.2. Les techniques de reformulation de requête

Plusieurs études ont traité les modèles de reformulation des requêtes sur le Web. Nous pouvons les classer dans les groupes suivants:

#### 2.2.1. Systèmes de reformulation de requêtes basés sur le profil utilisateur

Un profil est un modèle utilisateur qui spécifie le domaine d'intérêt de l'utilisateur et les préférences les plus générales qui distinguent cet utilisateur des autres. Toutes les requêtes émises par le même utilisateur sont évaluées par rapport à son profil spécifique. La même requête émise par différents utilisateurs peut avoir des résultats différents car elle est évaluée à l'aide de profils différents. [22]

#### 2.2.2. Reformulation des requêtes par réinjection de pertinence

La méthode de Rocchio est un algorithme classique pour la réinjection de pertinence. La reformulation des requêtes en utilisant la réinjection de pertinence est un processus interactif, dirigé par l'utilisateur avec l'objectif de générer une nouvelle requête plus appropriée que celle initialement exprimée par l'utilisateur. Son principe fondamental est d'utiliser la requête initiale afin de commencer la recherche, puis de la modifier à partir de jugements de pertinence et/ou aucune pertinence par l'utilisateur. La nouvelle plainte obtenue dans chaque réinjection d'itération, peut rectifier la direction de la recherche dans le sens des documents pertinents dans le sens exprimé explicitement par l'utilisateur (Baeza-Yates et coll., 2004). [22]

### 2.2.3. Dés ambiguïté de Requête

Les techniques de dés ambiguïté visent à identifier précisément le sens visé par les termes de la requête et à se concentrer sur les documents contenant les mots cités dans le contexte défini par le sens correspondant. [22]

### 2.2.4. Expansion de requête utilisant des ressources externes de termes

On Commence cette section par définir les deux termes : thésaurus et ontologie.

**Thésaurus**

Le thésaurus est une liste organisée de termes contrôlés et normalisés (descripteurs et non descripteurs) représentant les concepts d'un domaine de la connaissance.

**Ontologie**

L'ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances.

L'expansion de requête est le processus d'ajout de termes à la requête d'origine afin d'améliorer les résultats en incluant des termes qui conduiraient à retourner des documents plus pertinents. Toutefois, dans un sens plus général, il se réfère également à des méthodes de reformulation de requête, c'est à dire, tout type de transformation appliquée à une requête pour faciliter une récupération plus efficace. Dans ce groupe d'approches, la requête initiale est développée en utilisant des ressources externes de termes, tels que thésaurus ou ontologie, qui contiennent le vocabulaire utilisé dans l'enrichissement de la requête. [22]

Beaucoup d'approches comme (Storey et coll., 2004) essaient de reformuler les requêtes Web basées sur une connaissance sémantique de différents domaines d'application de la recherche par exemple, d'autres utilisent l'information de sens (WordNet en général) pour étendre la requête.[30]

De nombreuses approches, développent la requête initiale de l'utilisateur à l'aide d'une ontologie afin d'extraire le domaine sémantique d'un mot et d'ajouter les termes associés à la requête initiale. [22]

### 2.3. Reformulation utilisant l'apprentissage

Il existe des approches qui sont utilisées pour assurer l'expansion de requête et qui sont basés sur l'apprentissage. Parmi ces approches on peut citer :

### 2.3.1. Apprentissage de l'expansion de requêtes par règles d'association

Le système utilisé dans cette approche est formé de deux composants principaux (Fig. 12). Le premier composant est consacré à la construction des exemples d'entraînement. Le deuxième composant a le rôle de construire un modèle de prédiction à partir des exemples d'apprentissages et en utilisant un algorithme de classification supervisée. Par exemple pour une requête 'q' et l'ensemble des règles d'association correspondant 'RAq', le modèle de prédiction identifie 'RA+q' qui représente le sous ensemble de règles d'association qui sont à utiliser pour générer la requête étendue de 'q' et qui est 'Eq' [24]

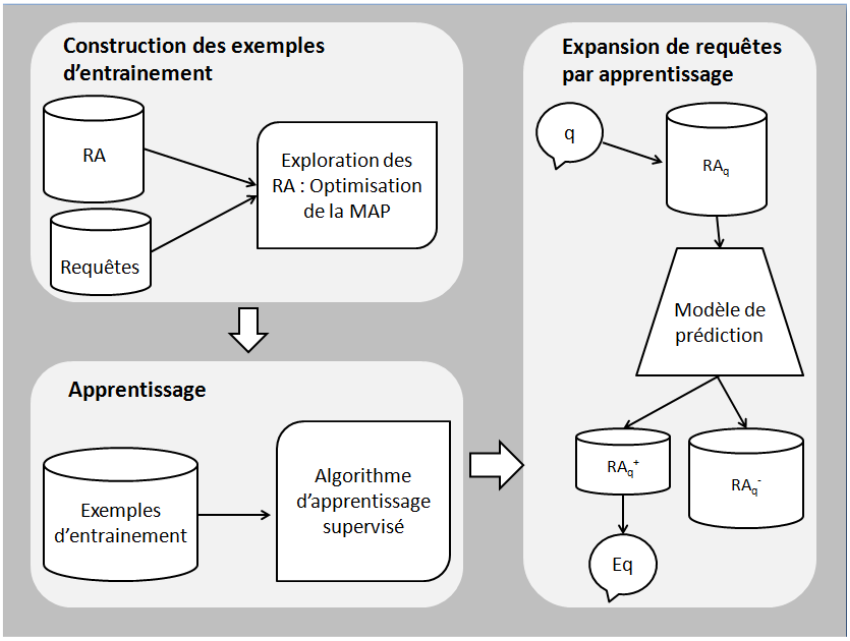


**Fig. 12 Schéma de l'approche**

## 3. Conclusion

Comme nous l'avons vu dans cette partie, nous concluons que l'apprentissage automatique enrichit la recherche d'information et ils se complètent. Le chapitre suivant contient la partie pratique de ce projet (la comparaison, la réalisation Del 'apprentissage automatique dans la recherche d'information).

