



جامعة أبو بكر بلقايد - تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie

Département de Génie Biomédical

Laboratoire de Recherche de Génie Biomédical

MEMOIRE DE PROJET DE FIN D'ETUDES

pour obtenir le Diplôme de

MASTER en Génie Biomédical

Spécialité : Informatique Biomédicale

Présenté par : Larab Farid et Keddour Faiza Nesrine

Recherche Intelligente d'Information Médicale

Soutenu le 28 mai 2015 devant le Jury

| | | |
|-------------------------------------|-----------------------|------------|
| Mme. Dali youcef Lamia | Université de Tlemcen | Présidente |
| Mr. Boudefla Amine | Université de Tlemcen | Examineur |
| Mr. Abderahim Mohammed Alaeddine | Université de Tlemcen | Encadreur |

Année universitaire 2014-2015

Remerciement

Avant tout, nous remercions Dieu, le miséricordieux, le Tout Puissant pour nous avoir donné la force et le courage d'accomplir ce travail avec beaucoup de courage et d'abnégation.

Nous adressons nos vifs remerciements

à nos encadrateurs Monsieur Abdérahim Mohammed alla edine et Abdérahim Mohammed El Amin de nous avoir proposé ce sujet et de nous avoir encadré tout au long de ce travail.

Nous rendons aussi un vibrant hommage

À tous nos Professeurs, Nous saisissons cette occasion pour vous exprimer notre profonde reconnaissance tout en vous témoignant notre respect et notre humble dévouement.

Nous remercions également,

Les membres du jury de nous avoir fait l'honneur d'expertiser ce mémoire Veuillez accepter l'expression de notre vive gratitude et notre parfaite déférence. Nous tenons à remercier également, tous ceux qui ont contribué par leurs conseils ou leurs encouragements à l'aboutissement de ce modeste travail.

Merci à tous et à toutes.

Dédicaces

A ma mère,

A la plus merveilleuse des mères. J'espère réaliser, en ce jour, l'un de tes rêves. Aucun mot ne saurait exprimer mon respect, ma considération et l'amour que je te porte. Ta présence constante à mes côtés, tes encouragements et tes prières m'ont été d'une aide précieuse et m'ont permis d'atteindre le but désiré. Puisse Dieu le tout puissant te donner santé et longue vie afin que je puisse te combler à mon tour...

A mon père,

Aucune dédicace ne saurait exprimer à sa juste valeur tout l'amour, le respect, l'attachement et la reconnaissance que je te porte. Tu m'as enseigné la droiture, le respect et la conscience du devoir. Ce travail est le fruit de tous tes sacrifices, tes encouragements, ton désir de me voir arriver et ton soutien permanent durant ce long parcours. Puisse Dieu, le tout puissant, te procurer santé, bonheur et longue vie...

J'espère

Avoir répondu aux espoirs que vous avez fondés en moi. Je vous rends hommage ici par ce modeste travail en guise de ma reconnaissance éternelle à jamais... Que Dieu, le Tout Puissant vous garde et vous procure santé, bonheur et longue vie.

Toute ma famille,

Grands et petits qui attendent ce travail avec impatience et surtout mes frères Takfarinas et Dihia Trouvez ici l'expression de ma profonde affection et mon respect.

Mes professeurs,

Qui m'ont enseigné depuis ma première scolarité. Trouvez ici l'expression de mes respects et mon éternelle reconnaissance. A tous mes amis, collègues, mon binôme keddour FAIZA NESRINE et toute la promotion master en génie biomédicale Que tous vos rêves soient exaucés et la réussite comble votre vie.

M .Larab Farid

Dédicaces

A ma mère,

L'école de la vie qui m'a enseigné mes premiers pas. Maman, je ne connais pas une personne aussi adorable, tendre que toi, tu as été toujours à mes côtés comme un ange guidant mes pas à chaque moment de ma vie, me couvrant de ta tendresse et de ton amour éternel.

A mon père,

Tous les mots du monde ne sauraient exprimer l'immense amour que je te porte, ni la profonde gratitude que je te témoigne pour tous les efforts et les sacrifices que tu n'as cessé de consentir pour mon instruction et mon bien-être.

Mon cher fiancé

Mohcine pour sa compréhension, ses encouragements, et sa confiance.

A mes sœurs

Meriem, Chaimaa et mon frère Zkaria

J'espère Avoir répondu aux espoirs que vous avez fondés en moi. Je vous rends hommage ici par ce modeste travail en guise de ma reconnaissance éternelle à jamais... Que Dieu, le Tout Puissant vous garde et vous procure santé, bonheur et longue vie. Toute ma famille, Grands et petits qui attendent ce travail avec impatience, Trouvez ici l'expression de ma profonde affection et mon respect.

Mes professeurs,

Qui m'ont enseigné depuis ma première scolarité. Trouvez ici l'expression de mes respects et mon éternelle reconnaissance. A toutes mes amies d'études, mes copines, à Nasreddine, Djamel, et Amine à mon binôme Farid larab et toute la promotion master génie biomédicale Que tous vos rêves soient exaucés et la réussite comble votre vie.

Mlle .kēddour faiza nesrine

TABLE DES MATIERES

| | |
|----------------------------|------|
| Remerciements..... | i |
| Dédicace..... | ii |
| Table des matières..... | iv |
| Table des figures..... | vii |
| Glossaire..... | viii |
| Introduction générale..... | 10 |

CHAPITRE I. les Système de recherche d'information

| | |
|--|----|
| 1. Introduction :..... | 14 |
| 2. Définition | 14 |
| 3. Notions de base dans un SRI :..... | 14 |
| 4. Architecture générale des systèmes de recherche d'information | 15 |
| 4.1. Processus de représentation (indexation) | 16 |
| 4.1.1. Indexation manuel :..... | 16 |
| 4.1.2. Indexation semi-automatique :..... | 16 |
| 4.1.3. Indexation automatique..... | 17 |
| 4.2. Le résultat de l'indexation : L'index | 20 |
| 4.3. La reformulation de la requête | 21 |
| 4.3.1. La reformulation manuelle..... | 21 |
| 4.3.2. La reformulation automatique | 21 |
| 4.3.3. La reformulation interactive..... | 21 |
| 5. La notion de pertinence..... | 22 |
| 6. Les modèles de recherche d'information..... | 22 |

| | | |
|------|-------------------------------------|----|
| 7. | Evaluation d'un système..... | 23 |
| 7.1. | Précision | 23 |
| 7.2. | Rappel..... | 23 |
| 7.3. | La courbe de Précision-Rappel | 24 |
| 8. | Conclusion : | 25 |

CHAPITRE II. Les ontologies

| | | |
|--------|---|----|
| 1. | Introduction..... | 27 |
| 2. | Les ontologies | 27 |
| 2.1. | Définition des ontologies | 27 |
| 2.2. | Rôles des ontologies..... | 28 |
| 2.3. | Les principaux types d'ontologies | 29 |
| 2.4. | Les ontologies les plus connu du domaine médical | 29 |
| 2.4.1. | UMLS | 30 |
| 2.4.2. | SNOMED CT | 30 |
| 2.4.3. | MeSH | 30 |
| 2.4.4. | Galen | 31 |
| 3. | Ontologies et recherche d'information | 31 |
| 3.1. | Quelles ontologies choisir | 31 |
| 3.2. | Principe d'utilisation des ontologies par un système de recherche d'information | 32 |
| 3.3. | La reformulation de la requête par utilisation des ontologies..... | 32 |
| 3.3.1. | Appariement à partir d'ontologies | 33 |
| 3.3.2. | Indexation des documents par les ontologies..... | 33 |
| 4. | Conclusion | 33 |

CHAPITRE III. Conception et mise en œuvre

| | | |
|--------|-------------------------------------|----|
| 1. | Introduction..... | 35 |
| 2. | Problématique | 35 |
| 3. | Environnement de développement..... | 36 |
| 3.1. | Langages utilisés..... | 36 |
| 3.1.1. | Java | 36 |

| | | |
|--------|--|----|
| 3.1.2. | MySQL..... | 36 |
| 3.2. | OUTILS | 37 |
| 3.2.1. | Eclipse..... | 37 |
| 3.2.2. | L'API JENA..... | 37 |
| 3.2.3. | Lucene | 38 |
| 4. | Solution proposé..... | 38 |
| 4.1. | Ontologie MeSH | 39 |
| 4.2. | Système proposé de recherche..... | 40 |
| 4.3. | Réalisation pratique | 40 |
| 4.3.1. | Application1 : Recherche simple | 40 |
| 4.3.2. | Application2 : Recherche Intelligente d'Information Médicale | 44 |
| 4.4. | Discussion | 47 |
| 5. | Conclusion | 49 |
| | Conclusion générale..... | 50 |
| | Référence..... | 52 |

Table des figures

| | |
|---|----|
| Figure I.1: Le Processus en U de la Recherche d'Information [MEHIDI et RABAH, 14]..... | 15 |
| Figure I.2: Les étapes de l'indexation..... | 17 |
| Figure I.3: Le rapport entre le rang × fréquence et l'importance du terme) [Yaël, 09]..... | 19 |
| Figure I.4: Allure d'une courbe de rappel-précision. | 24 |
| | |
| Figure II.1: Une partie de l'ontologie des formes géométriques [Bendaoud, 09]..... | 28 |
| | |
| Figure III.1: Page d'accueil jena (http://jena.sourceforge.net/)..... | 37 |
| Figure III.2: Architecture de notre système de recherche. | 38 |
| Figure III.3: Fenêtre principale de la recherche. | 41 |
| Figure III.4: Ajouter un document..... | 42 |
| Figure III.5: Affichage des résultats..... | 43 |
| Figure III.6: Affichage du contenu de l'index. | 44 |
| Figure III.7: Fenêtre principale de la recherche. | 45 |
| Figure III.8: Ajouter un document..... | 45 |
| Figure III.9: Ajouter une ontologie..... | 46 |
| Figure III.10: Affichage des résultats selon l'ordre décroissant. | 47 |
| Figure III.11: Résultat de la recherche simple..... | 48 |
| Figure III.12: résultat de la recherche intelligente..... | 49 |

Glossaire

SRI: système de recherche d'information

Tf: Terme Frequency

Idf: Inverted Document Frequency

IA: Intelligence Artificiel

SNOMED-CT: Systematized Nomenclature of Medicine, Clinical Terms

MeSH: Medical Subject Headings

SQL: Structured Query Language

INTRODUCTION GÉNÉRALE

Introduction Générale

Avec la croissance des données, une recherche d'information repose sur l'utilisation d'une liste des mots clés deviens insuffisante pour avoir des résultats pertinents. En effet, les systèmes de recherches ne disposent pas d'information automatique de nature sémantique à propos du contenu. L'utilisation des termes ce diffères de façon morphologique, syntaxique...etc. influx sur la qualité des résultats retournés par les SRIs

En effet, les ontologies offrent des ressources sémantiques, qui peuvent constituer une solution (parmi d'autres) pour résoudre le problème des variations sémantiques. L'utilisation des ontologies dans un système de recherche d'information peut être envisagée à plusieurs niveaux. Avant d'être envoyée la requête de l'utilisateur pourra être enrichie par les concepts jugés proches dans l'ontologie et ceci, par le biais de l'utilisation des relations appropriées. Une telle recherche d'information a déjà été étudiée dans de nombreux projets pour différents langues comme l'anglais, Ontobroker ou WebKB (graphes conceptuels). Ces projets ont démontré l'utilité et l'avantage de la recherche d'information guidée par les ontologies. Notre projet consiste à implémenter un système de recherche d'information à base d'ontologie en utilisant le modèle vectoriel pour le domaine médicale. Il permettra de dépasser les limites d'une recherche classique par mots clés.

Ce système combinera le modèle de recherche d'information classique (modèle vectoriel avec une ontologie médicale MESH).

Objectifs :

Dans le cadre de ce mémoire, on se propose d'améliorer les résultats de la recherche d'information dans un Système de Recherche d'Information (SRI) en intégrant des ressources sémantiques comme les ontologies du domaine médical. Dans le présent document nous allons discuter les éléments suivants :

- ❖ Premièrement, il s'agit de dresser un état de l'art sur les modèles de recherche d'informations ainsi que leur mode de fonctionnement. Puis, étudier les effets des ontologies aux différents niveaux des SRIs.

- ❖ Deuxièmement, nous proposons un modèle de recherche en intégrant l'ontologie médicale au niveau de la requête et de l'indexation des documents afin d'améliorer les résultats d'un SRI.

Organisation du projet

Le mémoire est organisé selon trois chapitres :

Chapitre 1 : les systèmes de recherche d'information

Dans ce chapitre nous allons aborder les systèmes de recherche d'informations en général, on a donné une définition de la recherche d'information, ses notions de bases ainsi que les différentes approches adoptées, on a parlé aussi de l'indexation, évaluation d'un SRI et une bref description sur les modèles de recherche d'information (modèle booléen, probabiliste, le model logique et le model vectoriel).

Chapitre2 : les ontologies

Dans ce chapitre nous avons décrit les ontologies, leur rôle, leur type et leur principe d'utilisation dans un système de recherche d'information.

Chapitre3 : Conception et mise en œuvre

Ce chapitre présente la conception et réalisation complète de notre système de recherche. L'explication des différentes étapes d'implémentation ainsi que les résultats obtenus.

