REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département d'Informatique

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : Systèmes d'Informations Avancés

Par: Denidni Meriem

SUJET

Développement D'un Moteur De Recherche Sémantique

Soutenu publiquement le : / /2015 devant le jury composé de :

……	**Université de M'sila**	**Président**
Bouzaroura Ahlem	**Université de M'sila**	**Rapporteur**
……	**Université de M'sila**	**Examinateur**
……	**Université de M'sila**	**Examinateur**

Promotion : 2014 /2015

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

N° d'ordre : .....

UNIVERSITE DE M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

Département d'Informatique

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : Systèmes d'Informations Avancés

Par: Denidni Meriem

SUJET

Développement D'un Moteur De Recherche Sémantique

Soutenu publiquement le : / /2015 devant le jury composé de :

……	**Université de M'sila**	**Président**
Bouzaroura Ahlem	**Université de M'sila**	**Rapporteur**
……	**Université de M'sila**	**Examinateur**
……	**Université de M'sila**	**Examinateur**

Promotion : 2014 /2015

Remerciements

Comment ne pas remercier Allah, qui nous a donné la santé, la volonté et le courage de faire ce travail.

Nous tenons tout d'abord à remercier notre promoteur, Mme Bouzaroura Ahlem, á qui nous doit le sujet, pour sa disponibilité, et pour ses précieux conseils tout au long de ce travail.

Nous tenons également à remercier les membres du jury qui ont bien voulu accepter d'examiner ce travail.

Nous tenons surtout á remercier nos parents pour leurs sacrifices et leur patience, tout au long de nos vies.

Nous remercions toutes les personnes qui ont contribué de près ou de loin dans l'accomplissement de ce travail.

Et à tous ceux que nous n'avons pas cités, nous remercions toutes ma famille et mes amis.

Merci pour tout le monde

TABLE DES MATIÈRES

LISTE DES FIGURES

INTRODUCTION GÉNÉRALE

La Recherche d'Information (RI) est un domaine qui s'intéresse à la structure, à l'analyse, à l'organisation, au stockage, à la recherche et à la découverte de l'information. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur. L'architecture des outils de RI sur le web est généralement caractérisée par l'utilisation d'un index inversé et d'un ensemble de machines fonctionnant en parallèle. La pertinence des réponses est liée à un système de tri de pertinence construit sur la notion de lien existant entre les pages. Avec l'utilisation de ce principe de recherche les moteurs de recherche comme Google, Yahoo et Bing sont nés.

Avant de pouvoir lancer des requêtes sur le serveur d'un moteur de recherche, il est indispensable de remplir sa base de données. En utilisant les liens notés avant, le « crawler » (module de collecte) parcours le web pour collecter les informations afin de construire l'index du moteur de recherche. Pendant la construction d'index, des calculs sont faits pour simplifier la recherche et classer les résultats obtenus quand la requête a été lancée. Lors du lancement de la requête on remarque l'existence des résultats non pertinents à notre demande, ces résultats dus à l'absence de la sémantique entre les mots de la requête.

Le web sémantique est trouvé pour résoudre tels problèmes, les moteurs de recherche intelligents appartiennent au cadre du web sémantique permis entre autres de spécifier le sens des mots. Une recherche pourra alors être située dans un contexte sémantique avant d'être lancée, afin d'apporter des réponses plus précises aux requêtes demandées. Mais on ne peut pas parler du web sémantique sans parler d'ontologie. L'ontologie permet de regrouper les différents concepts du domaine ainsi les relations sémantiques entre eux, ce sont des outils puissants pour la représentation des connaissances.

Alors, notre mémoire s'inscrit dans le domaine du web sémantique dont l'objectif est de créer un moteur de recherche sémantique basé sur une ontologie qui contient le plus vaste nombre de concepts le plus possible avec des relations sémantiques, pour satisfaire les requêtes de recherche et pour obtenir des résultats plus pertinents.

Notre mémoire s'articule autour du squelette suivant :

- Dans le premier chapitre on présente un état de l'art clair et précis sur les concepts de recherche d'information, sur les systèmes de recherche d'information, ceux des outils de recherche sur le web, le moteur de recherche et un exemple de moteur de recherche intelligent.
- Dans le deuxième chapitre on présente des généralités sur le web sémantique, la définition et l'objectif d'une ontologie et la recherche d'information guidée par les ontologies dans le web sémantique.
- Le troisième chapitre est consacré à la conception avec l'explication détaillé de notre contribution.

CHAPITRE 1

LA RECHERCHE D'INFORMATION SUR LE WEB

1. Introduction

Pour trouver l'information désirée sur le Web, une stratégie de base consiste tout bonnement à « Surfer », c'est-à-dire à déambuler de lien en lien, au gré des pages. Souvent agréable et fructueuse, les résultats obtenus demeurent, par ailleurs, largement tributaires des pérégrinations de l'internaute et de l'inclusion subjective par les auteurs de liens dans leurs pages Web. C'est pourquoi il est plus courant, de nos jours, de recourir à ce que l'on appelle des outils de recherche, c'est-à-dire aux divers sites spécialisés dans le repérage de l'information sur Internet. Immensément populaires, ces outils sont également de plus en plus nombreux, et parmi ces outils les moteurs de recherche.

2. La recherche d'information

La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d'établir des représentations des documents dans le but d'en récupérer des informations à travers la construction d'index. L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l'information. On peut aujourd'hui dire que la recherche d'information est un champ transdisciplinaire qui peut être étudié par plusieurs disciplines utilisant des approches qui devraient permettre de trouver des solutions pour améliorer son efficacité.[2]

2.1. Définition

La recherche d'information (RI) concerne les mécanismes qui facilitent l'accès à une base d'informations. C'est une démarche faite par un utilisateur pour obtenir à l'aide du système de recherche d'information (SRI) les informations (ou les références vers les informations) qui peuvent répondre à son besoin.

3. Système de recherche d'information

Selon Alan Smeaton [3] « Le but d'un système de recherche d'information est de retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'usager ».

3.1. Définition

Un système de recherche d'information est défini par un langage de représentation des documents (qui peut s'appliquer à différents corpus de documents) et des requêtes qui expriment un besoin de l'utilisateur (sous forme de mots-clés par exemple), et une fonction de

mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur, c'est-à-dire répondant à son besoin d'information [5]. La figure 1.1, présente un système de recherche d'information.

Les SRI et les modèles sous-jacents se basent donc sur trois notions clés :

- Document : Un document peut être un texte, une page WEB, une image, une bande vidéo,…etc. Dans notre contexte, nous appelons document toute unité qui peut constituer une réponse à une requête d'utilisateur.
- Requête : Une requête exprime le besoin d'information d'un utilisateur.
- Correspondance : Le but de la RI est de trouver seulement les documents pertinents (qui correspondent le mieux à la requête).

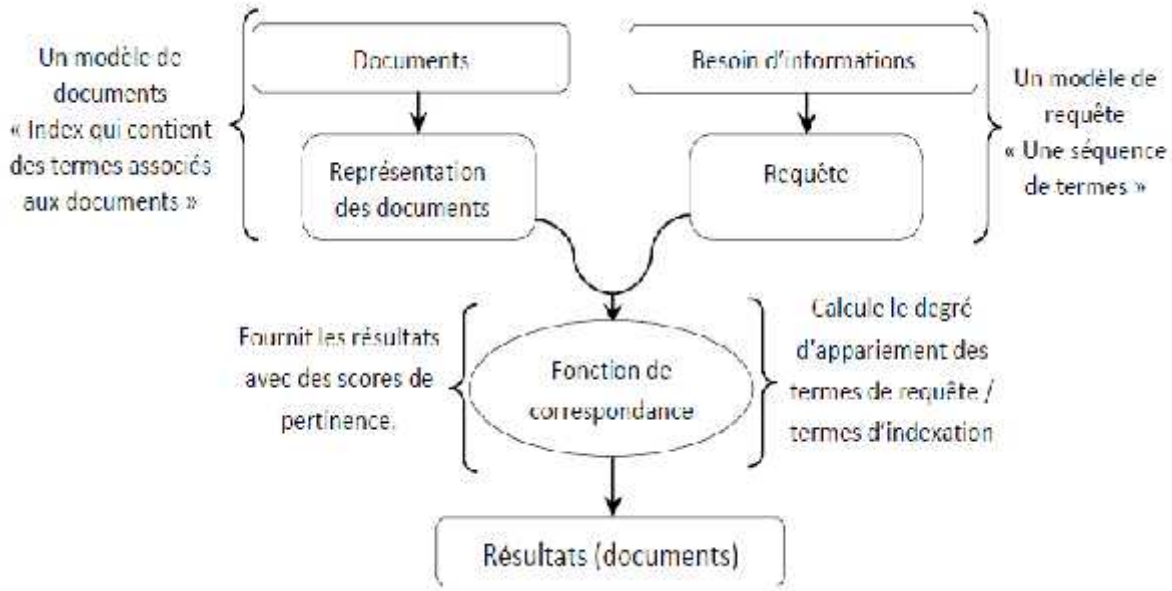

Figure 1.1 Système de Recherche d'Information.[2]

3.2. Processus de recherche d'information

Un système de recherche d'information manipule un corpus de documents qu'il transpose à l'aide d'une fonction d'indexation en un corpus indexé. Ce corpus lui permet de résoudre des requêtes traduites à partir de besoins utilisateur. Un tel système repose sur la définition d'un modèle de recherche d'information qui effectue ces deux transpositions et qui fait

correspondre les documents aux requêtes. La transposition d'un document en un document indexé repose sur un modèle de document. De même, la transformation du besoin utilisateur en requête repose sur un modèle de requête. Enfin, la correspondance entre une requête et des documents s'établit par une relation de pertinence [16]. La figure 1.2, présente les différentes étapes d'un processus de recherche d'information.

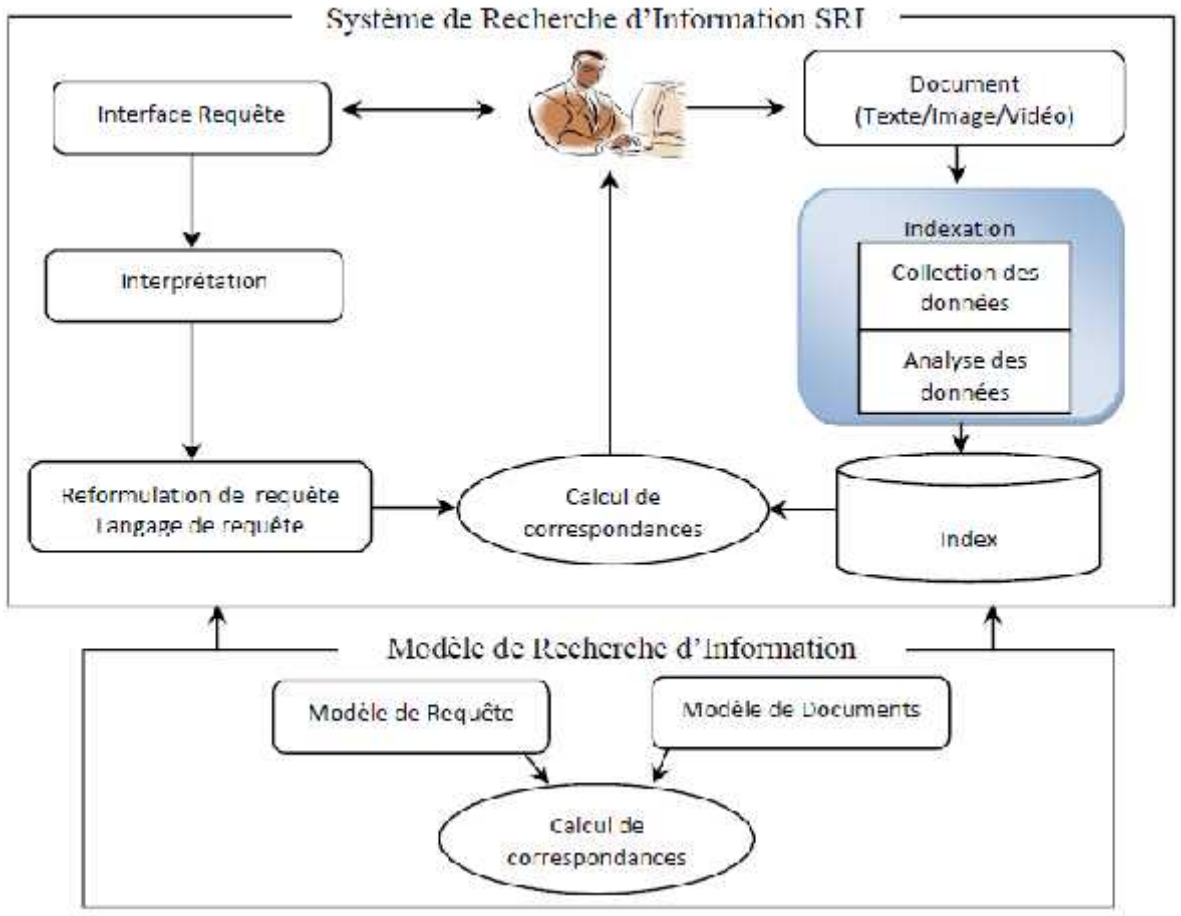

Figure 1.2 Processus de recherche d'information. [17]

3.2.1. Indexation :

L'objectif d'indexation est de trouver les concepts ou représentants les plus importants dans les documents et de créer une représentation interne en utilisant ces concepts, en pratique ces concepts peuvent être des mots simples ou composés (groupe de mots), L'idée d'utiliser des termes simples comme des représentants de concepts est assez naturelle. [13].