# Chapitre IV

## Expérimentation et Évaluation

## 1. Introduction

Notre objectif dans ce chapitre est de comparer deux modèles de recherche d'information classiques : (probabiliste et vectoriel). Le premier système choisi est 'VSM' qui est un système basé sur le modèle vectoriel, alors que le deuxième modèle est BM25 qui est un modèle probabiliste. En plus on comparé notre propre modèle.

## 2. Outils de développement

Les outils de développement utilisés pour atteindre notre objectif sont résumés en :

- Langage de programmation : Python 3.6.5
- Éditeur de code : Pycharm
- Bibliothèques importées :
  - BeautifulSoupparBS4
  - Un programme de prétraitement de Stemming : porter2
  - Counter par Collections
  - OS : pour les accès aux fichiers dans le système par python
  - Math: pour la réalisation des opérations arithmétiques
  - Operator
  - pickles
  - CSV : pour générer les résultats sous format exl.csv
- Un corpus appelé CACM contenant des documents et des requêtes.

## 3. Présentation du corpus

Le corpus choisi représente une collection de documents assemblés par une compagne appelée ACM (Association of Computing Machinery). Cette collection qui contient des articles publiés dans le Journal des communications de l'ACM entre 1958 et 1979 et des requêtes, a été utilisée dans de nombreux documents de recherche. L'ensemble des documents sont disponible sous format HTML. Les fichiers tar.gz sont archivés avec le programme GNU tar. Les utilisateurs Windows peuvent utiliser WinZip pour la décompression. Le corpus est disponible sous forme de trois catégories selon le volume du contenu:

- Wiki Small (6043 documents) [tar.gz, 26MB] [corpus, 36MB]
- Wiki Large (121,790 documents) [tar.gz, 524MB] [corpus, 715MB]
- CACM (3024 documents) [tar.gz, 1MB] [corpus, 1MB] [jugements de pertinence] [requêtes brutes] [requêtes traitées]

## 4. Modèle vectoriel choisi

Il existe plusieurs modèles vectoriels qui sont utilisés dans des systèmes de recherche d'information. Parmi ces modèles, on a choisi le VSM qui est un modèle très utilisé.

### 4.1. Présentation du modèle

Le modèle choisi c'est le modèle VSM (vector space model) qui se base sur la similarité cosinus et pour calculer cette dernière il utilise des tf*idf comme le poids de terme pour la requête et le document. (comme nous avons vu dans le premier chapitre ).