CHAPITRE I

Les systèmes de recherche d'Information

1. Introduction :

L'objectif principal d'un SRI est de mettre en œuvre un processus de comparaison entre le besoin utilisateur et les documents d'une collection dans le but de retrouver ceux qui sont pertinents. L'élaboration d'un mécanisme de recherche d'information pose alors des problèmes liés tant à la représentation qu'à la localisation de l'information pertinente. Tout au long de ce chapitre, notre intérêt se porte sur les principes de la recherche d'information, ses notions de bases ainsi que les différentes approches adoptées.

2. Définition

Un SRI est un système qui gère une collection d'informations organisée sous forme d'une représentation intermédiaire qui reflète aussi fidèlement que possible le contenu des documents. Ceci est réalisé grâce à un processus préalable d'indexation manuelle ou automatique. La recherche d'information désigne alors le processus qui permet, à partir d'une expression des besoins d'information d'un utilisateur, de retrouver l'ensemble des documents contenant l'information recherchée. Le résultat est donné par la mise en œuvre d'un mécanisme d'appariement entre la requête de l'utilisateur et les documents ou plus exactement entre la représentation de la requête et la représentation des documents [Ouaneche, 13].

3. Notions de base dans un SRI :

On distingue quatre notions de base dans un SRI:

- **La notion de document** : L'ensemble des documents sur lesquels portera la recherche est stocké dans une banque de données (sur le Web). Un document est le type d'objet de base géré par le système.
- **La notion de besoin d'information d'un utilisateur (requête):**

Ce besoin est exprimé par une requête spécifiée dans un formalisme propre au système. Le formalisme de spécification de la requête peut être en langage naturel.

- **Notion de correspondance entre la requête et les documents:**

Une fois la requête spécifiée, le système tente de retrouver les documents qui correspondent à la requête en se basant sur une mesure de similarité

- **La notion de contexte de l'application** : Le contexte de l'application représente l'univers dans lequel le système fonctionne. L'univers est nécessaire aux SRI pour une bonne compréhension des besoins des utilisateurs. Un SRI doit être capable de retrouver les documents pertinents à partir d'une banque de données, satisfaisant la requête posée par un utilisateur et traduisant un besoin d'information donné.

4. Architecture générale des systèmes de recherche d'information

Un système de recherche d'information possède trois fonctions principales présentées schématiquement dans la figure 1.1 par un processus en U :

La partie gauche désigne l'information accessible par le système tel que les collections des documents qui regroupe les documents de même domaine ou de domaine proche.

Dans la partie à droite nous trouvons le besoin en information de l'utilisateur : exprimé par une requête. Enfin, pour combiné les deux cotés ; le système propose un certain nombre de traitement pour trouver les informations pertinentes.

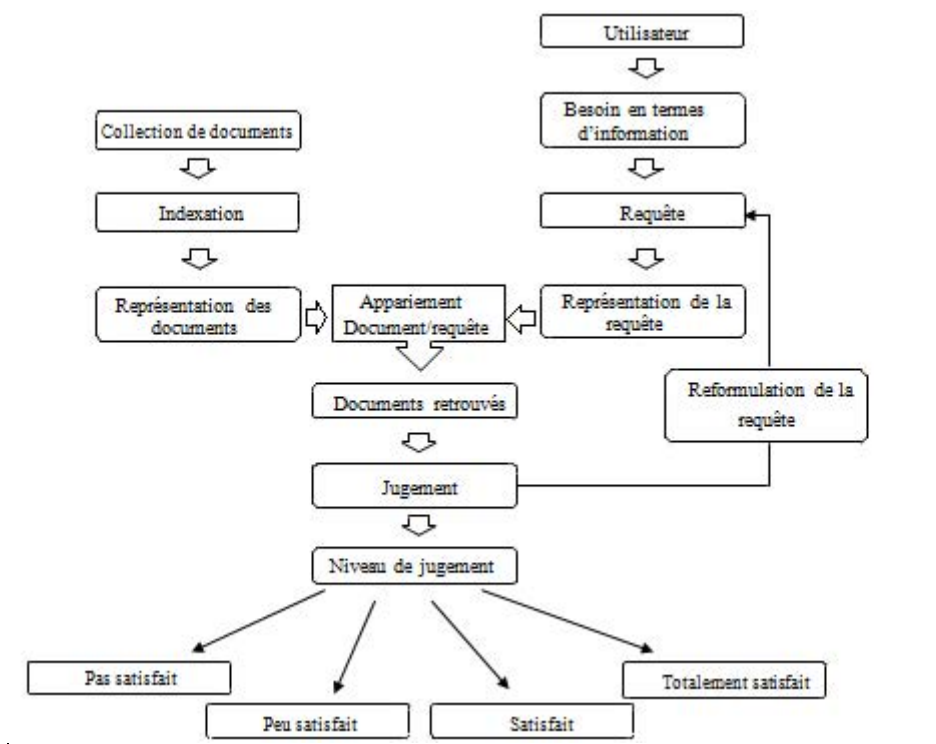


Figure I.1: Le Processus en U de la Recherche d'Information [MEHIDI et RABAH, 14].

Ces traitements sont de deux catégories : la représentation (indexation) et la recherche

4.1. Processus de représentation (indexation)

La recherche d'information(RI) par un parcours complet de tous les textes de la collection de documents n'est pas pratique [Bouidghaen, 11]. Donc on utilise une opération appelée l'indexation qui consiste à créer un ensemble des mots clés (termes) à partir de l'analyse d'un document qui se trouve dans une collection de documents pour que l'exploitation de ces mots clés ou descripteur par le système soit facile. Ces mots clés peuvent être regroupés dans un thésaurus (« en pratique, un thésaurus regroupe plusieurs relations de types linguistique (équivalence, association, hiérarchie) et statistique (pondération)»)[Baziz,05].

L'objectif de cette opération est de garder les termes significatifs de ce document.

On distingue deux types d'indexation :

- **l'indexation libre** : les termes sont extraits du texte à indexer
- **l'indexation contrôlée** qui s'appuie sur un ensemble prédéfini de termes: le thésaurus. Elle consiste à sélectionner les termes de ce thésaurus qui indexent ce texte.

L'opération d'indexation peut se dérouler en trois modes différents [Kompaoré, 08] [Nassr, 02] :

4.1.1. Indexation manuel :

Chaque document est analysé par un documentaliste ou par un spécialiste du domaine, qui extrait les mots basant sur un vocabulaire contrôlé (liste hiérarchique, thésaurus, lexique,...) [Bouidghaen, 11]. L'avantage de l'indexation manuelle est d'assurer un meilleur rapport entre les documents et les termes choisis par les spécialistes [Baziz, 05], mais l'inconvénient, elle est coûteuse en termes de temps (nécessite un temps important) et plus d'effort intellectuel (nombre de personnes).

4.1.2. Indexation semi-automatique :

Appelée aussi indexation supervisée. Ce type d'indexation fait une combinaison des deux modes précédant [Daoud, 09], les termes du document sont extraits en un premier temps par un processus automatique. Mais le choix final des mots clés est fait

par l'indexeur ou le spécialiste du domaine, généralement les indexeurs utilisent un vocabulaire contrôlé sous forme de thesaurus ou de base terminologique [Baziz, 05]

4.1.3. Indexation automatique

L'indexation dans ce cas-là est faite par un SRI basé sur des algorithmes et des méthodes, l'expert du domaine n'intervient pas [Daoud, 09]. Elle détecte d'une façon automatique les concepts significatifs d'un document en analysant le document mot par mot aussi l'élimination des mots vides, la lemmatisation, la pondération des termes, à la fin la création de l'index. Ce type d'indexation est souvent le plus utilisé. [Nassr, 02].

De manière générale, l'indexation automatique est réalisée selon les étapes suivantes [Hlaoua,07] :

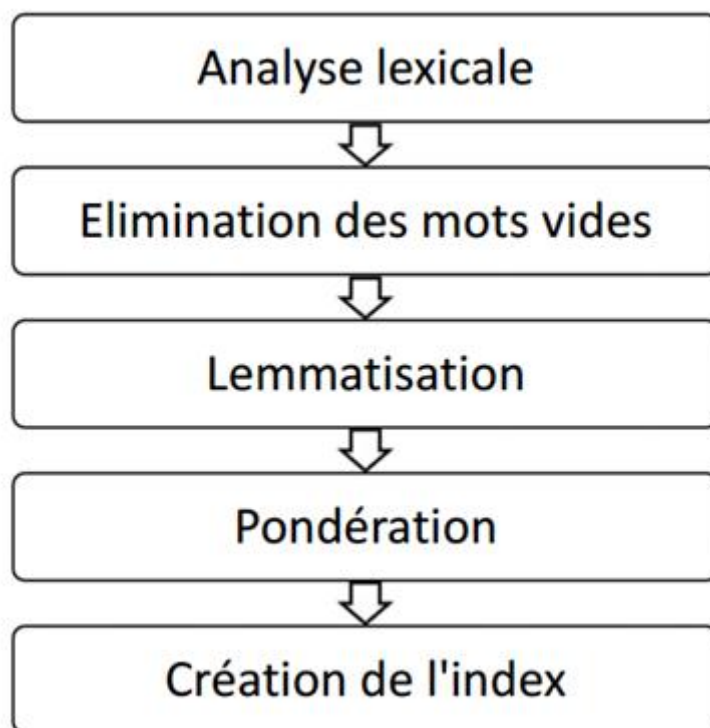


Figure I.2: Les étapes de l'indexation.

4.1.3.1. L'analyse lexicale :

C'est l'opération de transformer un document textuel en un ensemble de termes ou unité lexicale en reconnaissant les espaces de séparation des mots, des caractères spéciaux, des chiffres, les ponctuations ou une liste de séparateurs [Kompaoré, 08]. Il y a des cas

particuliers où les mots sont composés ou contiennent des séparateurs Ex : aujourd'hui, pomme de terre, Le Mans, 127.0.0.1, M. Durand, 14/07/1789.

4.1.3.2. Elimination des mots vides

C'est la suppression d'une liste des mots préétablie appelée des mots vides (tels que les pronoms personnels, les articles, les mots de liaison, ou les prépositions), pour ne garder que les termes importants. Plusieurs techniques peuvent être mises en œuvre parmi celles-ci, l'utilisation des stops liste ou des anti-dictionnaires (anti-lexiques) et l'utilisation des mesures statistiques. Le traitement lié à un anti-dictionnaire est très simple si un mot est apparait dans l'anti-dictionnaire et dans le texte a indexé, il n'est pas considéré comme un index, il peut cependant induire des effets de silence (par exemple, en éliminant le mot **a** de **vitamine a** [Kompaoré, 08]).

4.1.3.3. Lemmatisation (radicalisation)

C'est une étape qui a pour but de regrouper les différentes variantes d'un mot à sa forme canonique ou lemme dans l'exemple de [Yaël, 09] « il peut être utile de retrouver des documents contenant les mots « transmission », « transmis », « transmet », « transmettra », « transmetteur » à partir d'une requête comportant le mot « transmettre ». Pour cela il est possible d'éliminer les différences non significatives et de garder la partie commune. Sur l'exemple, les mots ont la même racine (le lemme) et une terminaison différente». Pour les verbes conjugués il suffit de rendre le verbe à l'infinitif pour les conjugués et le singulier pour les noms. Des fois le passage à la forme canonique supprime le sens du mot comme le verbe portera et le nom porte sera indexé de la même façon.

4.1.3.4. Pondération des termes

Le poids d'un terme dans un document traduit l'importance de ce terme dans le document. La pondération est une fonction fondamentale, elle est généralement basée sur l'association des valeurs numériques aux termes de manière à représenter le pouvoir de discrimination (degré d'informativité) de ces termes pour chaque document de la collection, l'objectif est de trouver les termes qui représentent le mieux le contenu d'un document.

La pondération d'un terme s'exprime en fonction de deux pondérations :

- Une pondération locale reflète l'importance locale du terme dans le document (mesures statistiques locales).

- La pondération globale exprime l'importance globale du terme dans la collection (mesures statistiques globale). La majorité de ces mesures tirent leur origine de la loi de Zipf et de la conjecture de Kuhn [Harrathi, 10].

4.1.3.4.1. Quelques méthodes de pondération

1. Loi de Zipf

La loi de Zipf également appelée loi rang-fréquence, c'est une loi empirique. Cette loi est mise en évidence par le linguiste américain George K. Zipf (1949) [Caron, 04]. Dans [Yaël, 09] l'auteur a dit que « la fréquence d'un mot est inversement proportionnelle à son rang dans la liste des termes classés par fréquence décroissante ou encore que le produit de la fréquence de n'importe quel mot par son rang est constant ». La relation est donnée par la formule suivante : $\text{Rang} \times \text{fréquence} = \text{constante}$ Le rang d'un mot est sa position dans cette liste [Baziz, 05].