L'indexation consiste à identifier l'information contenue dans tout texte et à la représenter au moyen d'un ensemble d'entités appelé index pour faciliter la comparaison entre la représentation d'un document et d'une requête. [6]

Donc l'indexation consiste à analyser les documents afin d'extraire un ensemble des mots clés servant comme descripteurs des documents. Elle peut être :

- Manuelle : la représentation du document se fait par un spécialiste (documentaliste).
- Automatique : la représentation du document est totalement automatisée.
- Semi-automatique : l'extraction des descripteurs s'effectue par le système et le choix des descripteurs est laisse au spécialiste.

L'indexation peut être basée sur un langage contrôle (lexique, ontologie, réseau sémantique, thesaurus) ou libre (les termes sont pris directement à partir des documents). L'indexation à base de langage contrôle permet une recherche par concepts (par sujets, par thème) plus intéressante que la recherche par terme. Le processus d'indexation se compose d'un ensemble de traitements : l'analyse lexicale, l'élimination des mots vides, la lemmatisation, la pondération et enfin la création de l'index [15]

a. Analyse lexicale

C'est la tache consistant à décomposer le contenu de document en mots simples ou composés, afin de trouve l'ensemble des termes appartenant à un document, cette étape se fait dépend fortement de la langue des documents à indexer. Cette extraction est effectuée en tenant compte des espaces, des chiffres et des ponctuations. Un terme peut être un mot simple ou composé, mais en RI on utilise souvent les mots simples.[9]

b. Elimination des mots vides

Les mots vides sont des mots trop fréquents peu significatifs et porteurs de peu de sens, augmentant ainsi la taille de l'index et rendant la recherche plus lente. L'élimination de ces mots permet de réduire l'index, on gagne alors en espace mémoire. [9]

On distingue deux techniques pour éliminer les mots vides :

- L'utilisation d'une liste prédéfinie de mots vides (aussi appelée anti-dictionnaire ou stop list), par exemple cette liste pourra contenir les termes (the, or, a, you, I, us, of, in...) pour l'anglais, (le, la, de, des, je, tu...) pour le français.
- L'élimination des mots dépassant un certain nombre d'occurrences dans le document.

c. Lemmatisation

Un mot peut avoir plusieurs formes dans un texte dont le sens est presque similaire. La lemmatisation est une technique qui permet de ramener un mot à sa racine. Par exemple,

programmes et programme, programmer et programmation, programmeurs et programmées font tous références à la racine 'programme'. Elle désigne l'analyse lexicale du contenu textuel regroupant les mots d'une même famille afin de réduire les mots à leurs racines grammaticales.[9]

L'algorithme de Porter est sans doute le plus connu dans ce domaine, elle est un procède pour éliminer les terminaisons plus communes morphologiques et flexionnelles des mots en anglais. Son utilisation principale est dans le cadre d'un processus de normalisation terme qui se fait habituellement lors de la mise en place des systèmes de recherche d'information.[18]

d. Pondération

Dans un document, certains termes sont plus représentatifs du contenu et de la sémantique du document par rapport l'autres. L'objectif de la pondération est de trouver les termes qui représentent le mieux le contenu d'un document. La pondération des termes permet de mesurer l'importance d'un terme dans un document, cette importance est souvent calculée à partir de considérations et d'interprétations statistiques (ou parfois linguistiques). Les termes importants doivent avoir un poids fort.[9]

e. Création de l’index

Les informations sélectionnées lors du processus d'indexation sont mémorisées et enregistrées dans une structure appelée index. Cette structure permet de sélectionner pour n'importe quel terme, tous les documents le contenant. Il y a deux solutions plus utilisée actuellement fichiers inverses et fichiers maitres.

- Les fichiers inverses : sont composes de deux éléments principaux : le dictionnaire et le fichier posting. Le dictionnaire consiste en une liste de tous les mots distincts de la collection. Pour chaque mot est assigne l'ensemble des documents dans lesquels ce dernier apparait (posting).
- Les fichiers maitres : Au lieu de donner pour chaque document les mots et les fréquences qui le constituent, on donne pour chaque mot les documents ou il apparait et sa fréquence dans chacun de ces documents.

Bien que l’indexation se base sur des techniques relativement établies, il peut y avoir plusieurs indexations différentes d’un même texte, aussi valables les unes que les autres, en fonction de l’usage qui doit en être fait et du public auquel elles s’adressent.[2]

L’indexation se décompose en trois phases schématisées dans la figure 1.3.

- L'extraction des termes du document.
- La sélection des termes discriminatifs pour un document.
- La pondération des termes.

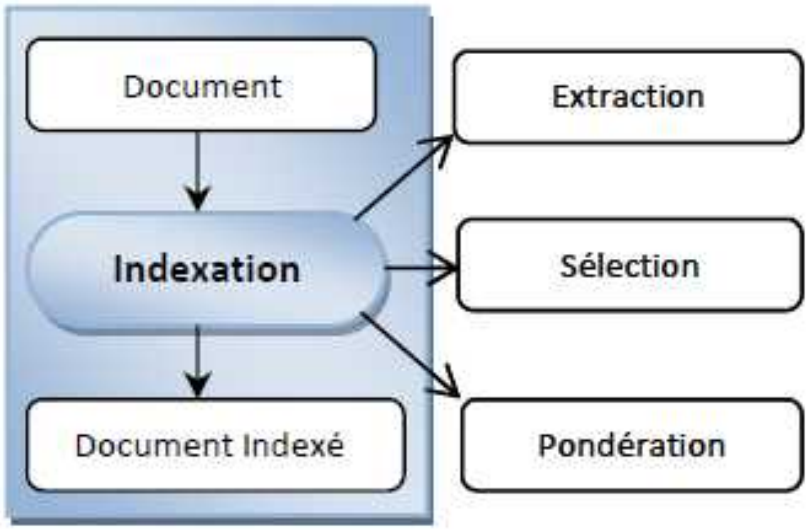

Figure 1.3 Indexation d'un document. [5]

3.2.2. Interrogation

Il s'agit de l'expression du besoin d'information de l'utilisateur dans la forme imposée par le système, la recherche dans le corpus, et la présentation des résultats. Cette phase nécessite un modèle de représentation du besoin de l'utilisateur, appelé modèle de requêtes, ainsi qu'une fonction de correspondance qui doit évaluer la pertinence des documents par rapport à la requête. La réponse du système est un ensemble de références à des documents qui obtiennent une valeur de correspondance élevée. Cet ensemble est généralement présenté sous la forme d'une liste ordonnée suivant la valeur de correspondance. [17]

3.2.3. Fonction de correspondance

Tout système de recherche d'information s'appuie sur un modèle de recherche d'information. Ce modèle se base sur une fonction de correspondance qui met en relation les termes d'un document avec ceux d'une requête en établissant une relation d'égalité entre ces termes. Cette relation d'égalité représente la base de la fonction de correspondance et, par la même, du système de recherche d'information.[5]

Il existe un certain nombre de modèles théoriques dans la littérature les plus connus étant le « Modèle Booléen», le «Modèle Vectoriel», et le «Modèle Probabiliste ». Dans le modèle booléen, les requêtes sont représentés sous forme de termes reliés par des opérateurs booléens (ET, OU, NON, . . .). Le modèle vectoriel considère les documents et les requêtes comme des vecteurs pondérés, chaque élément du vecteur représentant le poids d'un terme dans la requête ou le document. Le modèle probabiliste tente d'estimer la probabilité qu'un document donné soit pertinent pour une requête donnée [23].

4. Recherche d'information sur le WEB

On va résumer premièrement l'historique de la recherche sur Internet.

4.1. Historique de la recherche sur Internet

Le premier moteur de recherche apparait en 1990, crée par Adam Emtage, étudiant à Mc Gill (Québec). Ce moteur, dénommé Archie, comportait les principes de base du moteur de recherche : on remplissait une base de données, que le moteur faisait correspondre aux requêtes des utilisateurs. Le Web de l'époque comportait seulement quelques centaines de sites, et Archie resta un projet universitaire.

Mais le saut technologique le plus important fut introduit par Wanderer (« le Vagabond ») en 1993 par Matthew Gray. Il fut le premier moteur à déployer des robots d'indexation (spiders). L'idée de base, qui était de mesurer la croissance du Web, fut rapidement remaniée pour arriver au premier moteur de recherche à indexation automatique (Bot search) Ce moteur a d'ailleurs causé un certain nombre de problèmes, car il retournait plusieurs centaines de fois par jour sur certains sites et les ralentissait.

En octobre 2003, le successeur d'Archie fait son apparition : Aliweb (Archie-like indexing the web). Ce moteur repose sur la soumission manuelle de sites. Le moteur se basait sur les mots clés et les descriptions fournies au moment de l'inscription pour effectuer la recherche.

Le premier moteur intelligent fut Excite (1993). Construit par six étudiants de Stanford, il se base sur l'analyse statistique des mots.

Enfin, en 1994, c'est la naissance de Yahoo, le premier « grand » service de recherche, crée également par des étudiants de Stanford. Mais à la différence des outils de l'époque, Yahoo se base sur un annuaire, pas sur un moteur de recherche. Les résultats sont sélectionnés et indexés par l'homme. En quelques mois, Yahoo devient le plus important portail du Web.

Les années 1995-1997 voient l'apparition des grands moteurs de recherche (Excite, Hotbot, Lycos…). Altavista, crée par un français et jugé efficace et rapide, deviendra la star des moteurs de recherche du moment jusqu'aux années 2000, détrôné par Google.

De son coté, Inktomi développe la première activité de recherche destinée aux entreprises. C'est la première fois que les moteurs de recherche ciblent les professionnels.

Enfin, c'est en 1998 que naît Google, crée par Sergei Brin et Larry Page, encore une fois étudiants de Stanford. Google va littéralement révolutionner le monde de moteurs de recherche grâce à sa simplicité et son efficacité. L'interface dépouillée se charge instantanément sur les connexions bas-débit de l'époque, et la technologie d'indexation est inédite : Google se base sur le nombre de liens pointant sur une page pour en déterminer sa pertinence.

Vers 2001-2002, l'éclatement de la bulle internet fait disparaitre les premiers moteurs de recherche, et seuls les plus grands survivent. C'est l'ère moderne de la recherche internet. [20]

4.2. Outils de recherche sur le WEB

Il existe de nombreux outils de recherche d'information sur le Web, ces outils qui se spécialisent en fonction des services utilisés et du type d'information qu'ils recensent. Il existe à l'heure actuelle trois grandes familles d'outils de recherche :

- Les annuaires ou répertoires thématiques.
- Les moteurs de recherche.
- Les méta-moteurs.