### 4.2. Validation du modèle sur le corpus

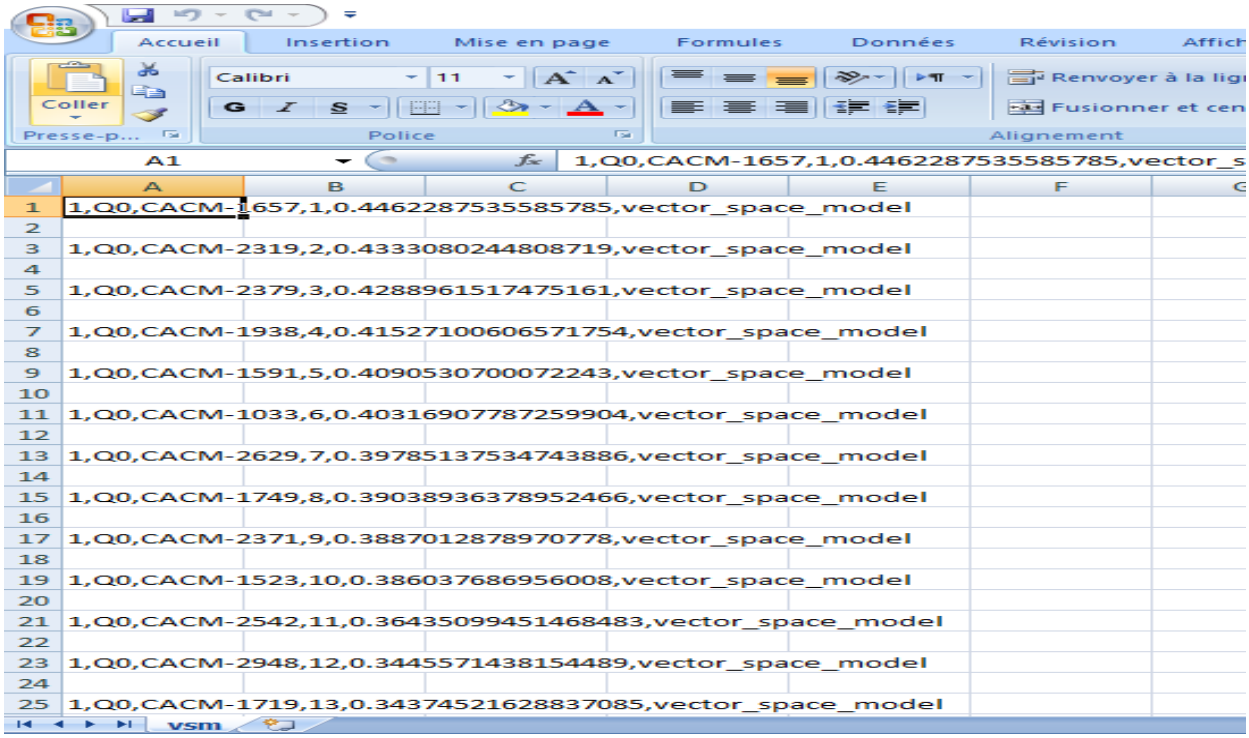


**Fig.13.les documents rapportés par le Modèle VSM**

Pour la validation, nous avons exécuté le premier modèle de recherche sur les documents du corpus choisi en considérant seulement trois requêtes choisies. Le moteur de recherche retourne un ensemble de documents pour chaque requête. En faisant référence aux documents pertinents pour chaque requête se trouvant dans le corpus, nous avons calculé le nombre de documents pertinents retournés pour chaque requête parmi les trois choisies. Les résultats de la validation sont indiqués dans le tableau suivant :

**Tableau.1.représentation des documents (retournés /pertinents) pour le modèle VSM**

| requête | Nb. Doc. Retournés | Nb. Doc. Pertinents (Corpus) | Nb. Doc. Pertinents (VSM) |
|---|---|---|---|
| Q1 | 99 | 5 | 2 |
| Q36 | 100 | 20 | 14 |
| Q58 | 99 | 30 | 7 |

### 4.3. Comparaison avec les résultats du corpus

Pour comparer les résultats du premier modèle de recherche avec les résultats du corpus, nous avons présenté dans le tableau suivant pour chaque requête, le nombre de documents pertinents se trouvant dans le corpus et le nombre de documents pertinents retournés par le système de recherche, puis nous avons calculé le rappel et la précision.

**Tableau.2. Calcul de mesure précision/rappel pour le modèle VSM**

| Requête | Nb. Documents pertinents | | Précision | Rappel |
|---|---|---|---|---|
| | Corpus | VSM | | |
| Q1 | 5 | 2 | 0,02 | 0,4 |
| Q36 | 20 | 14 | 0,14 | 0,7 |
| Q58 | 30 | 7 | 0,07 | 0,24 |

On remarque bien dans le graphe, que la précision pour les trois requêtes est très éloignée de la valeur 1, ce qui signifie que le taux des documents pertinents dans les documents rapportés est trop faible. On conclut que pour ces trois requêtes, le système VSM n'est pas précis.

D'un autre coté, On remarque dans le graphe, que le rappel pour la requête Q36 est proche de la valeur 1, ce qui signifie que le système VSM a renvoyé la plupart des documents pertinents du corpus pour cette requête, alors que pour les deux autres requêtes, Q1 et Q58, il n'a renvoyé que peut de documents pertinents.

## 5. Modèle probabiliste choisi

Il existe plusieurs modèles probabilistes qui sont utilisés dans des systèmes de recherche d'information. Parmi ces modèles, on a choisi le BM25 qui est un modèle très utilisé.

### 5.1. Présentation du modèle

Le modèle "Best match 25" (BM25) est un modèle issu du système de recherche de base d'Okapi dans les conférences de TREC.Ce modèle a été l'un des modèles de pondération de l'information les plus efficaces et largement utilisés au cours des trois dernières décennies. Contrairement à d'autres modèles probabilistes, le BM25 pourrait être calculé sans information pertinente.