Donc l'utilisation de loi de Zipf dans le domaine de (RI) c'est pour déterminer les mots qui représentent au mieux le contenu d'un document, par l'élimination des termes trop fréquents ou trop rares. Le rapport entre le rang \times fréquence et l'importance du terme est représenté par les courbes suivant (Figure I.3) [Yaël, 09] :

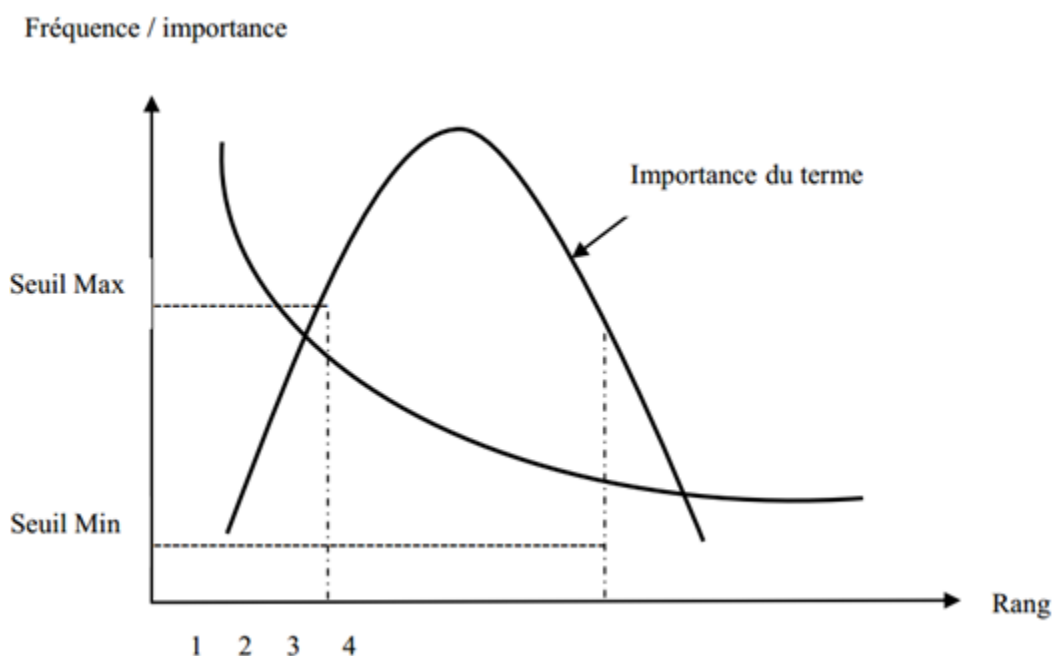


Figure I.3: Le rapport entre le rang \times fréquence et l'importance du terme) [Yaël, 09].

2. Loi de tf*idf

- Le nom tf*idf est très connu dans le milieu de la RI.
Cela désigne un ensemble de schémas de pondération (et de sélection) de termes.
- Par tf, on désigne une mesure qui a rapport à l'importance d'un terme pour un document.
- Par idf, on mesure si le terme est discriminant (ou non-uniformément distribué).
- En général, cette valeur est déterminée par la fréquence du terme dans le document.

On donne quelques formules de tf et d'idf souvent utilisées.

1. $tf = \text{fréquence d'occurrence du terme dans un document } f(t,d);$
 $tf = f(t,d) / \text{Max}[f(t,d)]$ où $\text{Max}[f(t,d)]$: fréquence maximale des termes dans d;
 $tf = \log(f(t,d))$
 $tf = \log(f(t,d) + 1)$
2. $idf = \log(N/n)$ où N est le nombre de documents dans le corpus, et n ceux qui contiennent le terme
3. Finalement, on peut aussi imposer certaine normalisation sur les valeurs calculées.

Une formule de tf*idf est donc la multiplication d'une tf par une idf.

- Avec une telle formule, on peut donc choisir à garder seulement les termes dont la valeur de tf*idf dépasse certain seuil

4.2. Le résultat de l'indexation : L'index

Après l'analyse lexicale, l'élimination des mots vides, la normalisation et la pondération. Le résultat de ces phases (l'index) est l'ensemble des descripteurs et leurs pondérations qui vont représenter au mieux le contenu de ces documents (ou bien requête), ces termes peuvent être soit un mot, soit une racine de mot, soit un terme composé, etc....

4.3. La reformulation de la requête

L'utilisateur ne sait pas choisir les bons termes qui expriment le mieux ses besoins d'information [Aliane, 01], [Ihadjaden, 94], [Smail, 98]. En introduisant la reformulation de requête, la RI est alors envisagée comme une suite de formulations et de reformulations de requêtes jusqu'à la satisfaction du besoin d'information de l'utilisateur, la requête initiale permettant rarement d'aboutir à un résultat qui satisfait ce dernier. Il s'agit en particulier d'ajouter des termes à la requête initiale de l'utilisateur et on parle alors d'expansion de la requête de l'utilisateur [Smail, 98], [Gauch, 92].

4.3.1. La reformulation manuelle

Cette approche est associée aux systèmes de recherche booléens. On peut procéder à la reformulation de requête en utilisant un vocabulaire contrôlé (thésaurus ou classification) pour permettre à l'utilisateur de trouver les bons termes pour compléter sa requête.

4.3.2. La reformulation automatique

Lorsque le feedback de pertinence s'accompagne d'une adjonction (et/ou) suppression de termes, on parle de reformulation automatique. La requête de l'utilisateur est remaniée automatiquement, pour intégrer les descripteurs des documents jugés pertinents ou rejetés.

On trouve différentes variantes de cette technique : celles qui sont utilisées automatiquement pour reformuler la requête en augmentant le poids des termes présents dans les documents jugés pertinents et inversement pour diminuer les poids des termes jugés non pertinents.

Le problème avec la reformulation automatique est l'estimation des « bons » termes qui peuvent conduire effectivement à une amélioration du processus de recherche car l'introduction des termes inappropriés peut entraîner un silence ou au contraire augmenter un bruit.

4.3.3. La reformulation interactive

Dans une reformulation interactive, l'utilisateur joue un rôle actif. A l'inverse de la reformulation automatique, ici, ce sont le système et l'utilisateur qui sont, ensemble, responsables de la détermination et du choix des termes candidats à la reformulation. Le système joue un grand rôle dans la suggestion des termes, le calcul des poids des termes et l'affichage à l'écran de la liste ordonnée des termes. L'utilisateur examine cette liste et

décide du choix des termes à ajouter dans la requête. C'est donc l'utilisateur qui prend la décision ultime dans la sélection des termes.

5. La notion de pertinence

Selon les premières définitions de la pertinence, la pertinence est la correspondance entre un document et une requête ; une mesure de l'informativité du document à la requête; un degré de relation (chevauchement, etc.) entre le document et la requête ; etc.

La notion de pertinence est le critère primaire pour l'évaluation des systèmes de recherche d'informations. Le processus de jugement de la pertinence de l'information est basé sur le degré de similitude de la représentation de la requête avec le contenu du document retrouvé par le système. [Daoud, 09]

On peut distinguer deux types de pertinence, la pertinence système et la pertinence utilisateur [Kompaoré ,08].

- ✚ **La pertinence système** : représente la relation entre la requête et l'information portée par les documents restitués pour le SRI. Elle définit <<une correspondance entre un document et une requête, ou encore une mesure d'informativité du document a la requête>> [Boughaneme& al, 08].
- ✚ **La pertinence utilisateur** : c'est une mesure subjective qui représente la satisfaction de l'utilisateur vis-à-vis des documents retournés par le système.

6. Les modèles de recherche d'information

Un modèle de RI a pour rôle de fournir une formalisation du processus de RI et un cadre théorique pour la modélisation de la mesure de pertinence. Il existe un grand nombre de modèles de RI textuelle développés dans la littérature. Ces modèles ont en commun le vocabulaire d'indexation basé sur le formalisme mots clés et diffèrent principalement par le modèle d'appariement requête-document.

Les principaux modèles de RI sont : le modèle booléen, le modèle vectoriel et le modèle probabiliste. Ces modèles sont largement décrits dans la littérature.

7. Evaluation d'un système

La qualité d'un système doit être mesurée en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir, plus les réponses du système correspondent à celles que l'utilisateur espère, mieux est le système.

La comparaison des réponses d'un système pour une requête avec les réponses idéales nous permet d'évaluer les deux métriques suivantes:

7.1. Précision

La précision mesure la proportion de documents pertinents retrouvés parmi tous les documents retrouvés par le système.

$$précision = \frac{\#documents\ pertinents\ retrouvés}{\#documents\ retrouvés}$$

7.2. Rappel

Le rappel mesure la proportion de documents pertinents retrouvés parmi tous les documents pertinents dans la base.

$$rappel = \frac{\#documents\ pertinents\ retrouvés}{\#documents\ pertinents\ dans\ la\ base}$$

Un système idéal obtient des taux de précision et de rappel proches de 1. Un système qui obtiendrait une précision de 1 et un rappel de 1 signifie qu'il retrouve tous les documents pertinents, et uniquement les documents pertinents : la pertinence système et la pertinence utilisateur seraient confondues.

Pour faire une évaluation de la qualité d'un système de recherche d'information, on utilise habituellement une collection de test. Une collection de test contient un corpus de documents, un ensemble de requêtes, et la liste des documents pertinents pour chaque requête. Cette liste de réponses idéales est établie par des experts ayant une grande connaissance du corpus et du domaine des documents. Pour chaque requête, on établit

alors une courbe de la précision en fonction du rappel. La moyenne de ces courbes constitue un profil visuel de la qualité d'un système.

7.3. La courbe de Précision-Rappel

Dans le cas d'un système idéal, le taux de précision est égal au taux de rappel, c'est-à-dire que, tous les documents pertinents dans ce cas, et que ceux-ci, sont sélectionnés. On aurait donc une droite.

En pratique, la courbe de Précision-Rappel à l'allure générale de la Figure suivante. Pour ce faire on procède comme suit :

Pour $i=1,2,3,\dots\#de_documents_dans_la_base$ faire :

Evaluer la précision et le rappel pour les i premiers documents dans la liste des réponses du système.

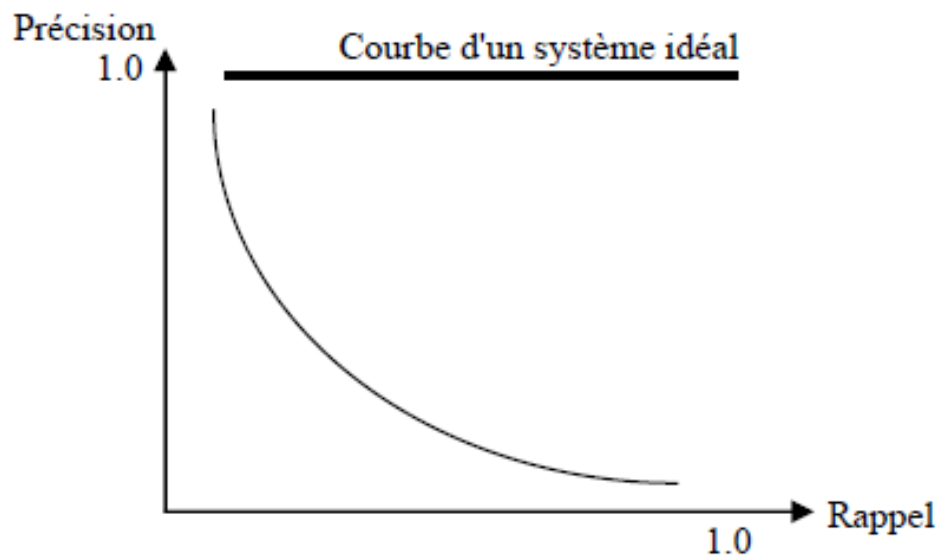


Figure1.4 : Allure d'une courbe de rappel-précision.

8. Conclusion :

Dans ce chapitre nous avons passé en revue les principaux concepts de la recherche d'information. nous avons, particulièrement , introduit des notion de base , telles que le besoin d'information ,la requête ,le document et la pertinence .Nous avons aussi décrit les processus de base de la recherche d'information , à savoir l'indexation, l'appariement requête document et la reformulation de la requête. Ensuite nous avons étudié les différents modèles de la RI. En fin l'évaluation des systèmes de recherche d'information est traitée.

Dans le chapitre suivant, nous dressons l'utilisation des ontologies et du sens des mots de façon générale en recherche d'information.

Nous développerons ci-après le concept d' « ontologie ».

CHAPITRE II

Les ontologies

1. Introduction

Dans ce chapitre nous donnerons des définitions et des notions fondamentales relatives aux ontologies. Les ontologies ont été introduites en Intelligence Artificielle (IA) il y a 25ans, le terme d'ontologie est cependant usité en philosophie depuis le XIXème siècle. Dans ce domaine, l'ontologie désigne l'étude de ce qui existe, c'est à dire l'ensemble des connaissances que l'on a sur le monde. En IA, de façon moins ambitieuse, on ne considère que des ontologies, relatives aux différents domaines de connaissances. C'est à l'occasion de l'émergence de l'Ingénierie des Connaissances que les ontologies sont apparues en IA, comme réponses aux problématiques de représentation et de manipulation des connaissances au sein des systèmes informatiques [ZIANI, 05].

2. Les ontologies

« Ontologie » est un mot de l'informatique issu de domaine philosophique apparut au début des années 90 [Gruber, 93]. Les ontologies sont introduit au champ de l'intelligence artificielle (IA) et de la RI afin de représenter les connaissances, partager l'information et faciliter la communication.

2.1. Définition des ontologies

Cette notion a été reprise par les chercheurs dans le domaine de l'intelligence artificielle et utilisée dans le cadre de construction des systèmes à base de connaissances. L'idée était de séparer, d'un côté, la modélisation des connaissances d'un domaine, et d'un autre côté, l'utilisation de ces connaissances (i.e. le raisonnement).