4.2.1. Les annuaires

L'annuaire (ou directory) est en fait une liste de liens subdivisés en catégories suivant une structure en arbre, accompagnée d'une brève description. Bien que ce procédé fût pionnier en la matière, il tend à disparaître. En effet, le fait de devoir sélectionner les catégories dans lequel on recherche suppose que l'on sache exactement où chercher. Et on peut se demander où se positionne le site qui appartient à plusieurs catégories. Mais à cette question, les moteurs utilisant ce procédé vous répliqueront qu'ils se trouvent dans toutes celles susceptibles de correspondre. Néanmoins, on doit lui reconnaître un gros avantage, celui de mettre en quelque sorte dans le contexte, ainsi les recherches dans la base de données sont diminuées, en plus d'obtenir des résultats plus pertinents.[20]

Quelques annuaires : Yahoo, Voilà, …

Les annuaires sont donc des outils basés sur le recensement humain de l'information. Ils signalent des sites et des ressources de l'Internet comme un catalogue de bibliothèque signale

des livres ou bien encore comme les pages jaunes signalent des entreprises. On distingue dans ce contexte deux catégories d'annuaires On distingue dans ce contexte deux catégories d'annuaires :

a. Les annuaires commerciaux

Ils se financent grâce à la publicité. Ils ont en principe une couverture dit "générale" (ils couvrent toutes les disciplines). Ils peuvent concerner le monde ou une zone régionale, nous citons parmi eux :

- Annuaires généralistes internationaux : le plus connu est sans doute 'Yahoo Directory', mais il existe aussi 'DMIZ' de l'Open Directory Project et l'annuaire de 'Lycos'.
- Annuaires régionaux commerciaux : ce sont les annuaires qui recensent des sites en fonction de leur langue. Dans le cas de des annuaires francophones nous citons la version française de 'Yahoo Directory' ou encore l'annuaire 'Francité'.
- Les annuaires qui recensent d'autre pays ou parties du monde: comme l'annuaire 'Wohaa' pour l'Afrique et l'annuaire russe 'Yandex'.

b. Les annuaires non commerciaux

Sont des annuaires élaborés par des individus de façon bénévole ou bien par des institutions. Ils sont soit généraux soit spécialisés. Leur préoccupation consiste toujours à identifier les ressources et les sites en tenant comptes de leur qualité :

- Annuaires à couverture (généraliste): comme le 'Vlib' (Virtual Library) et l'annuaire 'Resource Discovery network'.
- Annuaires à couverture thématique ou spécialisée : comme le répertoire en sciences humaines 'Voice of the Shuttle' et le répertoire de ressources juridiques 'Findlaw'. [2]

4.2.2. Les moteurs de recherche

Un moteur de recherche (Search Engine) est un outil , un logiciel qui parcoure le web et indexe automatiquement le contenu qu'il visite. Il permet d'accéder à différentes ressources comme des pages web, des images, de la musique, des vidéos,… Lorsque l'internaute effectue une recherche, le moteur lui retourne une liste de résultats classée selon leur pertinence avec cette requête, interrogation. Quelque moteurs de recherche : Google, MSN Search, Lycos, Altavista, Excite….

4.2.3. Les méta-moteurs

Certains moteurs ont opté pour une solution plus économique, puisqu'ils utilisent les bases de données des autres moteurs. Ainsi les métamoteurs rassemblent plusieurs moteurs de

recherche. L'un des avantages évidents de ce procédé pourrait être d'obtenir des résultats plus pertinents, puisque la recherche s'étend sur un plus grand nombre de sites indexés, sites figurant sur tel moteur, mais pas sur un autre. Néanmoins, la redondance de sites affichés peut-être un inconvénient gênant. De même que l'augmentation considérable de résultat qui peut engendrer un délai d'attente supérieur. De plus, le fait d'envoyer différentes requêtes à différents serveurs rallonge également le temps de réponse. [20]
Quelques métamoteurs : Infospace, Askjeeves, MyWay, Websearch.com…

5. Les moteurs de recherches

Un Moteur de recherche d'informations sur Internet est un outil permettant à un utilisateur, aussi bien novice que très expérimenté, d'accéder de manière simplifiée à des données dont la localisation lui est inconnue, ou dont les différentes parties sont disséminées sur le web.

Ceci est réalisable par la simple opération qui consiste à donner au Moteur les mots importants concernant le sujet des informations recherchées, dits mots-clés, ou la description même du document recherché, et ce Moteur a pour effet de retourner comme résultat de ses recherches la liste de toutes les pages web relatives à ces mots-clés, ou bien celles correspondant à la description passée en paramètre. [21]

De façon scientifique. Un moteur de recherche est une immense base de données continuellement mise à jour par des "Robots" ou "Spiders" qui scrutent le Web en allant de page en page et qui les sauvegardent dans son index. Son fonctionnement est donc totalement automatisé.

Le rafraîchissement de la base se fait selon certains délais. Une fois les pages repérées, le moteur de recherche classe les pages par ordre de pertinence, selon un ordre et un algorithme basé sur des critères de tri qui lui sont spécifiques. C'est à la pertinence des résultats obtenus que l'on juge la qualité d'un moteur de recherche. [9]

5.1. Principe de base

Les moteurs de recherche sont constitué de « robots », encore appelés bots, spiders, crawlers ou agents qui parcourent les sites à intervalles réguliers et de façon automatique (sans intervention humaine, ce qui les distingue des annuaires) pour découvrir de nouvelles adresses (URL). Ils suivent les liens hypertextes (qui relient les pages les unes aux autres) rencontrés sur chaque page atteinte. Chaque page identifiée est alors indexée dans une base de données, accessible ensuite par les internautes à partir de mots-clés.

Les moteurs de recherche ne s'appliquent pas qu'à Internet : certains moteurs sont des logiciels installés sur un ordinateur personnel. Ce sont des moteurs dits desktop qui combinent

la recherche parmi les fichiers stockés sur le PC et la recherche parmi les sites Web — on peut citer par exemple Exalead Desktop, Google Desktop et Copernic Desktop Search, etc.

Des modules complémentaires sont souvent utilisés en association avec les trois briques de bases du moteur de recherche. Les plus connus sont les suivants :

a. Le correcteur orthographique : il permet de corriger les erreurs introduites dans les mots de la requête, et s'assurer que la pertinence d'un mot sera bien prise en compte sous sa forme canonique.

b. Le lemmatiseur : il permet de réduire les mots recherchés à leur lemme et ainsi d'étendre leur portée de recherche.

c. L'anti dictionnaire : utilisé pour supprimer à la fois dans l'index et dans les requêtes tous les mots "vides" (tels que "de", "le", "la") qui sont non discriminants et perturbent le score de recherche en introduisant du bruit.

En ce qui concerne les caractéristiques, les moteurs de recherche ont un fonctionnement commun, mais diffèrent par un certain nombre de critères. Pour ce qui est de commun, rappelons simplement qu'ils procèdent tous des même étapes :

- D'abord l'exploration du web, durant laquelle ils vont collecter les informations sur chaque page rencontrée.
- Puis l'indexation, durant laquelle ils vont enregistrer dans une base de données les Informations collectées.
- Enfin la recherche, durant laquelle ils vont rechercher les données collectées en fonction des mots clés. [2]

5.2. Structure d'un moteur de recherche

Afin d'assurer le rôle d'un moteur de recherche et couvrir tous les aspects liés à la recherche documentaire, il est principalement constitué de trois parties :

- Le robot (appelé aussi spider)
- L'index ou la base de données.
- Le logiciel/interface d'interrogation.

5.2.1. Le robot

Le robot ou encore "spider" est la partie la plus importante du moteur de recherche, car celui-ci effectue la recherche directement sur le Web pour en extraire le plus grand nombre

d'informations relatives aux documents présents sur le Web et les indexer au sein de sa base de données.

Un moteur de recherche utilise un robot qui balaie la structure hypertexte du Web par suivi récursif des liens pour en archiver intégralement son contenu. Il assure ainsi la lecture des données des pages Web et le repérage des liens pointant vers d'autres pages afin de constituer l'index.

5.2.2. L’index

L'index est le lieu de stockage et d'indexation des pages Web visitées par le robot, c'est la phase de structuration et de classification de l'information rapatriée. En général, les informations répertoriées sont : l'adresse (URL), le titre des pages, les mots clés voir même l'intégralité des pages. L'entrée de pages Web dans l'index est également rendue possible par la visite du robot pages soumises volontairement par leurs créateurs. Afin d'actualiser le contenu des pages Web dans l'index, le robot se rend à intervalle défini sur celles-ci, pour mettre à jour leur contenu dans l'index.

5.2.3. L’interface

L'interface permet à l'utilisateur de saisir sa requête en utilisant un ou plusieurs termes significatifs qui seront ensuite recherchés dans l'index. L'interface sélectionne parmi les milliers de documents enregistrés dans l'index ceux qui satisfont cette requête et les propose sous forme d'une liste de pages Web énumérées selon un ordre de pertinence décroissant, pages sur lesquelles l’utilisateur peut ensuite se rendre via un lien hypertexte. [9]

5.3. Fonctionnement d’un moteur de recherche

Le mode de fonctionnement d’un moteur de recherche peut-être schématisé de la manière suivante :

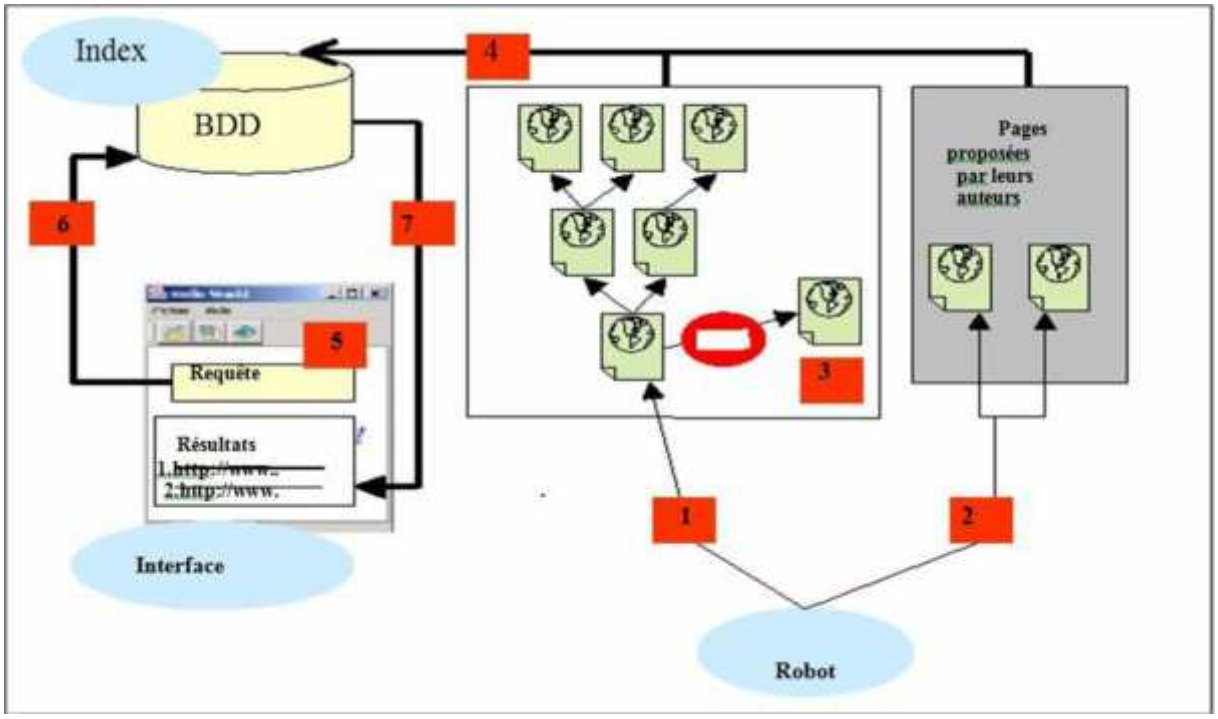

Figure 1.4 Fonctionnement d'un moteur de recherche.[9]

Le robot repère les pages Web par suivi récursif des liens appartenant aux pages présentes sur le réseau **(1)** ou proposées par les auteurs **(2).** Toutefois, les pages à accès réservé ne seront pas indexées (**3**) alors que les autres seront indexés au sein de la base de données (**4**). La requête est émise par l'utilisateur à travers l'interface de recherche (**5**) dont les termes seront recherchés dans l'index (la base de données) (**6**). Les résultats sont ensuite affichés à travers l'interface sous forme de liens hypertextes (**7**).