### 5.2. Validation du modèle sur le corpus

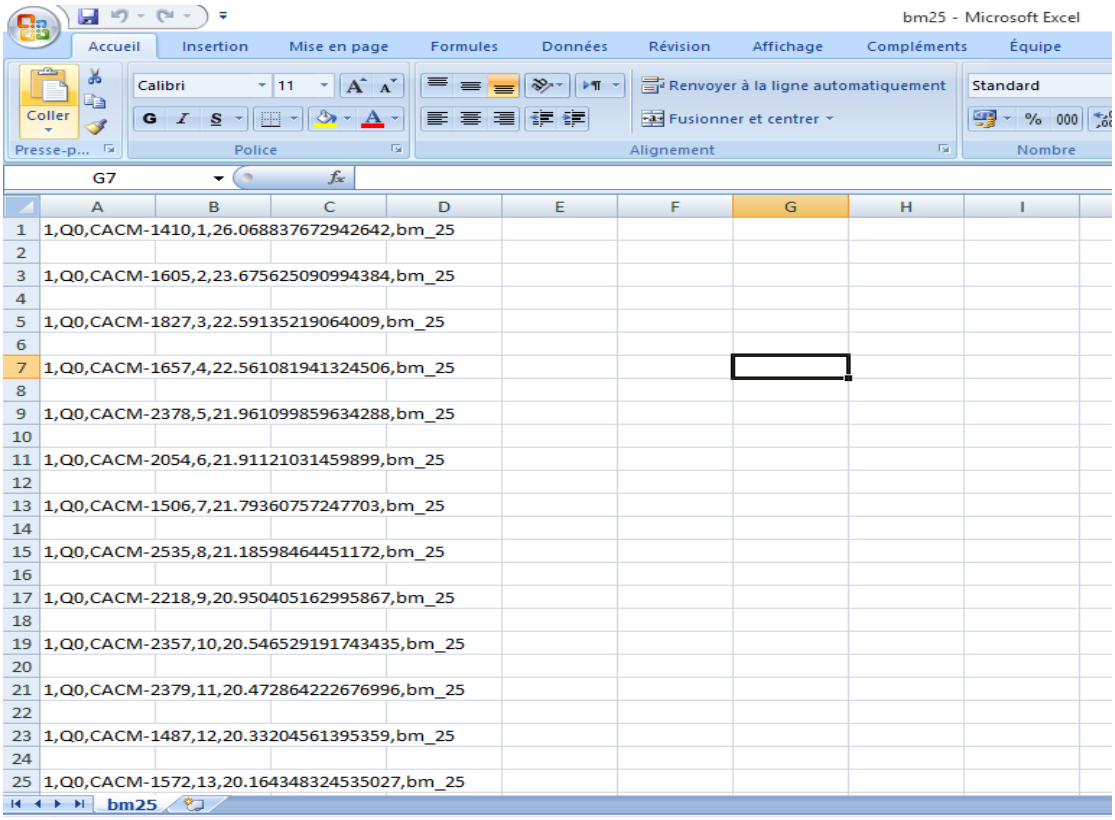


**Fig.14.les documents rapportés par le Modèle BM25**

Pour la validation, nous avons exécuté le deuxième moteur de recherche sur les documents du corpus choisi en considérant aussi les trois premières requêtes. Le moteur de recherche a retourné un ensemble de documents pour chaque requête. En faisant aussi référence aux documents pertinents pour chaque requête se trouvant dans le corpus, nous avons calculé le nombre de documents pertinents retournés pour chaque requête parmi les trois choisies. Les résultats de la validation sont indiqués dans le tableau suivant :

**Tableau.3. représentation des documents (retournés /pertinents) pour le modèle BM25**

| requête | Nb. Doc. Retournés | Nb. Doc. Pertinents (Corpus) | Nb. Doc. Pertinents (BM25) |
|---|---|---|---|
| Q1 | 100 | 5 | 4 |
| Q36 | 100 | 20 | 17 |
| Q58 | 100 | 30 | 14 |

### 5.3. Comparaison avec les résultats du corpus

Pour comparer les résultats du deuxième modèle de recherche avec les résultats du corpus, nous avons présenté dans le tableau suivant pour chaque requête, le nombre de documents pertinents se trouvant dans le corpus et le nombre de documents pertinents retournés par le moteur de recherche basé sur le modèle probabiliste, puis nous avons calculé le rappel et la précision.

**Tableau.4.Calcul de mesure précision/rappel pour le modèleBM25**

| Requête | Nb. Documents pertinents | | Précision | Rappel |
|---|---|---|---|---|
| | Corpus | BM25 | | |
| Q1 | 5 | 4 | 0,04 | 0,8 |
| Q36 | 20 | 17 | 0,17 | 0,85 |
| Q58 | 30 | 14 | 0,14 | 0,47 |

On remarque bien dans le graphe, que la précision pour les trois requêtes reste toujours éloignée de la valeur 1 tout comme dans le système VSM, ce qui signifie que le taux des documents pertinents dans les documents rapportés est trop faible pour le système BM25. On conclut que pour ces trois requêtes, le système BM25 n'est pas précis aussi.

D'un autre coté, On remarque dans le graphe, que le rappel pour les deux requêtes Q1 et Q36 est proche de la valeur 1, et pour la requête Q58 est proche de la valeur 0.5, ce qui signifie que le système BM25 a renvoyé la plupart des documents pertinents du corpus pour les deux premières requêtes et presque la moitié pour la dernière. On conclut que BM25 est meilleur que VSM de point de vue rappel.

## 6. Comparaison entre les deux modèles

En peut comparer entre les deux modèles en comparant entre le nombre de documents retournés, la précision el le rappel. Le tableau suivant fait l'objet.

**Tableau.5. Comparaison entre les deux modèles VSM/BM25**

| Système / Caractéristique | Nb Doc. Retournés | Précision | Rappel |
|---|---|---|---|
| VSM | 298 | 0,077 | 0,447 |
| BM25 | 300 | 0,117 | 0,707 |

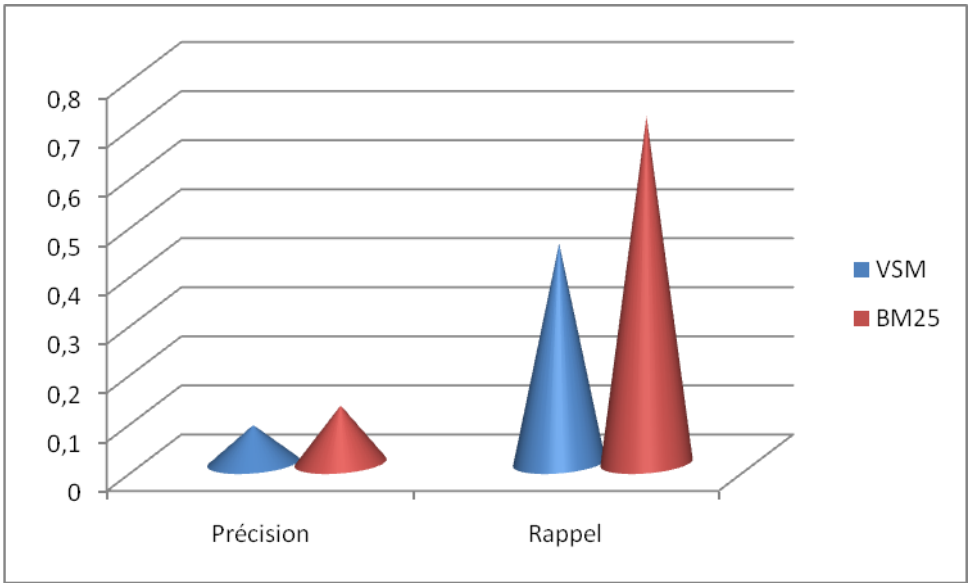


**Fig.15. Comparaison de mesure précision / rappel**

**entre les deux modèles**

D'après ce graphe, on remarque bien que le système BM25 et meilleur que le système VSM de point de vue précision et rappel, mais il reste toujours non précis mais avec un bon taux de rappel.