Dans ce contexte, plusieurs définitions des ontologies ont été proposées. La première a été proposée par [Neches, 91]: « Une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui permettent de combiner les termes et les relations afin de pouvoir étendre le vocabulaire ».

Cette définition descriptive donne un premier aperçu sur la manière de construire une ontologie, à savoir l'identification des termes et des relations d'un domaine ainsi que les règles pouvant s'appliquer sur ces derniers.

Deux années plus tard, [Gruber, 93] donne la définition qui est devenue la plus utilisée dans la littérature : «Une ontologie est une spécification explicite d’une conceptualisation».

Exemple d’ontologie:

L’ontologie de la figure II.1 présente un exemple d’ontologie sur les formes géométriques. Cette ontologie contient un ensemble de concepts, comme ‘Triangle’, un ensemble de relation, comme ‘estPartieDe’ entre ‘figure’ et ‘segment’, et d’attributs, comme ‘aPourLangueur’, des instances, comme ‘ABC’ et enfin des types de données, comme entier [Bendaoud, 09].

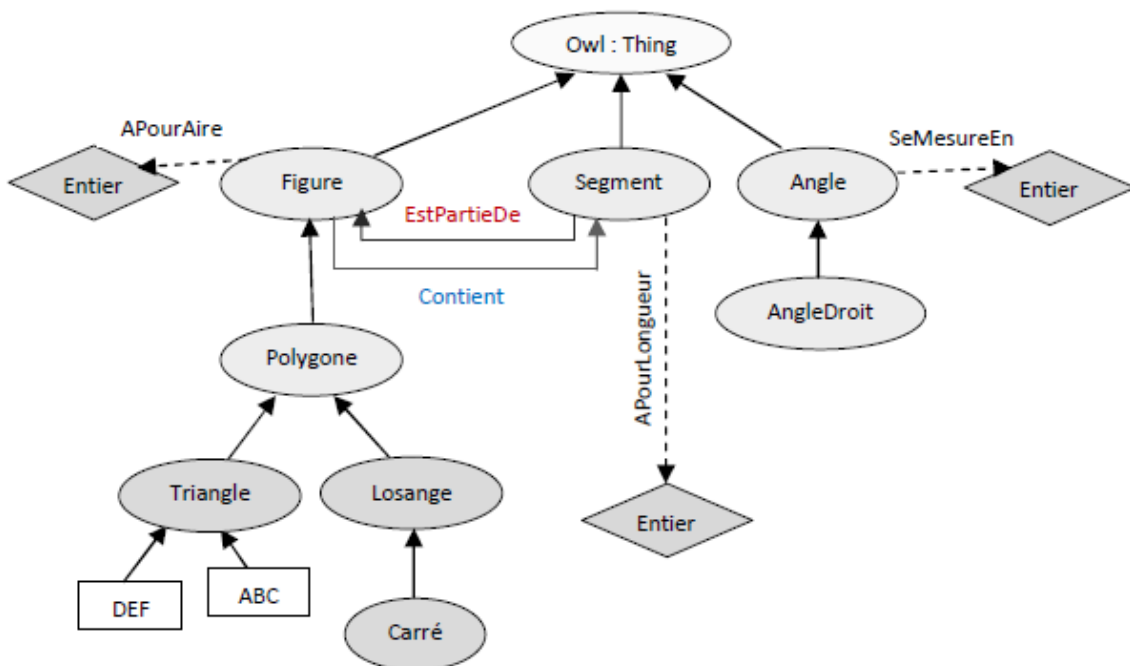


Figure II.1: Une partie de l’ontologie des formes géométriques [Bendaoud, 09].

2.2. Rôles des ontologies

Historiquement, la notion d’ontologie est apparue pour satisfaire des besoins d’interopérabilité dans les systèmes informatiques et de réutilisation. On attend d’elles qu’elles améliorent la communication non seulement entre machines, mais aussi entre humains et machines ou encore entre humains par le biais de logiciels. Les propriétés de ce type de structure de données ont permis de diversifier leur utilisation à différentes applications, en particulier la gestion des connaissances et le Web sémantique. Elles sont utilisées pour : [Ottens, 07]

- Résoudre des problèmes de compréhension et faciliter le partage des connaissances entre personnes de spécialités différentes ;
- Assurer l'interopérabilité entre applications à base de connaissances
- Accéder à des ressources hétérogènes.
- Permettre la réutilisation de modèles de connaissances;
- Faciliter la communication entre agents logiciels.
- Annoter des ressources à l'aide de méta-données.
- Améliorer les processus de recherche d'informations.

2.3. Les principaux types d'ontologies

[Van Heijst, 97] définit deux grandes typologies d'ontologies : une typologie fondée sur la structure de la conceptualisation et l'autre fondée sur le sujet de la conceptualisation.

Dans la première typologie, ils distinguent trois catégories à savoir

- Les ontologies terminologiques (lexiques, glossaires...);
- Les ontologies d'information (schéma d'une BD);
- Les ontologies des modèles de connaissances.

Dans la seconde typologie, qui est la plus citée, ils distinguent quatre catégories :

- **Les ontologies d'application** : elles contiennent toutes les informations nécessaires pour modéliser les connaissances pour une application particulière.
- **Les ontologies de domaine** : elles fournissent un ensemble de concepts et de relations décrivant les connaissances d'un domaine spécifique.
- **Les ontologies génériques** : elles sont similaires aux ontologies de domaine, mais les concepts qui y sont définis sont plus génériques et décrivent des connaissances tels que l'état, l'action, l'espace et les composants. Généralement, les concepts d'une ontologie de domaine sont des spécialisations des concepts d'une ontologie de haut niveau.
- **Les ontologies de représentation (méta-ontologies)** : elles fournissent des primitives de formalisation pour la représentation des connaissances. Elles sont généralement utilisées pour écrire les ontologies de domaine et les ontologies de haut niveau.

2.4. Les ontologies les plus connu du domaine médical

Il est nécessaire pour une communauté scientifique médicale de disposer d'un vocabulaire de consensus pour ce domaine, on trouve par exemple :

2.4.1. UMLS

UMLS est un recueil de plusieurs vocabulaires dans le domaine médical, il fournit une structure de cartographie de vocabulaires et permet la traduction entre des différents systèmes de terminologies.

Le Cœur d'UMLS est constitué d'un Meta thésaurus qui comporte plus de 1.300.000 concepts dans sa version 2006 et qui croît régulièrement de près de 200.000 concepts par an. C'est la partie la plus évolutive d'UMLS, et le rythme de sa croissance pose la question de sa maîtrise et de sa cohérence [HADJAOUI, 12] .

2.4.2. SNOMED CT

SNOMED-CT (Systematized Nomenclature of Medicine, ClinicalTerms) combine la SNOMED Reference Terminology (SNOMED RT) et la Version 3 de la United Kingdom'sClinicalTerms (anciennement connue sous le nom de Read Codes). Son usage se retrouve dans plusieurs contextes : l'indexation de documents cliniques, l'aide à la décision clinique, l'indexation d'images, etc. La terminologie est organisée en une hiérarchie de 18 catégories de premier niveau : les procédures, les entités observables, les structures du corps humain, les événements, etc. Ces entités, dites majeures, sont regroupées autour d'une racine appelée Top. Dans sa version de Juillet 2006, SNOMED CT totalise plus de 300.000 concepts et 770.000 descriptions en anglais [HADJAOUI, 12] .

2.4.3. MeSH

MeSh [Coletti et al.,01] est un vocabulaire normalisé utilisé dans l'indexation des document biomédicaux qui sont issus des bases de données de la National Library of Medicine(NLM).

Un terme dans le vocabulaire MeSh est composé d'un seul mot ou d'un ensemble de mots synonymes représentant un concept biomédical. Chaque concept dans ce thésaurus est désigné par un terme préféré (dit concept préféré) représentant le sens le plus large de ce concept. A titre d'exemple, le concept pain désigne le concept préféré des termes{Pain,Ache;Pain,Burning;Pain,Crustring;Pain,Migratory;Pain,Radiating;Pain,Splitting;Su ffring,Physical}.

2.4.4. Galen

GALEN est une initiative européenne pour le développement de systèmes multilingues de terminologies médicales. Galen est basée sur un modèle sémantique puissant de terminologie clinique, le GalenCoding Reference (CORE), et utilise un langage de représentation appelé GRAIL (GALEN Représentation And IntegrationLanguage) dont le noyau est basé sur les logiques de description. Ce modèle consiste en une hiérarchie de subsomption d'entités élémentaires (concepts primitifs des DL) et un ensemble de déclarations liant ces entités. Le modèle terminologique permet une définition explicite des concepts dans le domaine médical [HADJAOUI, 12].

3. Ontologies et recherche d'information

Un des enjeux actuels de la RI est de développer des systèmes capables d'intégrer plus de sémantique dans leurs traitements. L'idée est d'avoir une solution au problème de confusion entre le besoin de l'utilisateur exprimé par une requête et le domaine exprimé par une collection de mots c.a.d. (parlé le même langage). Pour cela les ontologies interviennent afin d'améliorer la qualité des mots restitués par les SRI à partir de document. Elles sont utilisées pour représenter des descriptions partagées et plus ou moins formelles de domaines et ainsi ajouter une couche sémantique aux systèmes informatiques.

La question qui se pose à ce niveau est : Dans le domaine de la recherche d'information électronique tel qu'il est connu actuellement en utilisant des SRI, comment une ontologie peut-elle être associée au processus de recherche d'information ? Autrement, à quel niveau de SRI l'ontologie peut intervenir ?

3.1. Quelles ontologies choisir

Une première solution vise à construire une ontologie à partir du ou des corpus sur lesquels les tâches de RI vont être réalisées. Cette solution assure a priori l'adéquation entre l'ontologie construite, le corpus et la tâche à réaliser. Cette solution n'est pas toujours adaptée: elle est coûteuse et ne prend pas en compte l'existence de ressources qui pourraient être réutilisées. Maintenant avec l'avènement de domaine des ontologies, elles sont devenues des standards à réutiliser. Une autre solution très utilisée par la majorité des approches de RI visant à intégrer ces ontologies dans ces approches [Baziz, 05].

A la fin nous pouvons dire que les ontologies choisies à être utiles dans la RI doivent être adaptées à la tâche de RI considérée et plus particulièrement apporter de la connaissance utile pour l'interprétation et la compréhension par le système des informations contenues dans les documents.

3.2. Principe d'utilisation des ontologies par un système de recherche d'information

L'ontologie peut intervenir aux différentes phases dans le processus de recherche d'information.

3.3. La reformulation de la requête par utilisation des ontologies

Les utilisateurs expriment leurs besoins difficilement à l'aide d'une requête mal écrite. A ce point, l'ontologie intervient pour aider l'utilisateur à formuler leur requête afin d'exploiter efficacement la collection de documents.

Il existe deux types d'expansion de requête dans la littérature : La première consiste à utiliser des ressources, internes ou externes, comme par exemple un dictionnaire [Moldovan&al., 99] ou bien WordNet [Voorhes, 94], elle est basée sur l'extension des requêtes par l'ajout de nouveaux termes en relation avec les termes de la requête. La seconde solution consiste à faire la réinjection de pertinence en reposant sur l'analyse des termes contenus dans le document jugés pertinents pour la requête initiale. L'idée est que l'ajout de termes liés aux termes initiaux de la requête peut permettre de retrouver les mots qui ne sont pas restitués auparavant.

Harman [Harman, 92] a prouvé que la reformulation de requêtes a des effets positifs en RI. L'objectif de la reformulation est soit de limiter le silence (Le silence fait référence aux documents pertinents mais qui ne sont pas retrouvés par le système) soit de réduire les risques de bruit (le bruit fait référence aux documents non pertinents retrouvés par le système).

Dans le premier cas, la requête est étendue à partir de termes similaires à ceux de la requête initiale. Dans le second cas la requête initiale est étendue ou modifiée à partir de termes qui ajoutent de l'information complémentaire à la représentation du besoin.

3.3.1. Appariement à partir d'ontologies

L'approche citée dans [Andreasen& al, 03] donne l'avantage à l'ontologie de mesurer la similarité entre la représentation des requêtes et des documents dans le cas où une ontologie unique décrit ces deux représentations.

Les documents et requêtes sont représentés à partir du langage et de l'ontologie. Cette ontologie contient un ensemble de concepts et de relations entre concepts, dont la relation de subsomption. Elle est considérée comme un graphe orienté. L'avantage du calcul de la similarité est de classer les documents restitués par rapport à leur similarité à la requête, cette similarité reposant sur l'organisation des concepts dans l'ontologie [Hernandez & al.,08].

3.3.2. Indexation des documents par les ontologies

La démarche issue du domaine de la RI consiste à choisir comme langage de représentation des documents, l'ensemble des concepts et instances de l'ontologie. Les documents sont alors indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambigus [Aussenac&Mothe, 04]. L'ontologie utilisée dans ce cas reflétant le ou les domaines de connaissance étudiés à la collection.

4. Conclusion

Les ontologies sont utilisées par les applications qui ont besoin d'ajouter la sémantique, d'où viennent l'importance d'utilisation des ontologies et les ressources lexicales dans les systèmes de recherche d'information.