5.4. Architectures des moteurs de recherche

5.4.1. Architecture générale des premiers moteurs de recherche

L'architecture originale utilisée par Altavista représente la première catégorie de systèmes. Il s'agit d'une architecture très simple qui se divise en deux parties distinctes. On retrouve d'une part un crawler et d'autre part l'interface d'interrogation du moteur de recherche et le système d'analyse des requêtes proposés par les utilisateurs du système.

Le cœur du système repose sur un index inversé permettant d'associer des mots à un ou plusieurs documents. La demande de l'utilisateur est traitée en interrogeant l'index inversé pour connaître les documents dans lesquels apparaissent le plus souvent les mots de la requête. [7]

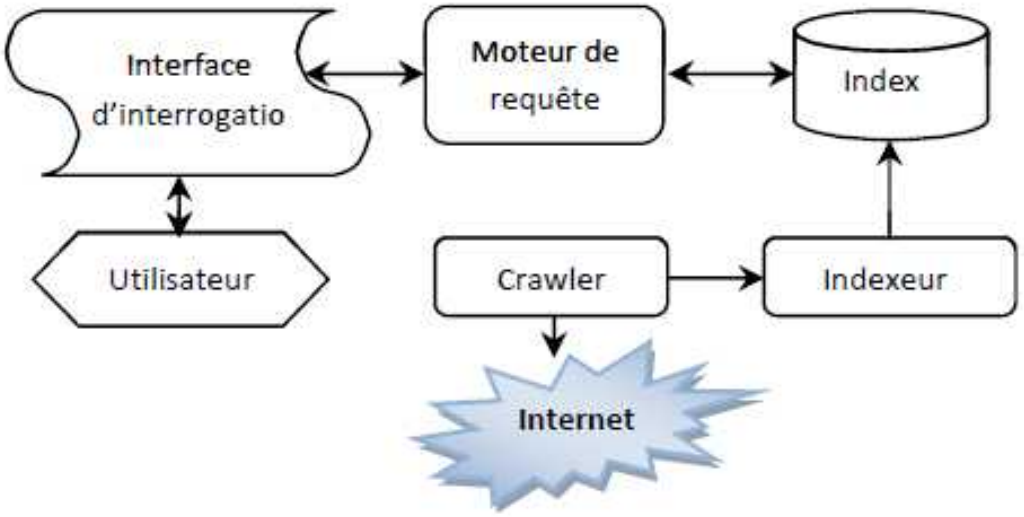

Figure 1.5 Architecture originale du moteur de recherche Altavista [7].

5.4.2. Architecture distribuée et adaptative

Des variantes de l'architecture précédente, basées sur le modèle indexeur-crawler, ont été imaginées pour gommer les défauts inhérents à sa conception. L'une d'entre elle, appelée Harvest s'est révélée très innovante en matière de distribution des ressources.

- Le récolteur : est chargé de collecter et d'extraire périodiquement des informations d'indexation -textes, images - depuis plusieurs sites Web.
- Le broker : quant à lui, fournit le mécanisme d'indexation et l'interface d'interrogation sur les données amassées par le récolteur.

On retrouve ici, le mécanisme indexeur-crawler identifié dans la section précédente. Cependant, plusieurs brokers et plusieurs récolteurs peuvent communiquer ensemble, chacun se spécialisant dans un domaine précis. Lorsqu'une requête est émise sur un broker dont le domaine traité ne correspond pas à ses capacités, celui-ci transmet la requête à une autre entité capable de la gérer.

C'est un système totalement adaptatif dans lequel il est possible de configurer les brokers et les récolteurs de manière à répartir le besoin en ressources sur un ou plusieurs domaines particuliers. [7].

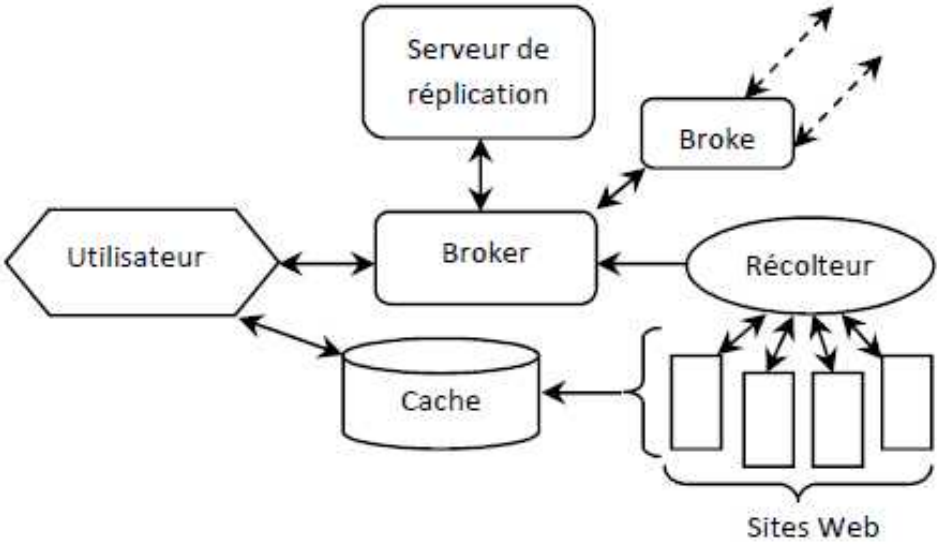

Figure 1.6 Architecture du système Harvest. [7]

5.4.3. Architecture moderne d'un moteur de recherche

L'architecture du moteur de recherche Google est certainement une des plus efficaces actuellement. Elle ne repose pas sur un système monolithique mais sur un grand nombre de machines classiques coopérant ensemble. Ce système peut se décomposer en plusieurs parties comprenant :

- Un sous-système d'exploration d'Internet
- Un indexeur
- Un analyseur de la topologie d'Internet formée par les liens hypertextes : et un sous-système de présentation et d'exécution de requêtes.
- Un serveur d'URL garde la mémoire des liens des pages à visiter. Des robots chargés d'explorer le Web récupèrent ces liens afin de télécharger les documents orrespondant et les stocker dans une base de données recensant la totalité des pages indexées. Cette opération est réalisée continuellement et alimente et met à jour en permanence la base de documents du moteur. Périodiquement, cette base est analysée pour réaliser un index inversé reliant des termes aux documents les contenant. D'autres informations sur les termes sont extraites comme leur position dans le document, la taille de la police utilisée ou sa fonte.

Cette analyse permet également d'extraire tous les liens hypertextes des documents rencontrés afin d'alimenter le serveur d'URL. Cette base de liens est utilisée afin de calculer le PageRank permettant de trier les documents de l'index par pertinence décroissante.[picarg.]

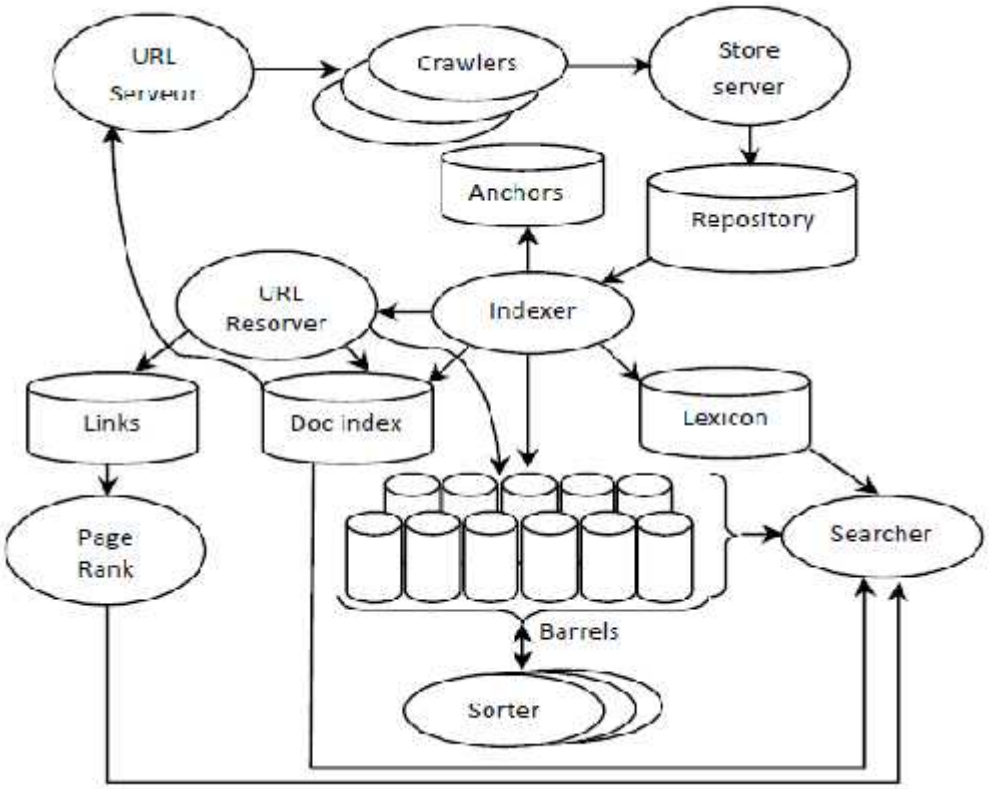

Figure 1.7 Architecture du moteur de recherche Google. [7]

5.5. Algorithme des moteurs de recherche

Différentes études ont suggéré de tenir compte de la popularité des documents afin d'améliorer les performances de la recherche d'information. Le PageRank [29] de Google et le HITS [12] de Kleinberg sont deux algorithmes fondamentaux qui utilisent les liens hypertextes pour classer les résultats d'une requête. Généralement, ces algorithmes fonctionnent en deux temps : Dans une première étape, un moteur de recherche retourne une liste de documents répondant à la requête posée, en fonction des termes de la requête et des termes d'indexation des documents. Dans une seconde étape, ces systèmes tiennent compte des liens hypertextes pour classer ces documents [11].

5.5.1. Hyperlink-Induced Topic Search (HITS)

Kleinberg fut un des premiers à s'intéresser aux propriétés de connectivité du graphe représentatif d'Internet et de son apport dans la détection de la pertinence d'une page à une requête [11]. Quelques constatations simples sont à l'origine de ses travaux dans ce domaine.

On retrouve d'une part les pages qui semblent être très importantes et jouent le rôle d'autorité sur un sujet donné et d'autre part les documents possédant un grand nombre de liens vers des pages faisant autorité sur un sujet. On distingue ainsi les pages *autorités* ayant un grand nombre de liens entrants et les pages *hubs* ayant un grand nombre de liens sortants et regroupant les autorités d'un même sujet. Le but de l'algorithme HITS est de déterminer les hubs et les autorités qui renforcent leurs relations mutuellement sur un sujet donné. Ainsi Kleinberg dénombre les bons hubs comme des pages pointant vers beaucoup de bonnes autorités et les bonnes autorités comme des pages pointées par beaucoup de bons hubs [7].

5.5.2. PageRank

Quelques moteurs de recherche, dont le plus connu est Google, ont pris le pari d'utiliser un autre mode de classement des résultats. Les pages Web sont ordonnées selon leur popularité, une page qui est la cible d'un très grand nombre de liens est probablement non seulement une page validée (page parcourue par un grand nombre de lecteurs, qui ont jugé bon de la citer en référence) mais aussi une page détenant un contenu utile à un grand nombre d'utilisateurs.

L'approche du PageRank qui a fait la spécificité du moteur de recherche Google, repose sur la notion de propagation de popularité. Le principe consiste à évaluer l'importance d'une page en fonction de chacune des pages pointant vers elle. La propagation met en avant les pages qui jouent un rôle particulier dans le graphe des liens, avec l'hypothèse suivante : *"une page est importante quand elle est beaucoup citée ou citée par une page très importante*".