# 7. Expansion de requête

## 7.1. Outils d'expansion

La faible performance des systèmes de recherche d'information est dû généralement à l'incapacité des utilisateurs de formuler les requêtes adéquates. En effet, la requête initiale de l'utilisateur est souvent exprimée par une liste de termes souvent très réduite qui exprime mal les besoins en information de l'utilisateur. Pour remédier à ce problème, une solution consiste à étendre automatiquement la requête initiale afin d'améliorer la qualité des documents retrouvés.

La reformulation d'une requête consiste en l'ajout et/ou retrait de termes de la requête initiale. Il existe deux classes d'approches pour la reformulation de requête. Pour assurer ceci, on a besoin de préciser l'ensemble de requêtes à formuler, le dictionnaire des synonymes à utiliser pour remplacer les termes d'une requête, ce dictionnaire est soit à réaliser, soit à le trouver et l'utiliser et enfin le logiciel qui va extraire de la requête terme par terme et rechercher dans le dictionnaire les synonymes appropriés.

## 7.2. Processus d'expansion

Pour fixer le processus d'expansion, on doit préciser l'approche que notre méthode va suivre. L'approche à suivre est la réinjection de la pertinence utilisateur qui consiste à modifier la requête utilisateur à l'aide des documents jugés pertinents et/ou non pertinents par l'utilisateur.

Basé sur les jugements de pertinence ou non pertinence par l'utilisateur, cette reformulation peut se faire par une repondération des termes de la requête et/ou l'ajout (et/ou retrait) de termes contenus dans les documents pertinents (non pertinents). Le processus de cette tache peut être décrit comme suit :

1. l'utilisateur effectue une première requête de recherche,

2. le système retourne un ensemble de documents,

3. l'utilisateur indique parmi les documents retournés ceux qui sont pertinents et/ou non pertinents,

4. Le système modifie alors automatiquement la requête de départ en fonction des jugements de l'utilisateur.

Dans notre exponentiation, le déroulement du processus d'expansion sera le suivant:

- Utilisation ou réalisation d un dictionnaire des synonymes (langue anglaise) de telle sorte que chaque terme de la requête possède un synonyme ou plus.
- Label : Pour l'itération de recherche courante, le terme de la requête sera changé par le(s) terme(s) synonyme(s) extrait(s) du dictionnaire.
- Le système récupèreun nombre de documents pertinents pour la requête.
- Prendre le terme suivant de la requête et aller exécuter à partir de Label jusqu'abalayertous les termes de la requêteen fin on obtenuplus possible de documents.