CHAPITRE III

Conception et mise en Œuvre

1. Introduction

Avec la croissance du domaine médicale, les informations partagées dans ce contexte deviens plus en plus énorme. Les outils de recherche actuels sont très pauvres et ne proposent que des solutions de dépannage. La recherche d'information médicale est devenu un véritable casse-tête d'où des résultats de mouvais qualités et non satisfaisant pour l'utilisateur qui en utilisant un moteur censé lui faciliter les recherches découvre une énorme listes de liens sans rapport évident avec ces besoins. De notre point de vue, la solution vient de l'utilisation des techniques de traitement sémantique et syntaxique des mots médicaux qui constitue un grand pas vers son intégration dans la technologie de l'information. La prise en charge de la richesse des mots médicaux lors de la recherche d'information est très nécessaire pour avoir des résultats de recherche plus satisfaisants et plus riche.

C'est également l'objectif de notre travail tel que dans ce dernier chapitre nous allons essayer de développer un système de recherche d'information spéciale pour les termes médicaux. On essaye de résoudre le problème de variation conceptuelle par l'intégration de l'ontologie médicale dans les différentes étapes de SRI, en vue d'obtenir des résultats satisfaisants.

2. Problématique

Le fonctionnement des systèmes de recherche d'information actuels opèrent essentiellement sur les termes (en mesurant leur présence ou leur absence) sans faire appel à des relations sémantiques. Ils ont tendance à accumuler des informations toujours plus nombreuses, provoquant ainsi un bruit certain, difficile à maîtriser par les utilisateurs.[Nassr,01]

Par exemple la requête avec le mot clé « malade » ne cherche pas les documents qui contiennent son synonyme « patient ». De même il n y a pas de distinction entre eux. La requête ne prévoit pas les cas où différents mots sont utilisés pour représenter le même sens. Pour ce phénomène, il n y a pas distinction dans les cas où des mots sont au singuliers, pluriel : « grippal », « grippaux » et de conjugaison « vomis », « vomissons », « vomirions ».

Le problème se pose : Comment découvrir le document rare ou caché, celui qui contient de réelles informations répondant à la requête ? Comment sélectionner des documents, les informations pertinentes pour des objectifs précis ?

Pour localiser et délivrer des documents à un utilisateur en fonction de son besoin en information, une manière pertinente est de chercher l'ensemble de contenu dans un document, et d'introduire plus de sémantique dans la recherche d'information sur les textes. On a intégré l'ontologie dans la phase de recherche. Ce qui permettra de combiner un modèle de recherche d'information classique qui est le modèle vectoriel avec l'exploitation de l'ontologie. Ceci permettra aussi d'améliorer la précision de recherche par la compréhension de l'objectif de recherche et la signification contextuelle des termes qui permet d'accéder rapidement et efficacement à l'information désirée.

3. Environnement de développement

3.1. Langages utilisés

3.1.1. Java

Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton de Sun Microsystems. Il permet de créer des logiciels compatibles avec de nombreux systèmes d'exploitation (Windows, Linux, Macintosh, Solaris). Le langage Java donne aussi la possibilité de développer des programmes pour téléphones portables et assistants personnels. Enfin, ce langage peut être utilisé sur internet pour des petites applications intégrées aux pages web (applet) ou encore comme langage serveur (JSP).

3.1.2. MySQL

MySQL est un serveur de bases de données relationnelles Open Source. Un serveur de bases de données stocke les données dans des tables séparées plutôt que de tout rassembler dans une seule table. Cela améliore la rapidité et la souplesse de l'ensemble. Les tables sont reliées par des relations définies, qui rendent possible la combinaison de données entre plusieurs tables durant une requête. Le SQL dans "MySQL" signifie "Structured Query Language" : le langage standard pour les traitements de bases de données.

3.2. OUTILS

3.2.1. Eclipse

Eclipse IDE est un environnement de développement intégré libre, le terme Eclipse désigne également le projet correspondant, lancé par IBM.

Il est extensible, universel et polyvalent, permettant potentiellement de créer des projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en Java à l'aide de la bibliothèque graphique SWT, d'IBM.

3.2.2. L'API JENA

Jena est une API java open source développée par le laboratoire de Hewlett-Packard permettant la lecture et la manipulation des ontologies décrites en RDFS ou en OWL.

La page d'accueil de Jena, illustrée ci-dessous, est <http://jena.sourceforge.net/>



Figure III.1: Page d'accueil jena (<http://jena.sourceforge.net/>).

3.2.3. Lucene

Lucene est une bibliothèque open source écrite en Java qui permet d'indexer et de chercher du texte. Il est utilisé dans certains moteurs de recherche. C'est un projet de la fondation Apache mis à disposition sous licence Apache. Il est également disponible pour les langages Ruby, Perl, C++,PHP, C#.

4. Solution proposé

Les systèmes de recherche d'information classiques essayent d'optimiser le temps de recherche en identifiant les mots selon des critères d'appariement entre les mots contenus dans les requêtes utilisateurs (et uniquement ceux-là) et ceux des concepts (les indexes).

La solution proposée dans ce mémoire consiste à intégrer l'ontologie médicale (MESH) aux modules de système de recherche. Plus précisément nous allons utiliser l'ontologie médicale MESH pour indexer les documents et la requête d'utilisateur.

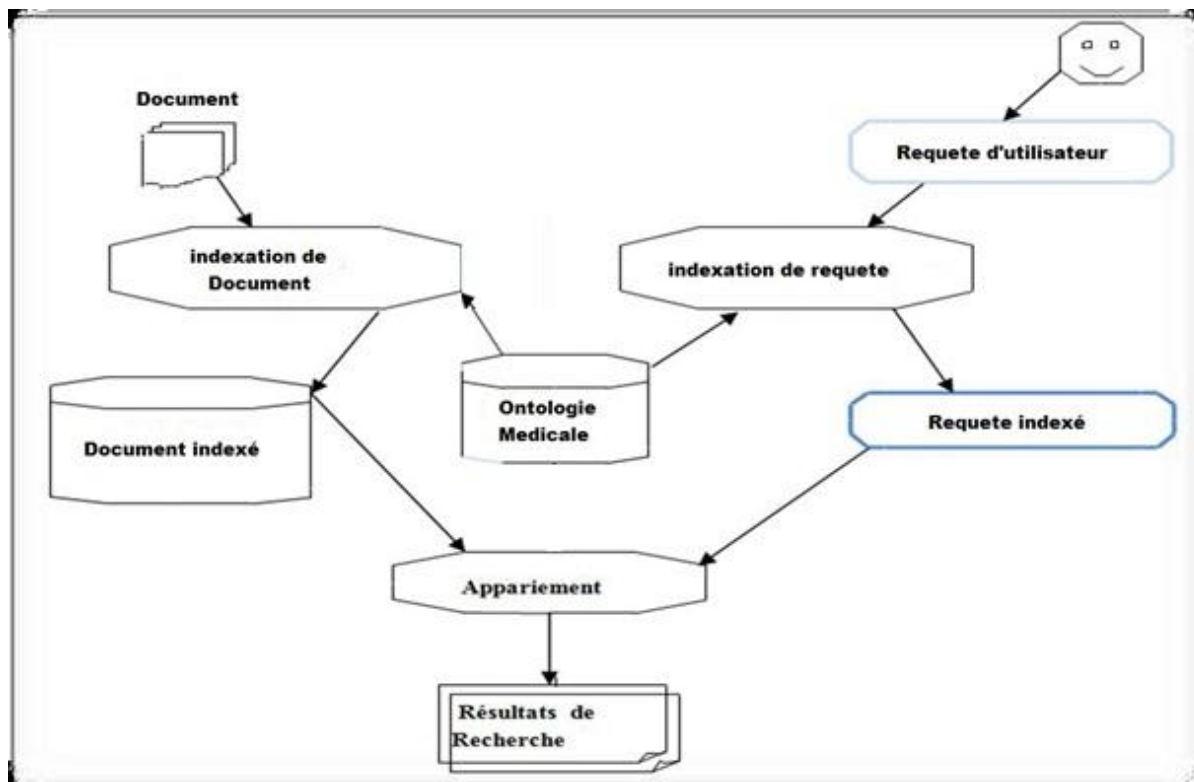


Figure III.2: Architecture de notre système de recherche.

4.1. Ontologie MeSH

MeSH (Medical Subject Headings) est un thésaurus de référence dans le domaine biomédical, publié en 1954 et régulièrement mis à jour par la NLM (National Library of Medicine). Il s'agit d'un vocabulaire normalisé, qui permet d'exprimer une notion donnée d'une manière indépendante du langage courant. Ce vocabulaire est utilisé pour indexer, classer et rechercher des documents, notamment ceux des bases de données de la NLM, dont MEDLINE/PubMed

Le thésaurus de la base MADELINE ou MeSH regroupe 25588 descripteurs pour l'année 2010.

Répartition des descripteurs

Le vocabulaire MeSH est organisé en 16 domaines principaux :

- A. Anatomie
- B. Organismes
- C. Maladies
- D. Produits chimiques et pharmaceutiques
- E. Techniques analytiques, diagnostiques et thérapeutiques et équipements
- F. Psychiatrie et psychologie
- G. Phénomènes et processus
- H. Disciplines et professions
- I. Anthropologie, enseignement, sociologie et phénomènes sociaux
- J. Technologie, industrie et agriculture
- K. Science humaines
- L. Sciences de l'information
- M. Groupes individualisés
- N. Soins de santé

- V. Caractéristiques d'une publication
- Z. Emplacements géographiques

Chaque descripteur est accompagné d'un code alphanumérique :

Une lettre = un domaine

Une série de nombres = position du descripteur dans l'arborescence.

- **Exemple :**

Prion Diseases [C10.228.228.800]

Creutzfeldt-Jakob Syndrome [C10.228.228.800.230]

- **Remarque**

Un descripteur peut appartenir à plusieurs domaines, dans ce cas plusieurs codes alphanumériques sont attribués au descripteur

- **Exemple**

MeSH Heading: Prion Diseases

Tree Number: C10.228.228.800

Tree Number: C10.574.843

4.2. Système proposé de recherche

Nous avons créés deux applications la première qui va faire la recherche intelligente d'information médicale en utilisant l'ontologie MeSH et la 2eme elle va faire la recherche simple ça veut dire sans utiliser l'ontologie.

4.3. Réalisation pratique

4.3.1. Application1 : Recherche simple

Pour montrer l'avantage de la recherche sémantique on a développé une application pour faire la recherche simple.

Notre application se compose d'une fenêtre principale dont laquelle on effectue la recherche

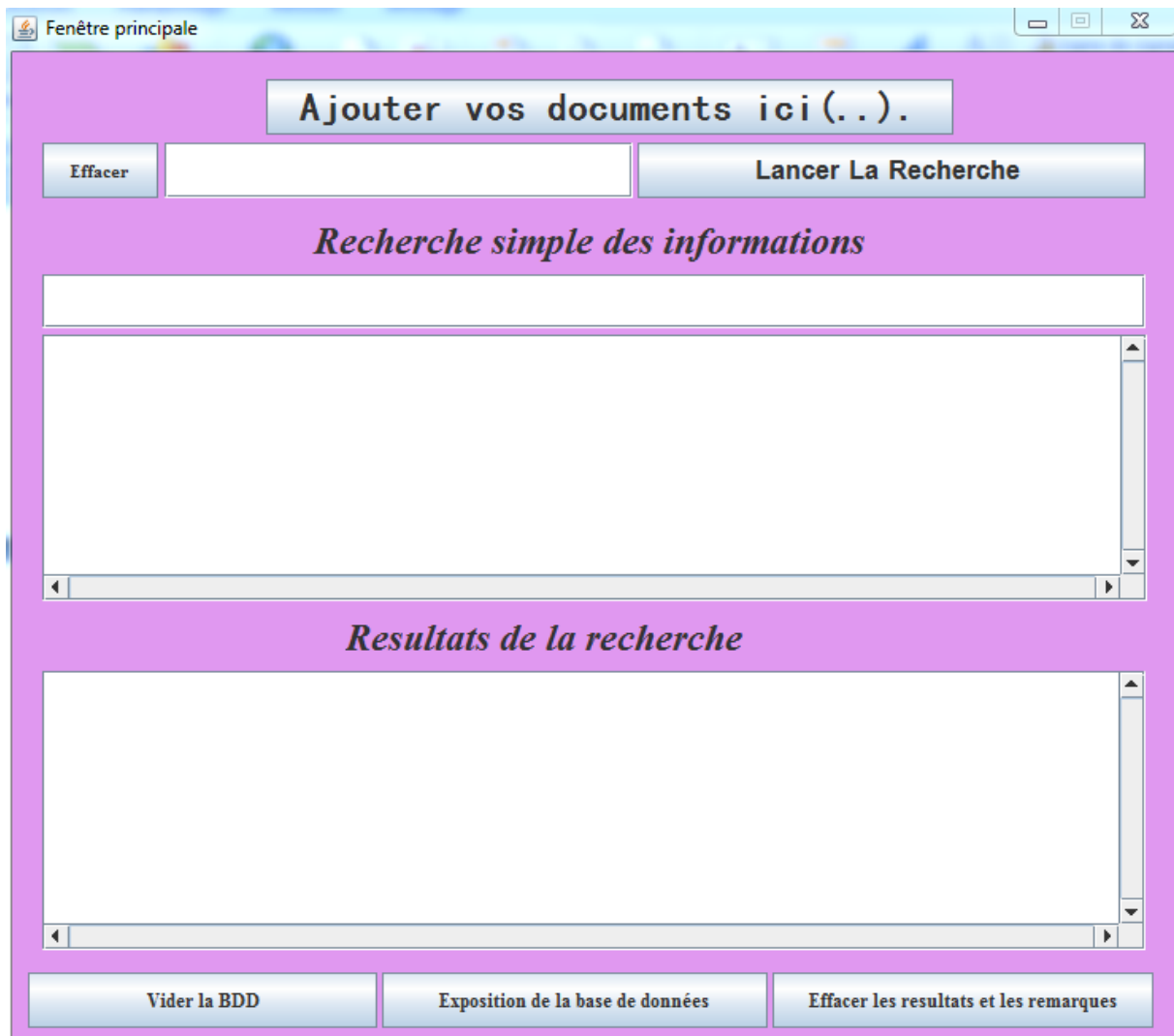


Figure III.3: Fenêtre principale de la recherche.