La mesure de PageRank (PR) proposée par [29] est une distribution de probabilité sur les pages. Elle mesure en effet la probabilité PR, pour un utilisateur navigant au hasard, d'atteindre une page donnée. Elle repose sur un concept très simple : un lien émis par une page A vers une page B est assimilé à un vote de A pour B. Plus une page reçoit de votes, plus cette page est considérée comme importante. Le PageRank se calcule de la façon suivante :

Soient $T1$, $T2$, ..., Tn : n pages citant une page A. Notons $PR(Tk)$ le PageRank de la page Tk, $S(Tk)$ le nombre de liens sortants de la page Tk, et d'un facteur compris entre 0 et 1, fixé en général à 0.85. Ce facteur d représente la probabilité de suivre effectivement les liens pour atteindre la page A, tandis que (1-d) représente la probabilité d'atteindre la page A sans suivre de liens. Le PageRank de la page A se calcule à partir du PageRank de toutes les pages Tk de la manière suivante :

$$PR(A) = (1-d) + d\frac{PR(T)}{S(T)} \qquad (1)$$

Initialement, toutes les pages sont équiprobables, leur valeur de PR est alors égale à 1/*n*, n étant le nombre de documents de la collection [11].

5.5.3. Ponderation TF.Idf

Le TF-IDF (Term Frequency-Inverse Document Frequency) est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur.[35]

a. Fréquence du terme TF

La fréquence d'un terme (term frequency) est simplement le nombre d'occurrences de ce terme dans le document considéré, normalisée par la somme des nombres d'occurrences de tous les termes du document.

Le nombre d'occurrence peut rendre compte de « l'importance » d'un terme dans un document. La normalisation du nombre d'occurrences d'un terme rend possible la comparaison de deux documents de longueurs différentes.

Soit le document $\mathbf{d_j}$ et le terme $\boldsymbol{t_i}$, alors la fréquence du terme dans le document est :

$$\mathrm{tfi,j} = \frac{\mathrm{ni,j}}{\sum_k \mathrm{nkj}} \qquad (2)$$

Où $\mathbf{n_{i,j}}$ est le nombre d'occurrences du terme $\mathbf{t_i}$ dans $\mathbf{d_j}$ .Le dénominateur est le nombre d'occurrences de tous les termes dans le document $\mathbf{d_j}$.

b. Fréquence inverse de document IDF

La fréquence inverse de document(inverse document frequency) est une mesure de l'importance du terme dans l'ensemble du corpus. Dans le schéma TF-IDF, elle vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme :

$$idfi = log\frac{|D|}{|\{dj:ti \epsilon dj\}|} \qquad (3)$$

Où : |D| : nombre total de documents dans le corpus

$|\{d_j : t_i \epsilon d_j\}|$: nombre de documents où le terme t_i apparaît (C'est-à-dire $n_{i,j} \neq 0$).

c. Calcul de TF-IDF

Finalement, le poids s'obtient en multipliant les deux mesures :

$$tfidfi,j = tfi,j * idfi,j \qquad (4)$$

6. Moteurs de recherche intelligents

Pour dire qu'un moteur de recherche est intelligent, il s'agit de comprendre les questions de l'utilisateur et lui donne des réponse spécifiques, sur mesure, comme le ferait un être humain. Mais il apparaît très vite que cette notion est assez difficile à définir. En effet, tous les jours, les gens utilisent le contexte dans leurs prises de décision. Prenant l'exemple de la phrase suivante ''S'il pleut, je vais prendre un parapluie pour me rendre à l'université''. Le fait de prendre un parapluie n'a rien à voir avec le fait d'aller à l'université mais pourtant cela contraint la manière d'exécuter la tâche d'aller à l'université. La mise en évidence du contexte est particulièrement visible quand il y a plusieurs méthodes pour accomplir une tâche. Dans ce cas, chaque personne choisit sa méthode en fonction de ses connaissances et des informations contextuelles qu'il possède. Il s'avère donc que le traitement des données contextuelles joue un rôle dans tous les domaines où le raisonnement intervient. [2]

Alors il est très important pour un moteur de recherche intelligent d'être capable de traiter ces langages naturels. Et ça se base sur la notion de Web sémantique. Aussi, les modules complémentaire (ensembles des agents intelligents intégrée) est un moyen d'ajouter l'intelligence au moteur de recherche.

Un moteur de recherche intelligents généralement contient les caractéristiques suivantes :

- ✓ Suggérer des mots proches : Certains moteurs de recherche proposent à l'utilisateur des termes proches de ceux de sa requête, ce qui permet de la préciser. Ces termes proches sont identifiés par des méthodes statistiques ou bien en se basant sur des dictionnaires ou ontologies.
- ✓ Les connaissances plutôt que les informations
  - Tentatives d'appréhender non plus les informations brutes (présence de tel mot dans telle page d'un site Internet) mais des informations qualifiées ou connaissances.
  - Nécessite des pages de sites où les informations sont qualifiées.
  - Recherches sur des informations qualifiées.

- ✓ Exploitation du sens sémantique des contenus.
- ✓ Standardisé et formalisé par Ontologie (OWL), Description des ressources, Métadonnées, annotation, URI, triplets
- ✓ machines exploitent, interprètent et combinent les ressources.
- ✓ Vaste espace d'échange de ressources entre machines et utilisateurs.[9]

6.1. Exemple d'un moteur de recherche intelligent :

6.1.1. WolframAlpha

Wolfram Alpha est un outil de calcul en langage naturel développé par la société internationale Wolfram Research. Il s'agit d'un service internet qui répond directement à la saisie de questions factuelles en anglais par le calcul de la réponse à partir d'une base de données, au lieu de procurer une liste de documents ou de pages web pouvant contenir la réponse. Son lancement a été annoncé en mars 2009 par le physicien et mathématicien britannique Stephen Wolfram et il a été lancé le 16 mai 2009 à 3 h 00 du matin. Wolfram Alpha contient environ 10 milliards d'informations, plus de 50 000 types d'algorithmes et de modèles, et des capacités linguistiques pour plus de 1 000 domaines**.** [22]

Les utilisateurs saisissent une question ou une demande de calcul. Le service calcule les réponses et les visualisations correspondantes à partir d'une base de connaissance. Grâce à l'utilisation de l'outil Mathématica, Wolfram|Alpha est capable de répondre à des questions mathématiques. La réponse est généralement présentée sous une forme lisible par un être humain.

Exemple : $\lim(x \to 0)\ x/\sin(x)$ fournit la réponse attendue, 1, ainsi qu'une façon de l'obtenir en utilisant la règle de L'Hôpital.

Wolfram|Alpha est aussi capable de répondre à des questions factuelles posées en anglais naturel, telles que « Where was Ségolène Royal born? »(« Où Ségolène Royal est-elle née ?»), ou des questions plus complexes telles que « How old was Nicolas Sarkozy in 1981? »
(« Quel âge avait Nicolas Sarkozy en 1981 ? »). Wolfram|Alpha affiche son interprétation de la question saisie (« Input interprétation ») à l'aide de phrases standardisées, par exemple « Ségolène Royal | place of birth » ou « age | of Nicolas Sarkozy (politician) | »

Wolfram|Alpha analyse des données issues de disciplines très variées, dont notamment les mathématiques, statistiques, analyse de données, physique, chimie, science des matériaux, ingénierie, astronomie, sciences de la vie et de la Terre (géologie), nouvelles technologies, dates et heures, lieux et géographie, données socioéconomique, météorologie, santé et

médecine, alimentation et nutrition, linguistique, culture, médias, personnalités, histoire, éducation, organisations diverses, jeux et sports, musique, couleurs, …etc.

Figure 1.8 Exemple d'un question dans Wolfram Alpha

7. Conclusion

Dans ce chapitre nous avons présenté les principales notions et concepts de la recherche d’information, des systèmes de recherche d’information et ceux des outils de recherche sur le Web.

A 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, il s’agit donc de retourner l’ensemble de documents pertinents. Cependant, nous constatons que la notion de pertinence dépend de la satisfaction de l’utilisateur d’une part, et des différents sens portés par les termes de la requête d’une autre part. Cette constatation constitue le point faible de la recherche d’information classique, elle représente également le point de départ pour de nouveaux paradigmes de recherche.

Dans le chapitre suivant, on va voire la recherche sémantique d’information, qui utilise des ressources externes, généralement les ontologies, pour avoir des résultats plus performants prenant en compte la sémantique.

CHAPITRE 2

LE WEB SÉMANTIQUE ET LES ONTOLOGIES

1. Introduction

Nées des besoins de représentation des connaissances, les ontologies sont à l'heure actuelle au cœur des travaux menés en Ingénierie des Connaissances (IC). Le terme « ontologie » est utilisé depuis le début des années 1990, et son champ d'application s'élargit considérablement. Un des plus grands projets basés sur l'utilisation d'ontologies consiste à ajouter au Web une véritable couche de connaissances permettant des recherches d'informations au niveau sémantique. A terme, il est prévu que des applications internet pourront mener des raisonnements on utilisant les connaissances stockées sur la Toile.

Au fur et à mesure des expérimentations, des méthodologies de construction d'ontologies et des outils de développement adéquats sont apparus. L'enjeu de l'effort engagé est de rendre les machines suffisamment sophistiquées pour qu'elles puissent intégrer le sens des informations.

2. Web Sémantique

2.1. Définition

Le Web sémantique désigne un ensemble de technologies visant à rendre le contenu des ressources du World Wide Web accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles, utilisant notamment la famille de langages développés par le W3C(World Wide Web Consortium).[34]

Concrètement, le web sémantique est une infrastructure qui permet l'utilisation de connaissances formalisées en plus du contenu informel que l'on peut trouver dans le web. Cette infrastructure s'appuie sur un certain niveau de consensus portant, par exemple, sur les langages de représentation ou sur les ontologies utilisées. Ainsi, elle permet, le plus automatiquement possible, l'interopérabilité et les transformations entre les différents formalismes et les différentes ontologies.

Grâce à la formalisation de connaissances, cette infrastructure peut faciliter la mise en œuvre de calculs et de raisonnements complexes tout en offrant des garanties supérieures sur leur validité. Mais restreindre le web sémantique à cette infrastructure serait trop limitatif. Sur la base de sémantiques bien définies pour ses ressources, le web sémantique pourra fournir aux utilisateurs, par le moyen d'agents logiciels, des services automatiques et avancés.[32]

2.2. Principales composantes du web sémantique

Le Web Sémantique a été proposé en se basant sur les critiques adressées au web. Ces critiques s'articulent autour des éléments suivants :

- Certes HTML a permis de tisser tout un réseau d'informations par ses liens hypertextes, mais il n'a donné aucune sémantique à ces liens ce qui les rend pratiquement inexploitables par les machines.
- Les métadonnées utilisées sont non structurées et limitées dans leurs usages.
- Il est difficile de faire des inférences et des raisonnements sur les connaissances décrites dans les documents publiés sur le web vu l'absence de modèles permettant la représentation sémantique de ces connaissances.[19]

Tous ces problèmes ont fait l'objet de différents travaux de recherche qui ont convergé vers plusieurs solutions parmi lesquelles celles qui semblent les plus essentielles :

- Proposer des langages et des formalismes de représentation et de structuration des connaissances (représenter le contenu sémantique des ressources du web).
- Rendre disponibles des ressources conceptuelles (des modèles) représentées dans ces langages modélisant les connaissances et facilitant leur accès et leur partage : les ontologies.
- Proposer des métadonnées explicites, c'est-à-dire qui suivent un modèle et qui sont exprimées dans des langages définis formellement.[2]

3. Les Ontologies

La notion d'ontologie a été introduit en Intelligence Artificielle (IA) il y a 25 ans, le terme d'ontologie est cependant usité en philosophie depuis le XIXème siècle. Dans ce domaine, l'ontologie est une étude de l'être entant qu'être, c'est-à-dire, une étude des propriétés générales de ce qui existe. 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.[27]

3.1. Définitions :

Le terme « ontologie » est employé dans des contextes très différents touchant la philosophie, la linguistique ou l'IA. De nombreuses définitions ont été offertes pour donner

un éclaircissement sur ce terme, mais aucune de ces définitions ne s'est explicitement imposée. Les définitions de ce terme ne sont pas toujours consistantes et cela dépend des domaines spécifiques **[**24**]**. Pour ne pas dévier de notre propos, nous avons recensé les définitions suivantes :

Définition 1 : « 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.[28]

Définition 2 : «Une ontologie est une spécification explicite d'une conceptualisation».[31]

Cette définition est devenue la plus utilisée dans la littérature.