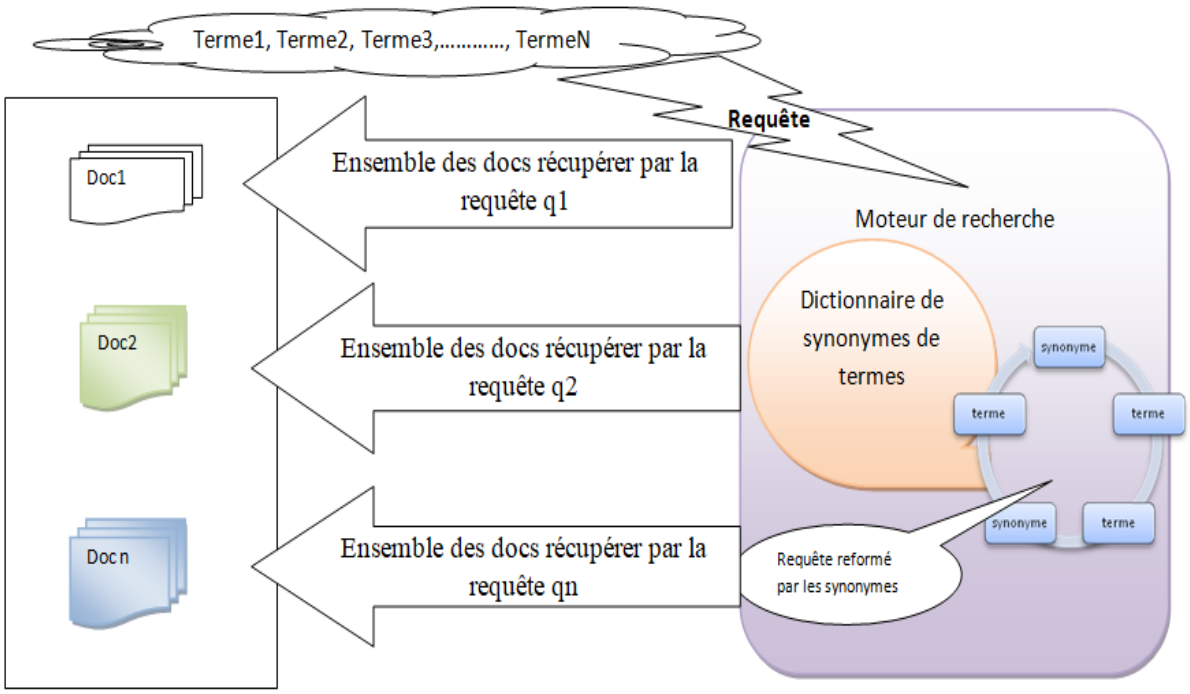


**Fig.16. Processus d'expansion**

### 7.3. Comparaison des résultats

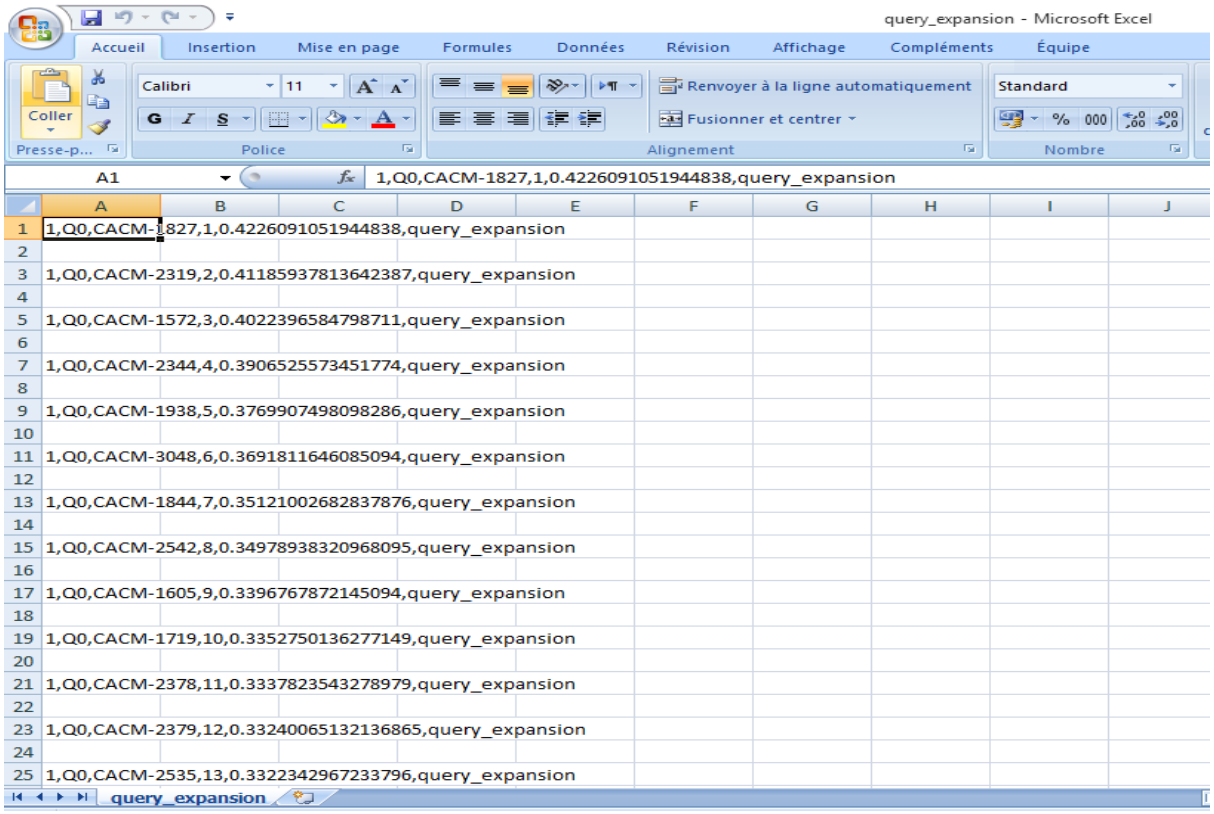


**Fig.17.les documents rapportés après l'exécution de l'expansion de requête**

On peut comparer entre les résultats obtenus avant l'application de l'expansion de la requête avec ceux obtenus après l'application de l'expansion. Pour catégorie de documents retournés, on peut calculer la précision el le rappel. Le tableau suivant fait l'objet.

- Avant application de l'expansion

**Tableau.6.les résultats obtenus avant l'expansion de requête**

| | | VSM | | BM25 | |
|---|---|---|---|---|---|
| Requête | Nb. Doc. Retournés | Précision | Rappel | Précision | Rappel |
| Q1 | 100 | 0,02 | 0,4 | 0,04 | 0,8 |
| Q36 | 100 | 0,14 | 0,7 | 0,17 | 0,85 |
| Q58 | 100 | 0,07 | 0,24 | 0,14 | 0,47 |

- Après application de l'expansion

**Tableau.7.représentation des documents pertinents et mesure précision /Rappel pour expansion de requête**

| Requête | Nb. Doc. Retournés | Précision | Rappel |
|---|---|---|---|
| Q1 | 100 | 0,03 | 0,6 |
| Q36 | 100 | 0,1 | 0,5 |
| Q58 | 100 | 0,08 | 0,27 |

D'après ces statistiques, on peut dessiner le graphe suivant de la précision indiquant la précision des trois requêtes pour le système VSM, le système BM25 et l'expansion de requête.