4.3.1.1. L'ajout des documents :

Pour ajouter des documents, un répertoire est choisi ainsi que les documents à ajouter.

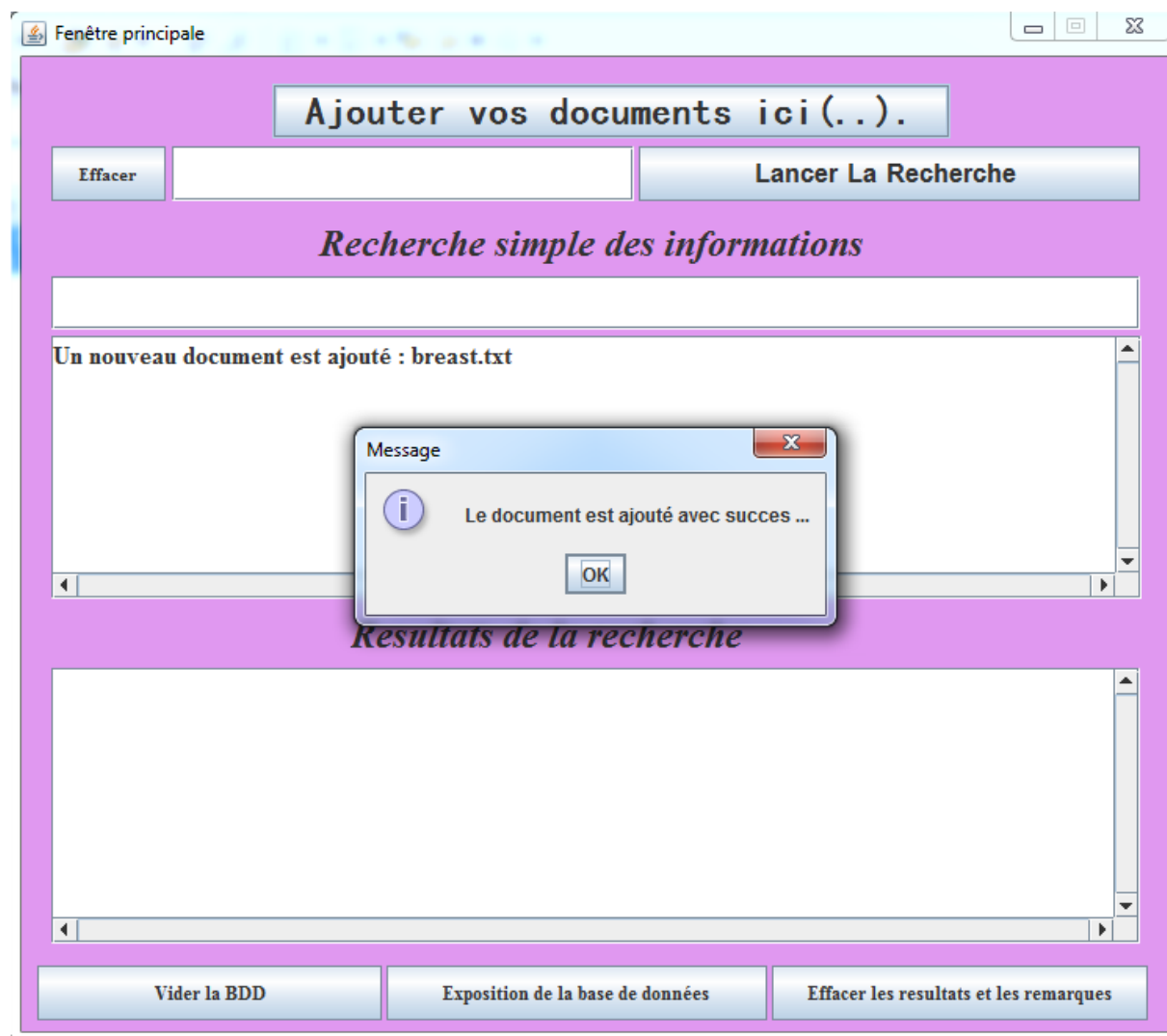


Figure III.4: Ajouter un document.

4.3.1.2. Affichage de résultats



Figure III.5: Affichage des résultats.

4.3.1.3. Affichage de l'index

L'affichage de l'index nous permet de voir la liste des documents qui existe dans l'index ainsi que l'ensemble des termes et leurs tf et idf

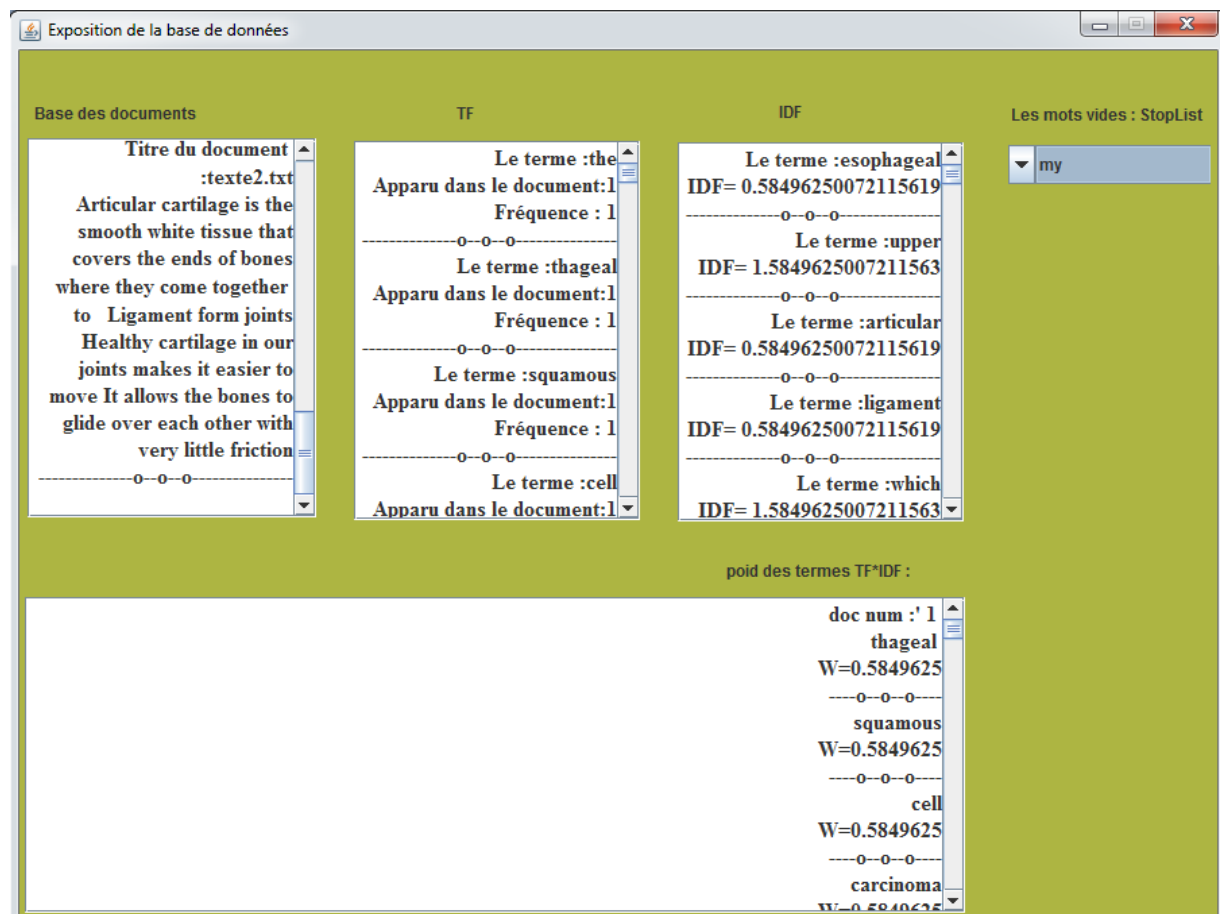


Figure III.6: Affichage du contenu de l'index.

4.3.2. Application2 : Recherche Intelligente d'Information Médicale

Notre application se compose d'une fenêtre dont laquelle on effectue la recherche. On doit entrer une requête pour effectuer cette tâche.

La figure III.7 contient une zone de texte dans laquelle nous allons écrire notre requête et le bouton rechercher pour activer l'opération.

Après la recherche l'application va calculer la Précision et Rappel et le temps d'exécution que nous affichons sur cette dernière.

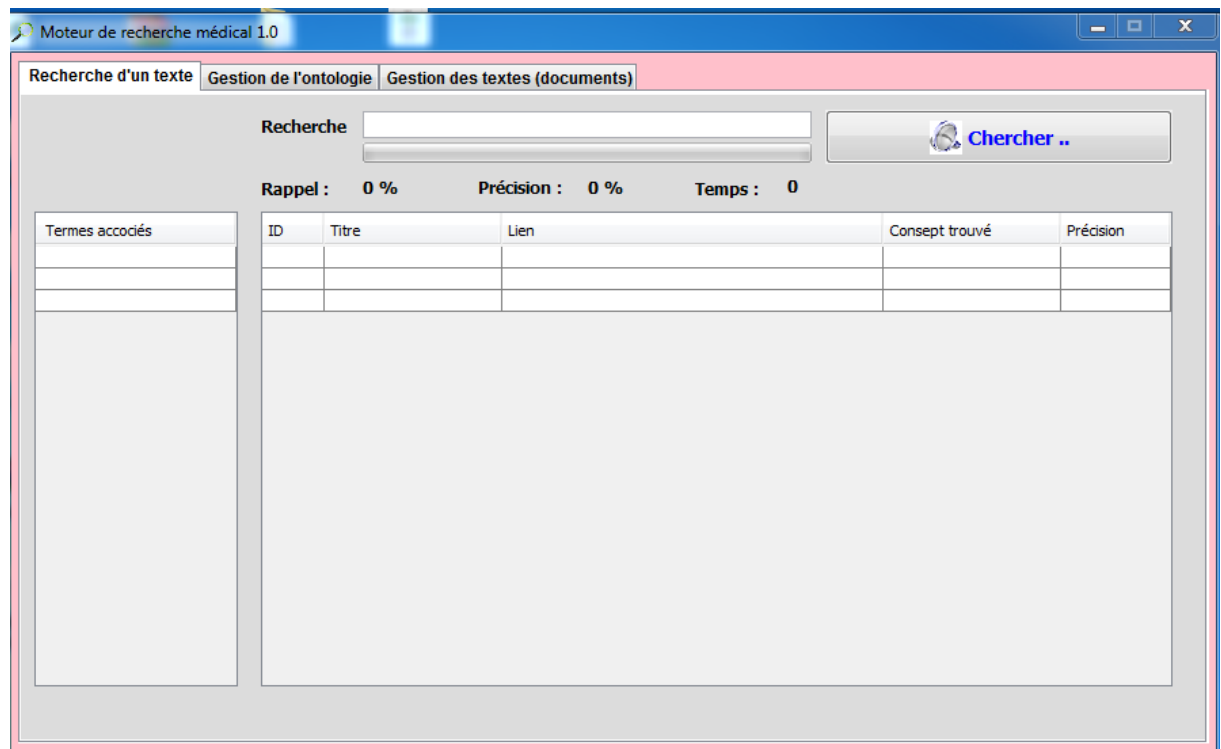


Figure III.7: Fenêtre principale de la recherche.

4.3.2.1. L'ajout des documents :

Pour ajouter des documents, un répertoire est choisi ainsi que les documents à ajouter. Après avoir choisi le chemin des documents, nous allons choisir les documents et cliquer sur Ajouter et indexer pour que les documents seront ajoutés et indexés.