La conceptualisation se réfère ici à l'élaboration d'un modèle abstrait d'un domaine du monde réel en identifiant et en classant les concepts pertinents décrivant ce domaine. La formalisation consiste à rendre cette conceptualisation exploitable par des machines.[9]

Définition 3 : « une ontologie est une spécification explicite et formelle d'une conceptualisation partagée ».

Le terme « conceptualisation » réfère dans cette définition à une abstraction d'un phénomène du monde, obtenue en identifiant les concepts appropriés à ce phénomène. Le terme « Formelle » indique que les ontologies sont interprétables par la machine. Cependant, « Spécification explicite» signifie que les concepts de l'ontologie et les contraintes liées à leur usage sont définis de façon déclarative. Enfin, le terme «partagé» signifie que l'ontologie capture la connaissance consensuelle. Mais cette définition laisse la porte ouverte à de nombreuses définitions.[33]

Définition 4 : « Les ontologies sont des spécifications partielles et formelles d'une conceptualisation commune ».[25]

En 1997, Guarino accentue l'ambigüité du terme conceptualisation qui doit être pris dans son sens intuitif. La spécification des ontologies est partielle, car une conceptualisation ne peut pas toujours être entièrement formalisée dans un cadre logique, du fait d'ambiguïtés ou du fait qu'aucune représentation de leur sémantique n'existe dans le langage de représentation d'ontologies choisi. « Commune » renvoie à l'idée qu'une ontologie rend compte d'un savoir consensuel, c'est-à-dire qu'elle n'est pas l'objet d'un individu, mais qu'elle est reconnue par un groupe.[8]

Exemple d'ontologie : La figure 2.1 présente un exemple d'ontologie sur les formes géométriques, Cette ontologie contient un ensemble de concepts, comme 'Carré', un ensemble de relation, comme 'Contient' entre 'figure' et 'segment', et d'attributs, comme 'SeMesureEn', des instances, comme 'DEF' et enfin des types de données, comme entier.[26]

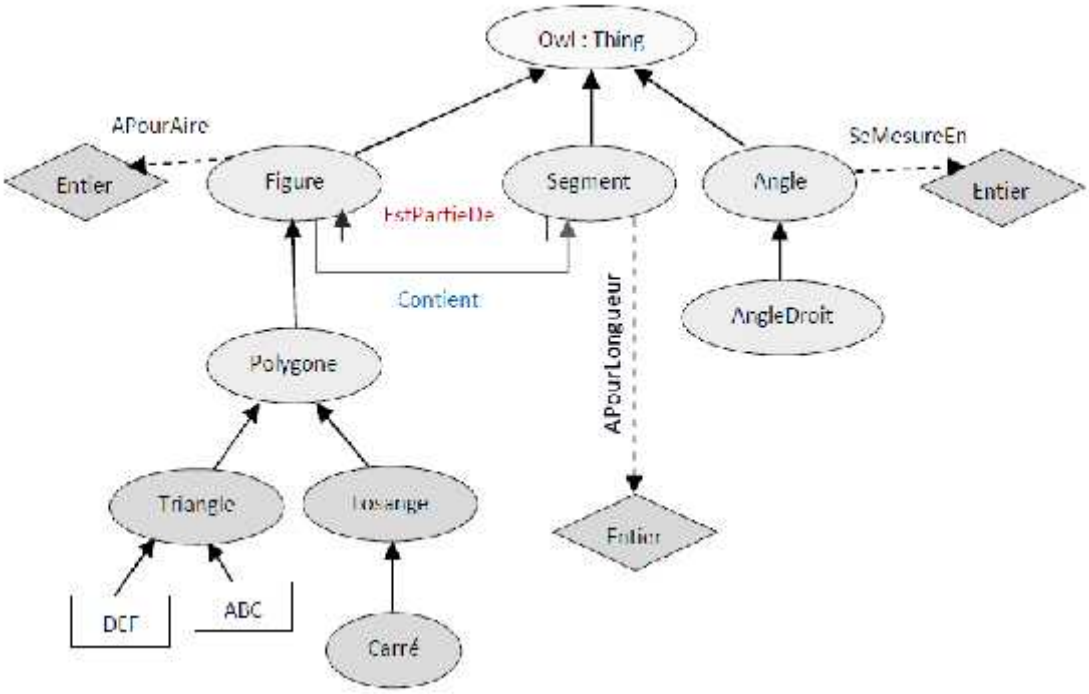

Figure 2.1 Une partie d'ontologie des formes géométriques.[26]

3.2. L'objectif d'ontologie

L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné, qui peut être réel ou imaginaire. Les ontologies sont employées dans l'intelligence artificielle, le Web sémantique, le génie logiciel, l'informatique biomédicale et l'architecture de l'information comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde.[34]

3.3. 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 :[14]

- 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.

3.4. Que représente-t-on dans une ontologie ?

Les ontologies produisent un vocabulaire commun d'un domaine et définissent, de façon plus ou moins formelle, la signification des termes et des relations entre eux. Les connaissances intégrées dans les ontologies sont formalisées en mettant en jeu cinq types de composants : concepts, relations, fonctions, axiomes, instances.[1]

3.4.1. Concepts :

Ils sont appelés aussi termes ou classes de l'ontologie. Un concept est un constituant de la pensée (un principe, une idée, une notion abstraite) sémantiquement évaluable et communicable. L'ensemble des propriétés d'un concept constitue sa compréhension ou son intention et l'ensemble des êtres qu'il englobe, son extension. [4]

3.4.2. Relations :

Représentent un type d'interaction, ou bien des associations existant entre les concepts d'un domaine. Elles se définissent formellement à partir d'un produit den concepts : R : C1× C2× …×Cn-1àCn. sous-classe-de (Spécialisation, généralisation), partie-de (agrégation ou composition), associée-à, instance-de sont des exemples de relations binaires. Voici quelques relations les plus courantes dans la littérature :

L'équivalence: une relation R est une relation d'équivalence si et seulement si : R est symétrique, réflexive et transitive. On écrit : une relation R est une relation d'équivalence si et seulement si : R est symétrique, réflexive et transitive. On écrit :
(R est une relation d'équivalence)ó((R symétrique) ^ (R réflexive) ^ (R transitive)).

la cardinalité: c'est le nombre possible de relations de ce type entre les mêmes concepts (ou instances de concept). Les relations portant une cardinalité représentent souvent des attributs.
Exemple : une pièce a au moins une porte, un humain a entre zéro et deux jambes.
L'incompatibilité: Deux relations sont incompatibles si elles ne peuvent lier les mêmes instances de concepts. Exemple : les relations «être rouge » et «être vert » sont incompatibles.
L'inverse: Deux relations binaires sont inverses l'une de l'autre si, quand l'une lie deux instances I1 et I2, l'autre lie I2 et I1.
Exemple : les relations « a pour père » et « a pour enfant » sont inverses l'une de l'autre .
L'exclusivité: Deux relations sont exclusives si, quand l'une lie des instances de concepts, l'autre ne lie pas ces instances, et vice-versa. L'exclusivité entraîne l'incompatibilité.
Exemple : l'appartenance et la non appartenance sont exclusives. Et bien d'autres relations…[9]

3.4.3. Fonctions :

Ce sont des cas particuliers de relations dans lesquelles le Nième élément de la relation est défini de manière unique à partir des n-1 premiers. Formellement, les fonctions sont définies ainsi : F: C1× C2× … ×Cn-1à Cn. Comme exemple de fonctions binaires, nous avons la fonction mère de et le carré, et comme exemple de fonction ternaire, le prix d'une voiture usagée sur lequel on peut se baser pour calculer le prix d'une voiture d'occasion en fonction de son modèle, de sa date de construction et de son kilométrage. [9]

3.4.4. Axiomes :

constituent des assertions, acceptées comme vraies, à propos des abstractions du domaine, traduites par l'ontologie. Ils ont pour objectif de représenter des concepts et des relations dans un langage logique permettant de représenter leur sémantique .Ils représentent les intentions des concepts et des relations du domaine et, de manière générale, les connaissances n'ayant pas un caractère strictement terminologique.[30] L'utilisation des axiomes sert à définir le sens des entités, mettre des restrictions sur la valeur des attributs, examiner la conformité des informations spécifiées ou en déduire de nouvelles.

3.4.5. Instances :

Elles constituent la définition extensionnelle de l'ontologie; ces objets véhiculent les connaissances (statiques, factuelles) à propos du domaine du problème.[10]

3.5. Types d'ontologies

Nous listons ci-dessous les différents types d'ontologies les plus utilisées:

3.5.1. 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.[31]

3.5.2. Les ontologies génériques (dites aussi de haut niveau) :

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. [2]

3.5.3. Les ontologies de domaine :

Elles fournissent un ensemble de concepts et de relations décrivant les connaissances d'un domaine spécifique. [2]

3.5.4. Les ontologies de tâches :

L'ontologie de tâche décrit les connaissances portant sur taches et/ou des activités particulières. Ces ontologies fournissent un ensemble de termes au moyen desquels on peut décrire au niveau générique comment résoudre un type de problème. Elles incluent des noms génériques (objectif, contrainte…), des verbes génériques (classer, sélectionner,…), des adjectifs génériques (assigné,..) et autres dans les descriptions de tâches. [9]

3.5.5. Les ontologies d'application :

Aussi appelée ontologie de domaine-tache : Ce sont les ontologies les plus spécifiques, elles contiennent les connaissances requises pour une application particulière permettant ainsi de modéliser une activité spécifique dans un domaine donné. [10]

3.6. Utilisation des ontologies

Même si le besoin de développer une ontologie est très varié et dépend du domaine d'application, on pouvons facilement énumérer un certain nombre d'utilités, notamment:

3.6.1. La connaissance du domaine :

Les ontologies permettent la modélisation des connaissances dans un domaine particulier, dans lequel opère le système à développer.

3.6.2. La communication:

Les ontologies assurent une communication fiable et hétérogène entre personnes et machines (agents logiciels ou organisations) du fait qu'elle permet de mettre en place un langage ou un vocabulaire conceptuel commun.

3.6.3. L'interopérabilité :

La représentation explicite des connaissances dans un domaine donné sous forme d'une ontologie, permet à son tour une plus grande réutilisation, un partage plus large et une interopérabilité plus étendue.

3.6.4. L'aide à la spécification des systèmes:

La représentation conceptuelle des éléments du domaine, permet aux systèmes de réaliser des raisonnements logiques qu'on appelle inférences, et de sortir avec des conclusions capables d'aider l'utilisateur ou le gestionnaire dans ses décisions.