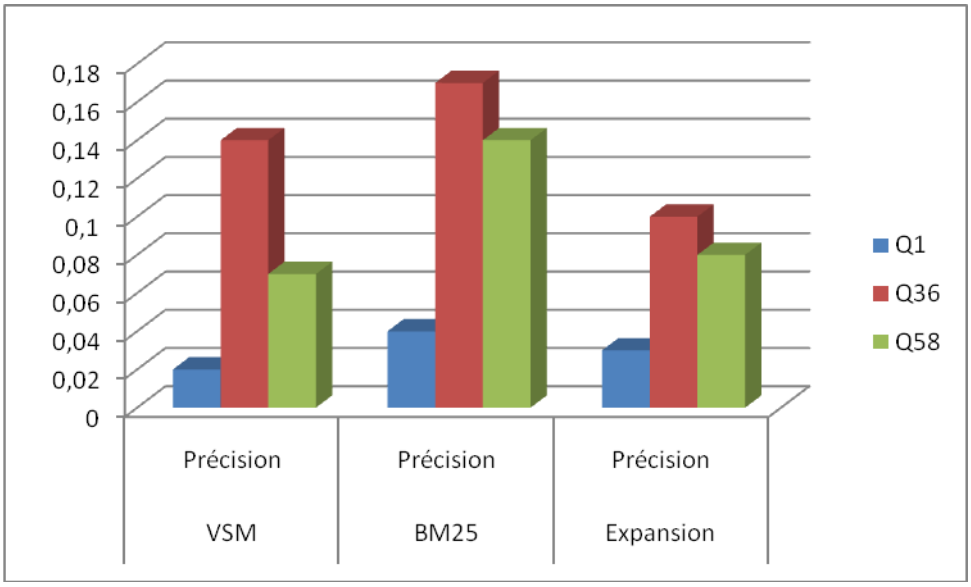


**Fig.18.mesure de précision pour les trois modèles**

On remarque que l'expansion a amélioré la précision pour les deux requêtes Q1 et Q58 concernant le système VSM alors qu'elle n'a pas amélioré la précision pour les trois requêtes concernant le système BM25

D'après les statistiques précédentes aussi, on peut dessiner le graphe suivant du rappel indiquant le rappel des trois requêtes pour le système VSM, le système BM25 et l'expansion de requête.

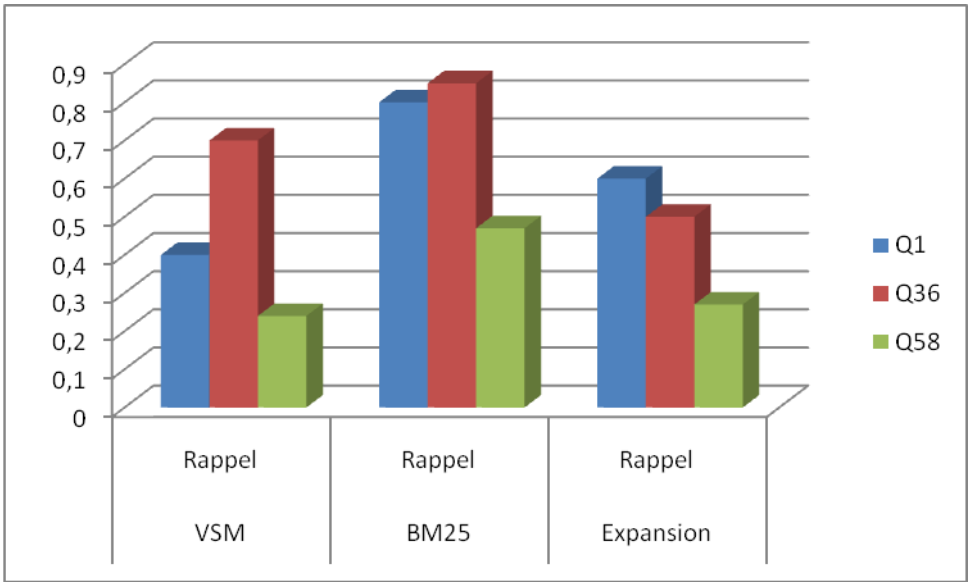


**Fig.18.mesure de rappel pour les trois modèles**

On remarque que l'expansion a amélioré le rappel pour les deux requêtes Q1 et peu pour Q58 concernant le système VSM alors qu'elle n'a pas amélioré le rappel pour les trois requêtes concernant le système BM25.

- Après la lecture et l'analyse des résultats obtenues des expériences effectuées sur le moteur de recherche et le corpus on remarqué que :

-le premier modèle VSM (vetor space model) est capable de récupérer un nombres important des documents et ce qui assure la récupération des documents pertinents et acceptable, dès qu'on considère que les documents pertinents et la fréquence de l'apparence des termes dans ces documents sont validés par le copus de test , on peut calculer la précision et le rappel par les équation correspondantes pour ce but .

-Dans le deuxième modèle BM25 (modèle probabiliste étendu 25) , on doit prendre en compte le nombre total des documents pertinents lors du calcul et cela pour assurer plus d'efficacité de ce modèle , la récupération des documents sera importante et assurée ,et ainsi l'apparition les documents pertinents qui sera très élevée , le calcule de précision et le rappel sont liés par le nombre des documents pertinents rapportés et le nombres des documents pertinents donnés .

- Après l'application de l'expansion de la requête qui se base sur le remplacement des termes de la cette dernière par les synonymes qui existent dans le dictionnaire des synonymes, On a observé que les résultats ne sont pas stables, ce problème traduit le voulu de l'utilisateur qui lui-même juge la pertinence des documents.

## 8. Conclusion

On s'est basé dans tous nos résultats sur les données des fichiers du CACM. À partir de l'analyse ci-dessus, On a constaté que le BM25 est le meilleure modèle dans la recherche d'information par rapport au VSM, et que l'expansion de requête dans ce cas là est moins efficace par rapport au BM25 et par contre plus efficace qu'au VSM.