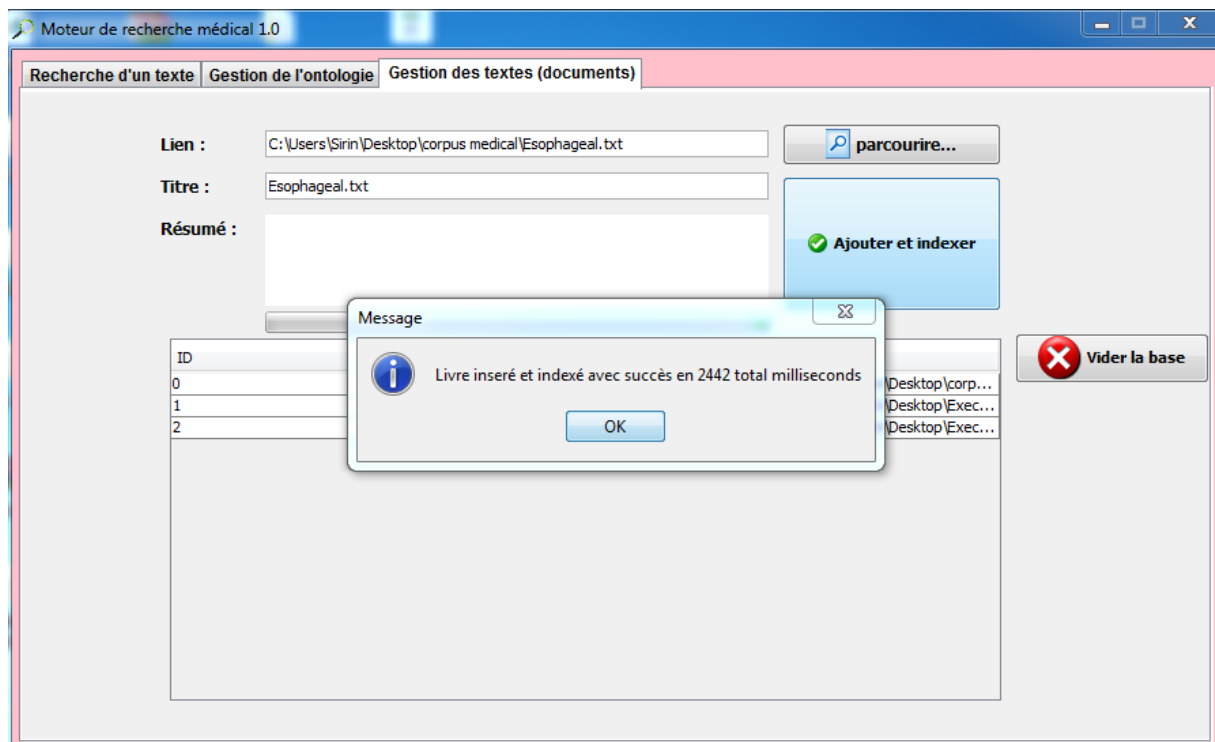


Figure III.8: Ajouter un document.

4.3.2.2. L'ajout de l'ontologie :

Pour ajouter une ontologie, un répertoire est choisi ainsi que l'ontologie à ajouter.

Dans notre cas nous avons utilisé l'ontologie médicale (MeSH).

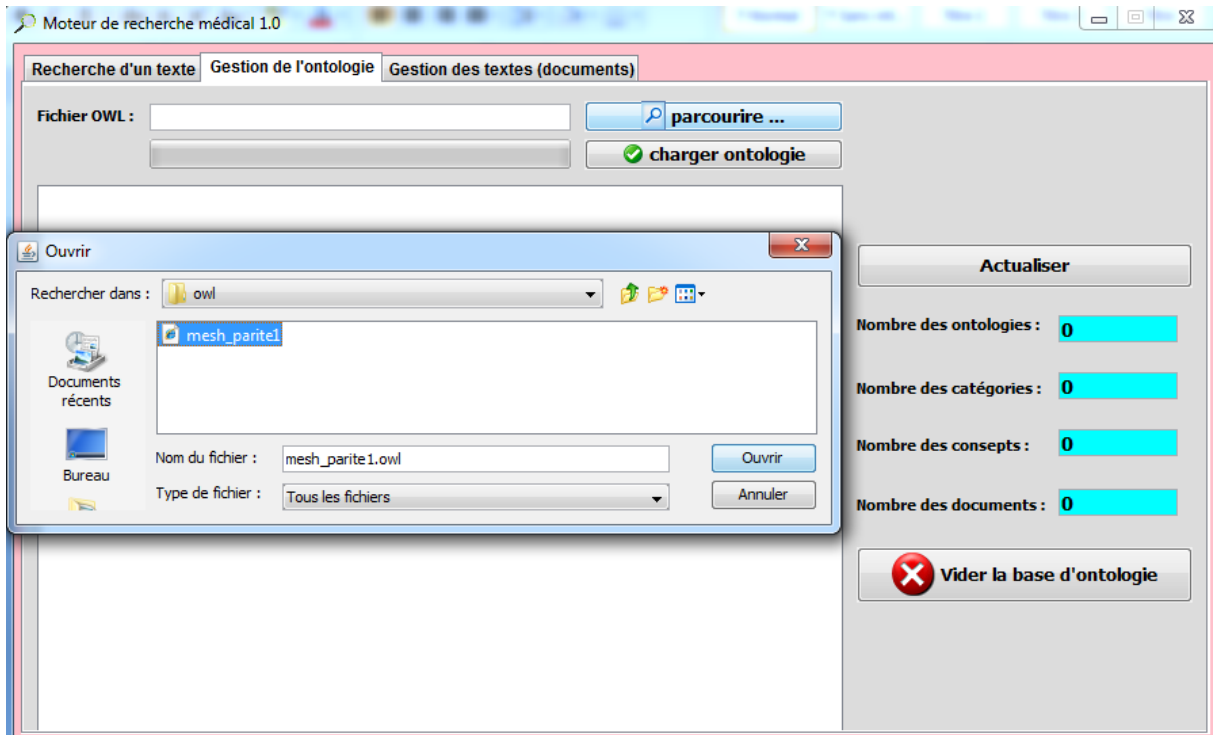


Figure III.9: Ajouter une ontologie.

4.3.2.3. Affichage de résultats

Exemple de recherche :

En effectuant la recherche, le système affiche une liste de documents correspondant à la requête avec l'ordre décroissant des coefficients de similarité.

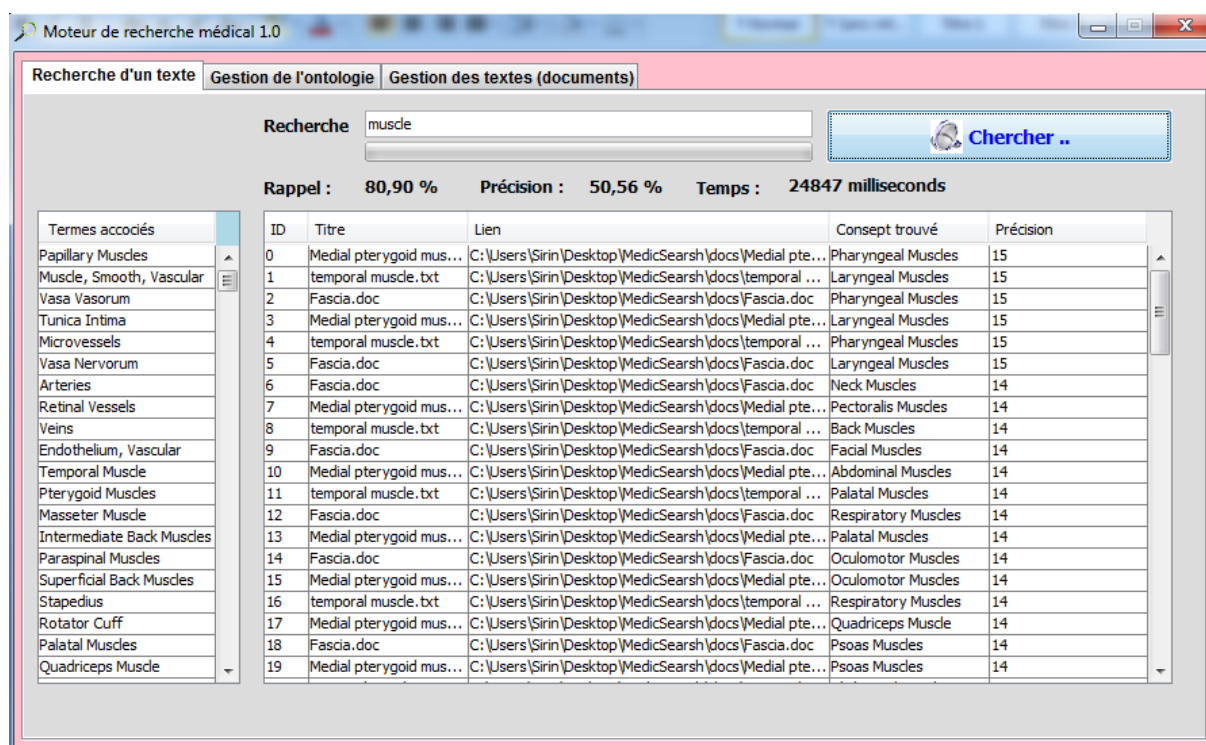


Figure III.10: Affichage des résultats selon l'ordre décroissant.

4.4. Discussion

Dans la deuxième application (Recherche Intelligente d'Information Médicale) nous avons fait une recherche sémantique à base d'ontologie car quand l'utilisateur saisie une requête, notre système va chercher les mots associés dans l'ontologie (les termes de même catégorie). Ensuite, faire la recherche dans les index des documents enfin notre application va nous afficher les documents pertinents à notre requête.

Dans ce cas, on utilise la recherche d'information sémantique, cette dernière ne se fait pas seulement avec les mots clés (comme le cas de la recherche simple) mais va prendre en considération la sémantique des termes en consultant les concepts d'ontologie.

Dans la première application (recherche simple) nous avons fait une recherche simple sans utiliser l'ontologie (se fait seulement avec les mots clés) et cela pour montrer l'avantage de la recherche intelligente et de faire la comparaison entre les deux résultats.

Exemple d'application

Recherche simple :

La requête "Head" ; le système ne retourne aucun document car le terme n'existe pas dans l'ensemble des documents de la collection.

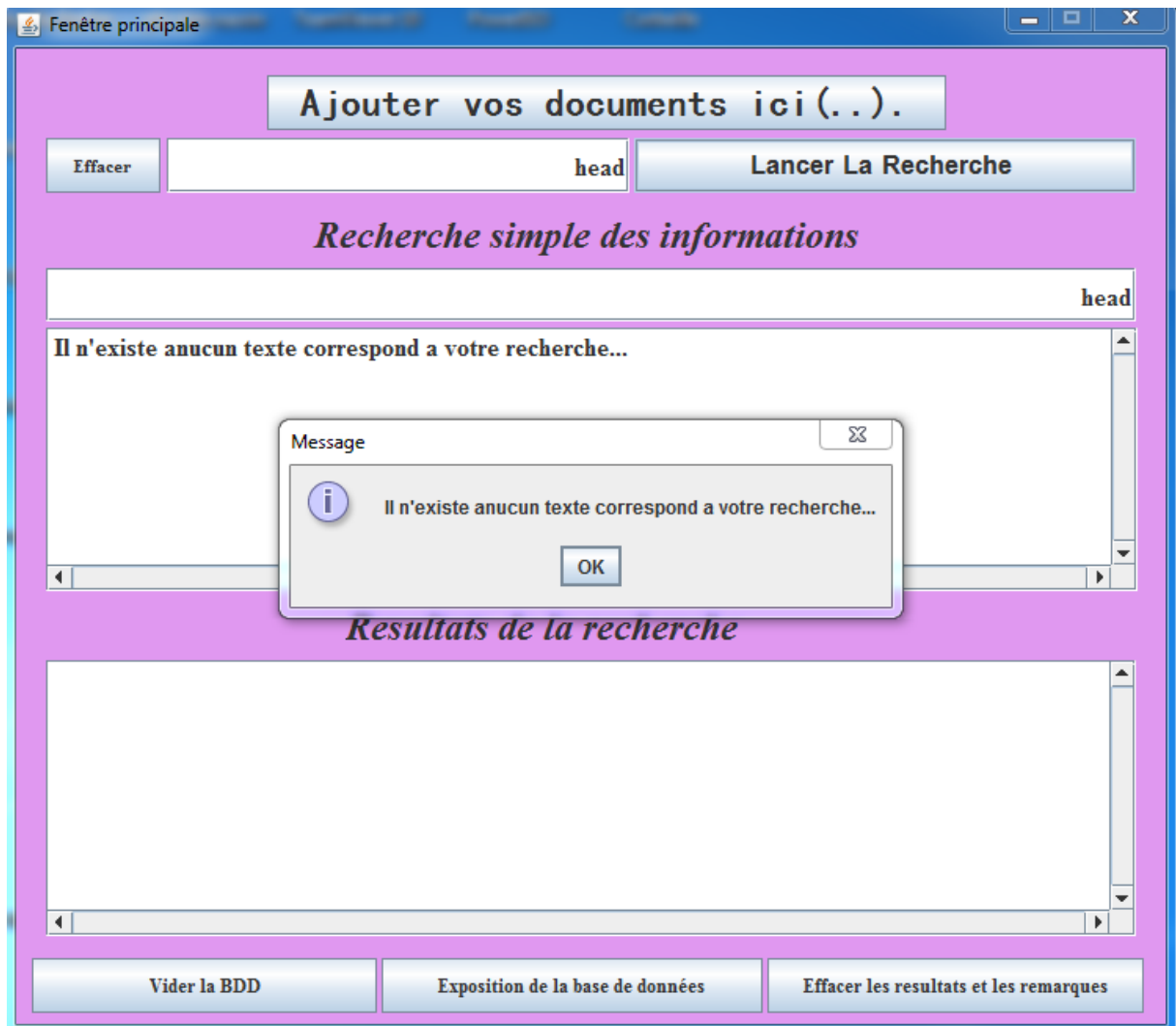


Figure III.11: Résultat de la recherche simple.