3.6.5. L'indexation et la recherche d'information:

Dans le web sémantique, d'une façon générale, dans certains cas en particulier, les ontologies sont utilisées pour indexer et décrire les ressources utilisées. Cela permet une plus grande précision dans les résultats des recherches ou d'assignation des ressources. [2]

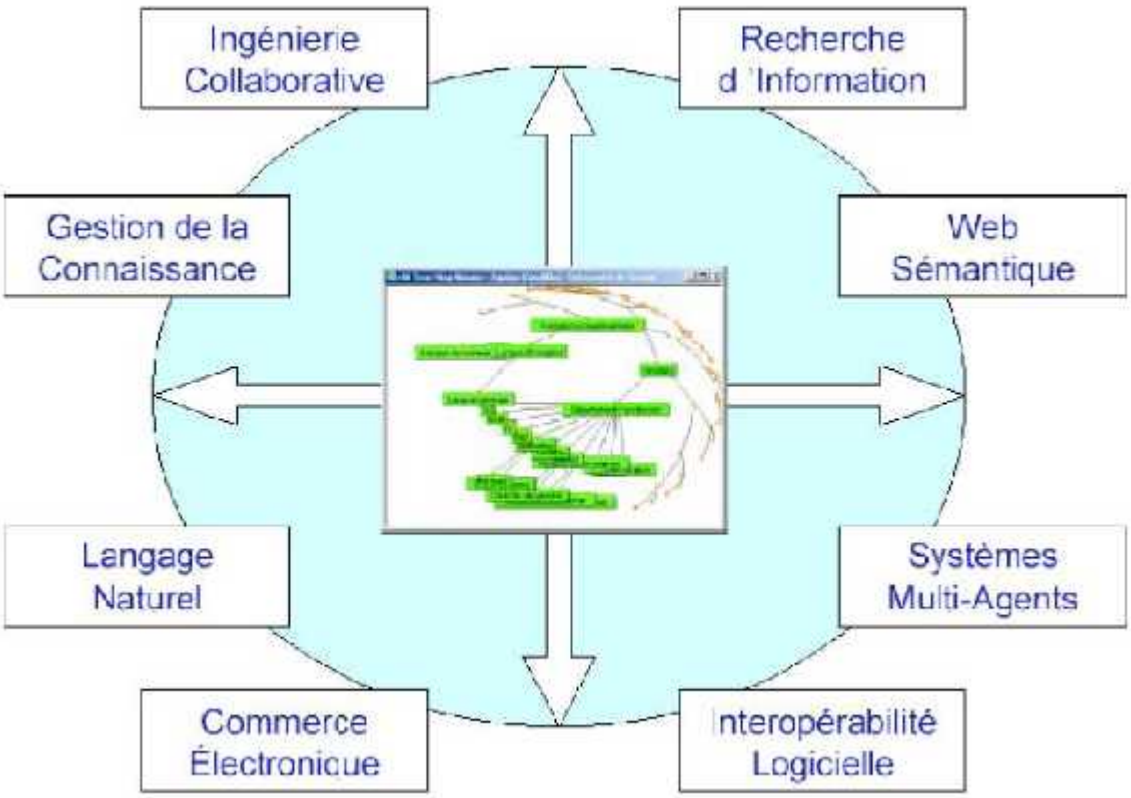

Figure 2.2 Quelque utilisation des ontologies. [9]

3.7. Construction d'une ontologie

La construction d'une ontologie, n'est pas quelque chose de facile. Elle demande un travail approfondi d'analyse et de compréhension du domaine et des utilisateurs du domaine. Le processus de construction d'une ontologie doit respecter certains principes de bases qui permettent d'obtenir une ontologie susceptible de répondre aux objectifs de l'ontologie. Le constructeur de l'ontologie, se doit donc de garder à l'esprit ces principaux critères tout au long du cycle de développement de son ontologie. [31]

La clarté et l'objectivité : l'ontologie doit fournir le sens des termes définis en offrant des définitions objectives ainsi que de la documentation associée en langage naturel.

L'exhaustivité: une définition exprimée par une condition nécessaire et suffisante est préférable à une définition exprimée seulement par une condition nécessaire ou par une condition suffisante.

La cohérence : afin de pouvoir formuler des inférences cohérentes avec les définitions.

L'extensibilité monotone maximale: Les nouveaux termes, qu'ils relèvent de la langue générale ou d'une langue de spécialité, devraient être inclus dans l'ontologie sans entrainer de modifications dans les définitions existantes. [9]

3.8. Recherche d'information guidée par les ontologies

L'utilisation typique du Web actuel consiste en la recherche 'information qui peut être d'ordre professionnel (veille stratégique/technologique, recherche d'articles…) ou d'ordre personnel (recherche de personnes ou de produits).

Pour faciliter ces tâches, plusieurs moteurs de recherche ont vu le jour (Google, Yahoo, Altavista…). Ces outils, bien qu'ils répondent à une bonne partie des besoins des utilisateurs, présentent quelques problèmes critiques :

- La masse énorme des documents retournés,
- La sensibilité au vocabulaire utilisé dans la requête,
- Le résultat fractionné en pages Web.
- La variabilité des langages utilisés sur le web et le non structuration des documents, ce qui rend cette tâche de plus en plus laborieuse.

Prenons l'exemple d'une personne anglo-saxonne qui cherche à trouver l'adresse d'un installateur de fenêtres ; en tapant la requête « Windows installation » dans n'importe quel moteur de recherche, elle obtiendra des milliers de pages traitant l'installation du système d'exploitation de Microsoft et les problèmes qui en résultent, mais elle aura beaucoup de mal à trouver l'information qu'elle recherchait. [19]

Avec l'utilisation d'une ontologie, un moteur de recherche fera la différence entre un site sur lequel 'Windows' désigne un logiciel et un autre sur lequel il désigne une fenêtre.

Cette recherche basée sur les ontologies se présente comme une recherche intelligente qui repose sur la sémantique des ressources et sur les concepts contenus dans les documents qui leur sont associés. Ces ontologies peuvent ainsi, d'une part, guider la création d'annotations sous la forme de métadonnées sur les ressources, et d'autre part, décrire leurs contenus de manière à la fois formelle et signifiante pour être exploitable aussi bien par les humains que par les machines. [2]

La recherche sémantique ou 'guidée par les ontologies' est une inférence exécutée par un raisonneur sur un ensemble de règles et de relations entre les instances modélisées par un langage formel telles que les logiques de descriptions et les règles SWRL. Cette recherche permet d'accéder aux ressources selon leur contenu plutôt que par mot clés. Les annotations des documents et la requête sont exprimés en utilisant le vocabulaire de l'ontologie. [2]

3.8.1. Exemple de l'utilisation d'ontologie dans la recherche d'information : moteur de recherche sémantique

La recherche sémantique est une inférence exécutée par un raisonneur sur un ensemble de règles et de relations entre les instances. Dans la plupart des cas, les règles sont modélisées par un langage formel tel que les logiques de descriptions et les règles SWRL(Semantic Web Rule Language) ou par des graphes conceptuels (GC).

A titre d'exemple, citons le moteur de recherche sémantique ***Corese*** synonyme de COnceptual REssource moteur de recherche. Il s'agit d'un moteur basé sur RDF Conceptual Graphs (CG). Il permet le traitement de RDF et RDF Schéma déclarés dans le formalisme CG. Les principales fonctionnalités de Corese sont dédiées à extraire des ressources Web annotées dans RDFS, en utilisant un langage de requête basé sur SPARQL et un moteur de règles d'inférence. La recherche est basée sur les annotations sémantiques qui sont des instanciations des schémas RDFS. [9]

4. Conclusion

Nous avons présenté dans ce chapitre les notions liées aux ontologies et au Web Sémantique. Cette présentation, bien que n'étant pas exhaustive car ce domaine est assez vaste.

Aujourd'hui, les ontologies apparaissent désormais comme une clé pour la manipulation automatique de l'information au niveau sémantique. Au fur et à mesure des recherches, des idées se dégagent autour du contenu des ontologies, des méthodes à utiliser pour les construire et des modèles et langages servant à leur représentation. Les ontologies contribuent à l'excellent travail dans le développement de moteurs de recherche sémantique.

CHAPITRE 3

CONCEPTION ET IMLÉMENTATION

1. Introduction

Les moteurs de recherche sont des outils informatiques qui ont pour but la mise en relation des informations contenues dans le corpus documentaire d'une part, et les besoins de l'utilisateur d'autre part. Dans ce chapitre nous présentons notre contribution basée sur la réponse du problème cité dans notre mémoire décrivant les langages et outils exploités pour créer les différentes parties de ce travail.

2. Fonctionnement du Système

2.1. L'objectif de notre travail

Moteur de recherche sémantique, où on a conçu un moteur de recherche classique relié à une ontologie générale qui représente des connaissances avec des relations sémantiques.

2.2. La conception du moteur de recherche

Pour la conception et la création il y a les étapes prédéfinis suivants :

2.2.1. L'exploration :

effectuée par le « crawler » pour construire une base de données contenant des liens restant à analyser, afin de se construire un index.

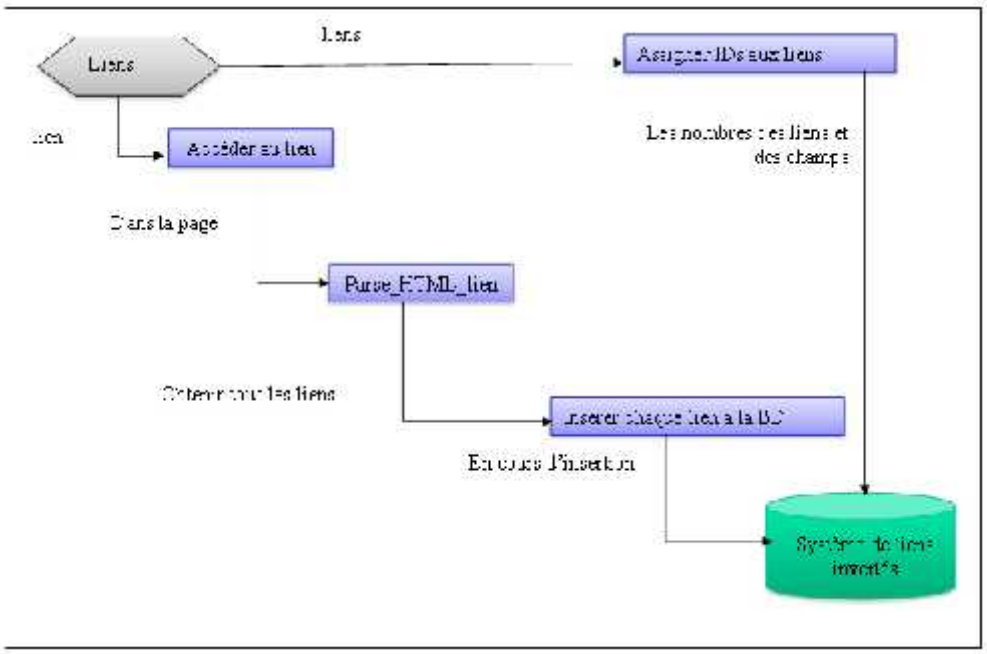

Figure 3.1 Le processus d'exploration.

❖ **L'algorithme d'exploration :**

```
Algorithme : Crawler

Input : L : ensemble de liens initiaux.

Output : BD de liens

Début

Contenu(BD)= L ;

Liens_pour_visiter=L ;

Tant_que liens_pour_visiter pas vide faire

    url= obtenir le lien suivant

      page= téléchargerPage (url) ;

      nouveau_urls= Parse_HTML_liens(Page) ;

      Pour chaque  nouveau_url : nouveau urls faire

            Si Contenu(BD) ne contient pas (nouveu_url) alors

               Insérer en queue (nouveu_url) dans BD

            Fin_Si

      Fin_pour

  Fin_Tant_que

Fin
```

2.2.2. L'indexation :

par accéder à la base de données des liens construit par le « crawler » et tenir les pages web des liens situées dans cette base de donnée pour les traiter à façon automatique passant des étapes prédéfinis afin de construire un index. Ces étapes sont comme suit :

- Détecter la langue : pour nous aider à poursuivre le traitement correctement.
- Faire l'analyse lexicale : consiste à décomposer notre fichier en mots simples.
- Eliminer les mots vides : allant de la langue détectée on peut déterminer la liste des mots vides à éliminer.
- Lemmatiser les mots du fichier : permet de ramener un mot à sa racine.
- Calculer la pondération : permet de mesurer l'importance d'un terme dans le fichier utlisant le principe TF-IDF.
- Créer l'index : insérer les termes traités avec ses informations dans une base de données pour l'utiliser dans la recherche.