# Conclusion général

# Conclusion général

Les travaux présents dans ce mémoire s'inscrivent dans le cadre de la Recherche d'Information (RI). Nous nous sommes principalement intéressés à la comparaison de la pertinence des documents dans plusieurs modèles de recherche d'information, pour déterminer le meilleur modèle.

La comparaison réalisée dans notre projet, a consistée sur le choix des deux modèles de recherche d'information classiques, le modèle probabiliste et le modèle vectoriel. Après les validations de chaque modèle sur le corpus, nous avons obtenu le résultat qui affirme que le modèle probabiliste est le meilleur dans la récupération des documents pertinents. Par suite nous avons développé un modèle de recherche qui se base sur l'expansion de requête, ce dernier consiste à passer d'une requête initiale à une autre requête autrement dit une requête reformulée par le fonctionnement d'un dictionnaire de termes synonymes qui sont utilisés pour remplacer les termes initiaux, après nous avons validé ce modèle sur le corpus.

Après cette dernière validation, le résultat obtenu qui affirme que le modèle probabiliste (BM25) est supérieur dans la pertinence des documents par rapport au modèle d'expansion de requête et le modèle vectoriel qui a montré une infériorité dans la pertinence.

# Références

# *Références*


**[1]** *MR. ABDELKRIM BOURAMOUL«RECHERCHE D'INFORMATIONCONTEXTUELLEET SEMANTIQUE SUR LE WEB » thèse de doctorat en informatique (2011)*

**[2]** *Hernandez N. « Ontologie de domaine pour la modélisation du contexte en recherche d'information », thèse de doctorat en informatique. (2006)*

**[3]** *ASMA HEDIA BRINI«Un Modèle de Recherche d'Information basé sur les Réseaux Possibilistes » thèsede doctorat en informatique (2005)*

**[4]** *V Singh ,B Saini«Probabilistic Ranking of Documents Using Vectors in Information Retrieval»*

**[5]** *G. KowalskiM. Maybury « INFORMATION STORAGE AND RETRIEVAL SYSTEMS » Theory and Implementation Second Edition*

**[6]** *M. Sanderson ,W. B Croft« The History of Information Retrieval Research »*

**[7]** *S. Boucham« Une approche basée ontologie pour l'indexation automatique et la recherche d'information multilingue » Mémoire de magister(2008)*

**[8]***G. LE TARGAT« Langages classificatoires et recherche d'information sur les portails d'entreprise : quels apports pour les utilisateurs ? » (2005)*

**[9]** *MANIEZ, Jacques. «Actualités des langages documentaires ». Paris ( 2002)*

**[10]***L.Bouabdallah « Expansion-de-requete-pour-un-systeme-de-recherche-dinformation-par-croisement-de-langues »(2012)*

**[11]***P .Bruno, D.Denis, B . Pierre « Indexation de textes médicaux par extraction de concepts, et ses utilisations » (2002) .*

**[12]***k. MECHACH « Etude de l'impact des méthodes de localisation dans les systèmes de recherche distribuées » thèse de doctorat (2015)*

**[13]** *Jitendra Nath Singh Sanjay Kumar ,Dwivedi Lucknow, India. Babasaheb Bhimrao Ambedkar University Department of Computer Science''A Comparative Study on Approaches of Vector SpaceModel in Information Retrieval '' International Journal of Computer Applications (0975 – 8887) International Conference of Reliability, Infocom Technologies and Optimization, (2013)*

**[14]***M .Pannu«A Comparison of Information Retrieval Models»(2014)*

**[15]** *A.Sheth« History of Machine Learning » site web (2017)*

**[16]** *T.O. Ayodele «Types of Machine Learning Algorithms» site web*

**[17]** *C. Lee , Y. Lin , R. Chen«Query Formulation by Selecting Good Terms » (2010)*

**[18]** *http://www.riemysore.ac.in/ict/unit__7__elearning.html#7.3.1*

# *Références*

**[19]** *N. K. Shah«E-Learning and Semantic Web»(2012)*

**[20]** *https://blogs.msdn.microsoft.com/continuous_learning/2014/11/15/end-to-end-predictivemodel-in-azureml-using-linear-regression/*

**[21]***A. Yengui «Système de recherche d'information sémantique pour les bases de visioconférences médicales à travers les graphes conceptuels» thèse de doctorat en informatique( 2016)*

**[22]** *O. Asfari «Personalized Access to Contextual Information by using an Assistant for Query Reformulation » thèse de doctorat en informatique (2011)*

**[23]** *B. Audeh«Reformulation sémantique des requêtes pour la recherche d'information ad hoc sur le Web » thèse de doctorat en informatique(2014)*

**[24]***A .Bouziri, C. Latiri ,E. Gaussier « Expansion de requêtes par apprentissage »(2014)*

**[25]** *Maron and Kuhns, J. «On relevance, probabilistic indexing and information retrieval» Journal of the Association for Computing Machi- nery 7 (1960)*

**[26]***R. Korfhage (1997). «Information storage and retrieval». Wiley Computer Publishing*

**[27]** A.Nahar *«cours apprentissage automatique » (2017)*

**[28]** *Rosenberg, M.J. «E-Learning: Strategies for Delivering Knowledge in the Digital Age. Vol. 9, McGraw-Hill, New York » (2001)*

**[29]***H.Brandon « E-learning, A research note by Namah»,( 2011)*

**[30]** V.C.Storey, V.Sugumaran, , A.Burton-Jones, «The Role of User Profiles in Context-Aware Query Processing for the Semantic Web» Conference on Applications of Natural Language to Information Systems (2004)