Recherche intelligente d'information médicale

Le cas de la recherche sémantique à base d'ontologie, le système va chercher dans la liste des concepts, puisque "Head" fait partie de la liste des concepts. Le système affiche les documents ayant les termes associés de mot "Head" dans l'ontologie MeSH.

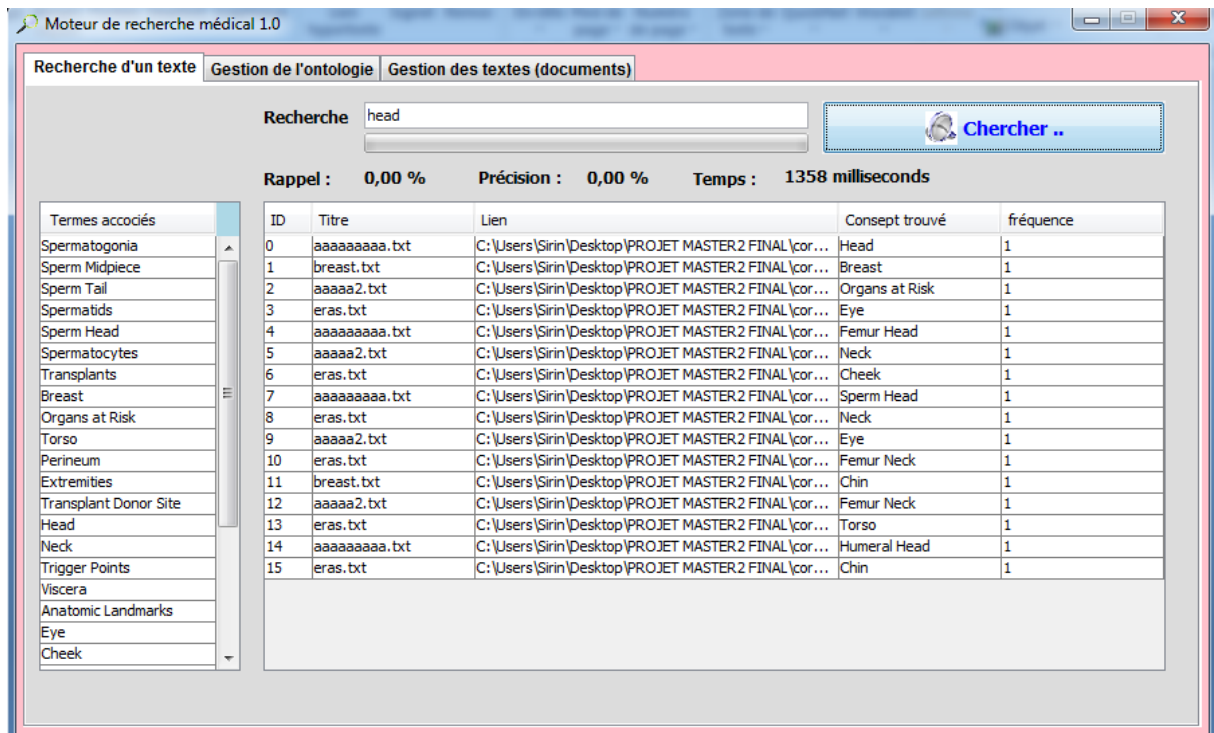


Figure III.12: résultat de la recherche intelligente.

5. Conclusion

Dans ce chapitre, nous avons décrits la partie implémentation de notre système. Cette étude nous a permis aussi de mettre en évidence les différents modules du système. Nous avons par ailleurs, illustré l'architecture globale du système. à savoir :

- les modules de construction : module d'indexation, module de gestion d'ontologie et enfin le module de recherche,
- pour la mise en œuvre on a cité le langage utilisé ainsi que les outils.

CONCLUSION GÉNÉRALE

Conclusion générale

Suite aux grandes quantités d'informations diffusées en médecine, nous sommes obligés de développer des systèmes de recherche plus intelligents pour répondre aux besoins des utilisateurs.

Notre travail s'inscrit dans le cadre d'intégration des ontologies aux différents niveaux des systèmes de recherche d'informations médicale, autrement dit, l'indexation des documents et des requêtes utilisateurs en utilisant une ontologie médicale comme par exemple MeSh.

Les travaux de ce mémoire se situent dans le contexte général de la recherche d'information médicale et plus particulièrement dans le cadre de la recherche d'information sémantique en utilisant un des modèles de la recherche d'information classique qui est le modèle vectoriel à base d'ontologie pour restituer des documents répondant à l'interrogation d'utilisateur.

L'indexation conceptuelle de la requête utilisateur consiste à trouver des nouveaux termes à partir d'un des termes initiaux en utilisant l'ontologie MeSh. Cela permet d'étendre le champ de vision de la requête utilisateur.

Les résultats confirment le fait que la recherche conceptuelle permet d'élargir le champ de recherche qui permet d'enrichir les informations trouvés pour l'utilisateur.

BIBLIOGRAPHIE

Références

- [HADJAOUI, 12] Melle. HADJOUI Fatima Zohra, **Gestion des Ontologies Médicales (G.O.M)**. Thèse de master, université abou bark belkaid,(2012)
- [AZZOUG, 13] Melle. Wassila Azzoug , **contribution à la définition d'une approche d'indexation sémantique de documents testuels** . Thèse de magister, université M'hamad bougara.Boumerdes,(2013)
- [OUANECHÉ, 13] Melle. Ouaneche Linda , **Système de recherche d'information pour la langue arabe : Utilisation du modèle vectoriel à base d'ontologie**. Thèse de master, université Hassiba Ben Bouali,(2013)
- [HAMMACHE,] M.Hammache arezki, **recherche d'information: un modele de langue combinant mots simple et mots composés**. Thèse de doctorat, université Mouloud mammeri
- [MEHIDI & RABAH, 14] M. MEHIDI Tawfiq et RABAH Zakarya, **Conception et implémentation d'un système de recherche à base d'annotations sociales**. Thèse de master, université abou bark belkaid,(2014)
- [ABDERRAHIM, 10] M. Mohammed Alaeddine ABDERRAHIM, **Apport des ontologies dans un système de recherche d'informations arabe** . Thèse de magister, université abou bark belkaid,(2010)
- [LAMRAOUI, 11] M. Lamraoui younes , **Recherche intelligente des informations dans le coran**. Thèse de magister, université abou bark belkaid,(2011)
- [BOURAMOUL, 11] M. ABDELKRIM BOURAMOUL, **RECHERCHE D'INFORMATION CONTEXTUELLE ET SEMANTIQUE SUR LE WEB**. Thèse de Doctorat, université MENTOURI de Constantine,(2011)
- [BAZIZ, 05] M. Mustapha BAZIZ, **INDEXATION CONCEPTUELLE GUIDÉE PAR ONTOLOGIE POUR LA RECHERCHE D'INFORMATION**. Thèse de Doctorat, université PAUL SABATIER,(2005)
- [BOUCHAM, 09] M. Boucham Souhila, **Une approche basée Ontologies pour l'indexation automatique et la Recherche d'Information Multilingue(RIM)**. Thèse de Magister, université Bougara de Boumerdes,(2009)
- [DOUZIDIA, 04] M. Fouad Soufiane Douzidia, **Résumé automatique de texte arabe**. Thèse de M.Sc, université de Montréal, (2004)

- [ROUSSEY, 01] Mlle. Catherine Roussey, **Une méthode d'indexation sémantique adaptée aux corpus multilingues**. Thèse de doctorat, L'Institut National des Sciences Appliquées de Lyon, (2001)
- [MESFAR, 08] M. Slim Mesfar, **Analyse Morpho_Syntaxique Automatique et Reconnaissance des Entites Nommees en Arabe Standard**. Thèse de Doctorat, université de Franche-Comte, (2008)
- [ELBYED, 09] M. Abdeltif Elbyed, **ROMIE, une approche d'alignement d'ontologies à base d'instances**. **Computer Science**. Thèse de M.Sc, Institut National des Télécommunications, France, (2009)
- [ZIANI, 05] M. Ziani Radouane, **Une Indexation A Base d'Ontologies Pour Le Filtrage d'Informations Sur Le Web**. Thèse de Magister, université de Badji Mokhtar Annaba, (2005)
- [Harrathi, 10] M. Harrathi, **recherche d'information conceptuelle dans les documents semi structurés**. Thèse de doctorat de l'institut nationale des sciences appliquées de Lyon, septembre (2010)
- [Yaël, 2009] Yaël Champclaux, **Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information**, Université Toulouse, thèse de doctorat, (2009).
- [Caron, 2004] Yves Caron, **Contribution de la loi de Zipf à l'analyse d'images**, thèse de doctorat, Université François Rabelais de Tours (2004).
- [Aliane,01] H.Aliane, "Towards a knowledge based plat-form for automatic indexing and information retrieval" Séminaire sur l'automatisation du trésor de la langue arabe, alger, (2001).
- [Ihadjaden ,94]M. Ihadjaden, **Conception, réalisation et évaluation d'un système de recherche et de catégorisation automatique d'information textuelle sur Internet**, Thèse de l'université ParisIV, (1994).
- [Smail ,98] M.Smail «Vers des systèmes évolutifs de recherche d'information : un état de l'art » Technique et Science Informatiques, vol. 17, n°10, (1998).
- [Gauch ,92] **An expert system for automatic query reformulation**, Technical report, University of north california, (1992).
- [Boughaneme& al, 08]M.Boughaneme,J.Savoy,editors. **recherche d'information états des lieux et perspectives**. Hermès Science Publication (2008)

[Daoud, 09] M.Daoud, **Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche**, thèse de doctorat, Université Paul Sabatier,(2009).

[Kompaoré, 08] N. Kompaoré, **Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes : vers un processus de RI adaptatif**, thèse de doctorat, Université Paul Sabatier, (2008).

[Hlaoua, 07] L. Hlaoua, **Reformulation de Requêtes par Réinjection de Pertinence dans les Documents Semi-Structurées**, thèse de doctorat, Université Paul Sabatier, (2007).

[Nassr, 02] N. Nassr, **Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes**, thèse de doctorat, Université Paul Sabatier, (2002).

[Bouidghaen, 11] O.R.Bouidghaen, **Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes**, thèse de doctorat, Université Paul Sabatier,(2011).

[Yaël, 09] C.Yaël, **Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information**, Université Toulouse, thèse de doctorat, (2009).

[Gruber, 93] T. R. Gruber, "Toward Principles for the design of Ontologies used for Knowledge Sharing," in *Proc of International Workshop on Formal Ontology*, Padova, Italy, March (1993).

[Neches, 91] R. Neches, R.E. Fikes, T. Finin T, T.R. Ruber, R. Patil, T. Senator, W.R. Wartout. "Enabling technology for knowledge sharing". *AI Magazine*, 12(3), 16- 36, 1991.

[Bendaoud, 09] R. Bendaoud. "Analyses formelle et relationnelle de concepts pour la construction d'ontologies de domaines à partir de ressources textuelles hétérogènes".page 15, Juillet 2009.

[Ottens, 07] K. Ottens. « Un système multi-agent adaptatif pour la construction d'ontologies à partir de textes ». page 14, Octobre 2007.

[Van Heijst, 97] G. Van Heijst, A. Schreiber, B. Wielinga. "Using explicit ontologies in KBS development". *Int. J. of Human-Computer Studies*, 46(2/3):183–292, 1997.

[Moldovan & al., 99] Moldovan D., Harabagiu S., Pasca M., Mihal-cea R., Goodrum R., Girju R., Rus V., **LASSO: A tool for surfing the answer net**. Proceedings of

the 8th Text Retrieval Conference (TREU-8), (1999).

[Voorhes, 94] Voorhes E. M., Query expansion using lexical-semantic relations,
Proceedings of the 17th International ACM SIGIR Conference on Research and
Development in Information Retrieval, pp 61-69, (1994).

**[Andreasen & al., 03] Andreasen T., Bulskov H., Knappe R., Similarity for Conceptual
Querying,** Proceedings for the 18th International Symposium on
Computer and Information Sciences, pp 268-275,(2003).

**[Hernandez & al., 08] Hernandez N., Hubert G., Mothe J., Ralalason B. : RI et
Ontologies –**
Etat de l'art 2008, RAPPORT INTERNE N° IRIT/RR—2008-14—FR JUILLET (2008)

**[Aussenac & Mothe, 04] Aussenac-Gilles N., Mothe J., Ontologies as Background
Knowledge to
Explore Document Collections,** Actes de la Conférence sur la Recherche
d'Information Assistée par Ordinateur (RIAO), pp 129-142, (2004).

Description de Lucene, : <http://fr.wikipedia.org/wiki/Lucene> (consulter le 13 février
2015)

Description phpMySQL : , [https://cours.etsmtl.ca/gpa775/Documents/phpMyAdmin-
MySQL-23-juin-2009.pdf](https://cours.etsmtl.ca/gpa775/Documents/phpMyAdmin-MySQL-23-juin-2009.pdf) (consulter le 13 février 2015)

Description XAMPP : <http://fr.wikipedia.org/wiki/XAMPP> (consulter le 17 février
2015)