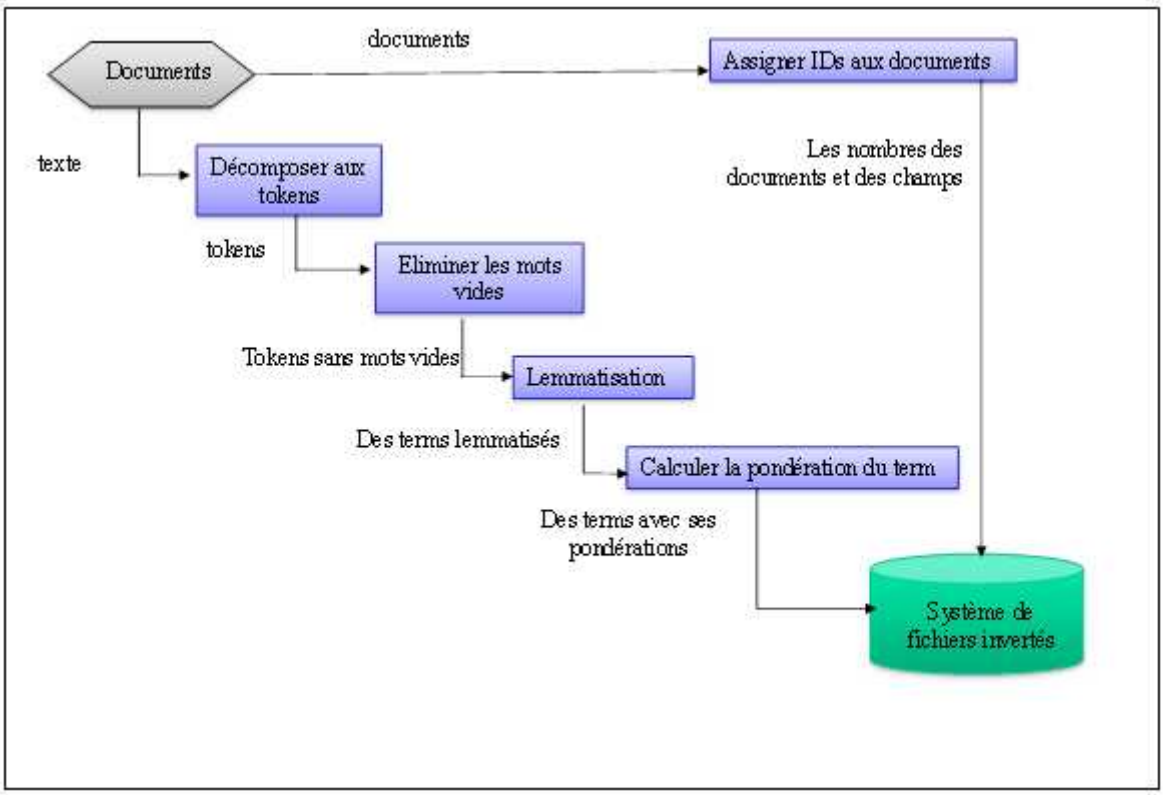

Figure 3.2 Le processus d'indexation

❖ **L'algorithme d'indexation**

Algorithme : Indexation

Input : L :BD_liens

Output : BD_Kwords

Début

url= le premier lien du BD_liens ;

Tantque url<> queue(BD_liens) **faire**

Page= téléchargerPage(url) ;

K_words=extraire_kwords(Page) ;

Pour chaque K_word :K_word **faire**

Si DB_Kwords contient(K_word)

Incrementer_nbr_occurence(

Update BD_Kwords set(nbr=nbr++) where BD_Kwords(Kword)=K_word ;

Sinon

Insérer_à_BD_Kwords(K_word ,1) ;

Fin_Si

Fin_pour

url= lien_suivant(BD_liens) ;

Fin_Tant_que

Fin

a. La recherche : qui présente l'interface entre l'utilisateur et notre système par permettant à ce dernier de saisir sa requête que va presque passer au même traitement d'un fichier avec des ajouts :

- Détecter la langue.
- Correcte l'orthographe : il permet de corriger les erreurs introduites dans les mots de la requête.
- Faire l'analyse lexicale.
- Eliminer les mots vides.
- Lemmatiser les mots de la requête.
- Obtenir l'ensemble des fichiers pertinents.
- Classer les fichiers selon leurs importances.

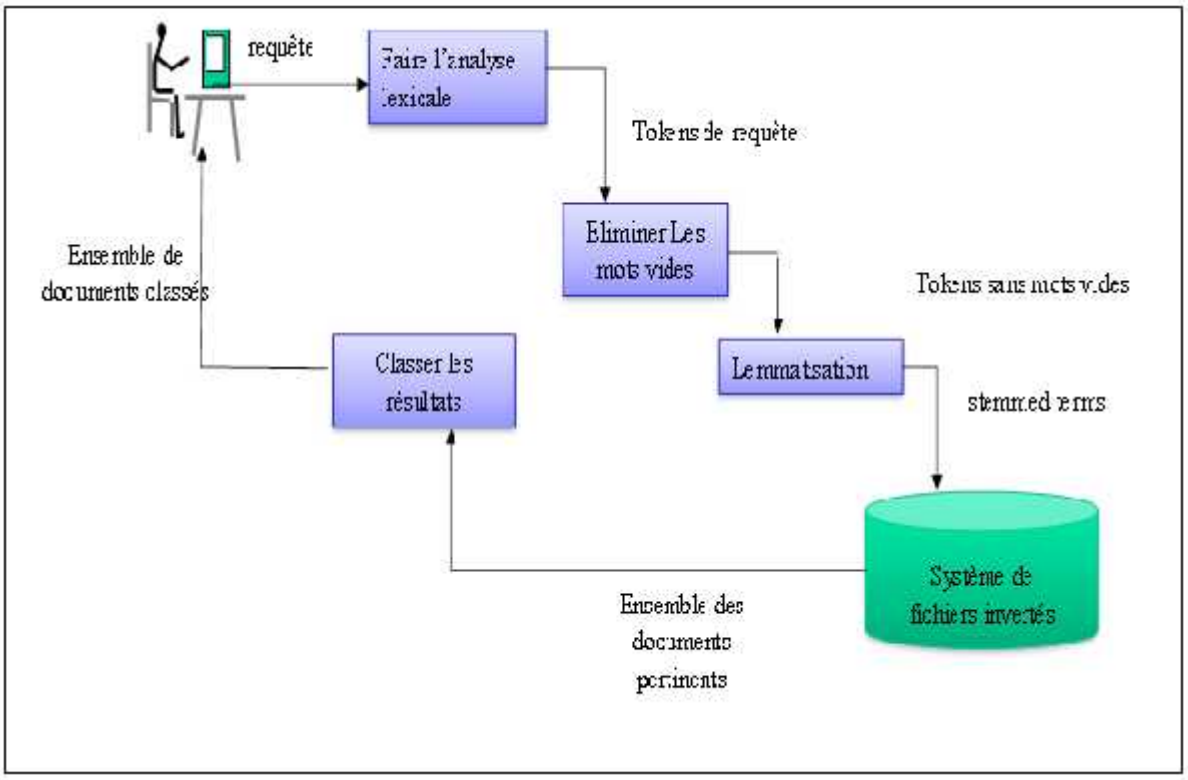

Figure 3.3 Le processus de recherche.

❖ **L'algorithme de recherche**

Algorithme : Search_Engine

Input :requête

Output :ensemble de fichiers.

Début

langue=detecter_la_langue (requête)

tokens=Analyse_lexicale(requête)

Suivant langue **faire**

Fr : Eliminer_les_mots_vides_française(tokens) ;

En : Eliminer_les_mots_vides_English(tokens) ;

Fin_Suivant

Pour chaque token : tokens **faire**

Lemmatiser(token) ;

Requête=concaténer(token) ;

Fin_pour

Docs=Extraire_doc_pertinent(requdoc_pertinent(requête) ;

Classer_doc(docs) ;

Afficher(docs) ;

Fin

2.3. L'ontologie

Pour réaliser un moteur de recherche sémantique on a besoin d'une ontologie qui couvre des concepts généraux, mais malheureusement on n'a pas trouvé ce type d'ontologie, donc on a fusionné un ensemble d'ontologies simples dans une.

2.4. Différences entre Moteur de Recherche Classique et Moteur de Recherche Sémantique

Moteur de Recherche Classique	Moteur de Recherche Sémantique
L'exploration Comme indiqué avant	**L'exploration** Comme indiqué avant
L'indexation Comme indiqué avant	**L'indexation** Comme indiqué avant
La Recherche -Détecter la langue. -Correcte l'orthographe. -Faire l'analyse lexicale. -Eliminer les mots vides. -Lemmatiser les mots de la requête. Obtenir l'ensemble des fichiers pertinents. -Classer les fichiers selon leurs importances.	**La Recherche** -Détecter la langue. -Correcte l'orthographe. -Faire l'analyse lexicale. -Eliminer les mots vides. -Lemmatiser les mots de la requête. Vérifier la sémantique de la requête par tester si les concepts existent dans une seule classe de notre ontologie ou s'il y a une relation entre eux utilisant les requêtes SPARQL. -Si les mots de la requête sont sémantique entre eux :obtenir les fichiers contenants seulement les concepts de notre ontologie. -Classer les fichiers selon leurs importances.

Figure 3.4 La différence entre Moteur de recherche classique et moderne

2.5. L'algorithme de recherche de moteur de recherche sémantique :

Algorithme : Search_Engine

Input :requête

Output :ensemble de fichiers.

Début

langue=detecter_la_langue (requête)

tokens=Analyse_lexicale(requête)

Suivant langue **faire**

Fr : Eliminer_les_mots_vides_française(tokens) ;

En : Eliminer_les_mots_vides_English(tokens) ;

Fin_Suivant

Pour chaque token : tokens **faire**

Lemmatiser (token) ;

Requête=concaténer (token) ;

Fin_pour

Si trouver_sémantique (requête) **alors**

requête= concepts_ontologie_pertinents(requête) ;

Docs=Extraire_doc_pertinent (requdoc_pertinent(requête)) ;

Classer_doc (docs) ;

Afficher (docs) ;

Fin_Si

Fin

3. Les outils de programmation

3.1. Choix du langage de programmation

Pour la langage de programmation on a choisi java

3.1.1. Java

Java est un langage de programmation orienté objet, développé par Sun Microsystems. Il fut présenté officiellement en 1995. Selon les développeurs de Sun, Java (qui signifie café en argot américain) est un langage : simple, orienté-objet, distribué, interprété, robuste sécurisé, neutre vis à vis de l'architecture, portable, à haute performance, multi-threaded et dynamique .

Le langage Java était à la base un langage pour Internet, pour pouvoir rendre plus dynamiques les pages (tout comme le JavaScript aujourd'hui). Mais le Java a beaucoup évolué et est devenu un langage de programmation très puissant permettant de presque tout faire. Contrairement à la plupart des autres langages (sauf la plateforme .Net), Java met à la disposition du développeur une API très riche lui permettant de faire de très nombreuses choses.[9]

3.2. Choix des éditeurs

3.2.1. NetBeans IDE 7.1.2

NetBeans est à l'origine un EDI Java qui fut développé par une équipe d'étudiants à Prague racheté ensuite par Sun Microsystems. En 2002 Sun a décidé de rendre NetBeans open-source. NetBeans n'est pas uniquement un EDI Java, c'est également une plateforme qui permet d'écrire des applications Swing ou C++, Python ou autres langages en lui incluant les plugins adéquats. Sa conception est complètement modulaire ce qui fait de lui une boite à outils facilement améliorable ou modifiable.

La License de NetBeans permet de l'utiliser gratuitement à des fins commerciales ou non. Elle permet de développer tous types d'applications basées sur la plateforme NetBeans.

3.2.2. Protégé 4.3

Protégé est un éditeur qui permet de construire une ontologie pour un domaine donné, de définir des formulaires d'entrée de données et d'acquérir des données à l'aide de ces formulaires sous forme d'instances de cette ontologie. Protégé est également une librairie Java

qui peut être étendue pour créer de véritables applications à bases de connaissances en utilisant un moteur d'inférence pour raisonner et déduire de nouveaux faits par application de règles d'inférence aux instances de l'ontologie et à l'ontologie elle-même (méta-raisonnement).

3.3. Choix des outils et des technologies supplémentaires

3.3.1. SPARQL

SPARQL est l'équivalent de SQL car comme en SQL, on accède aux données d'une base de données via ce langage de requête alors qu'avec SPARQL, on accède aux données du Web des données. Cela signifie qu'en théorie, on pourrait accéder à toutes les données du Web avec ce standard. L'ambition du W3C est d'offrir une interopérabilité non pas seulement aux niveaux des services, comme avec les services Web, mais aussi aux niveaux des données structurées ou non qui sont disponible à travers l'Internet Ce standard a été créé par le groupe de travail DAWG du W3C, SPARQL est considéré comme l'une des technologies clés du Web sémantique et le 15 Janvier 2008, la version 1.0 est devenue une recommandation officielle du W3C. La version 1.1 permettra d'enregistrer des données et de fusionner des données de sources différentes. Le dernier appel à contribution a eu lieu donc la version 1.1 sera bientôt candidat à la recommandation pour devenir un nouveau standard du Web.[9]

3.3.2. Jsoup (Java HTML Parseur)

Jsoup est une bibliothèque Java pour travailler avec le monde réel HTML. Il fournit une API très pratique pour extraire et manipuler des données, en utilisant le meilleur des DOM, CSS, et les méthodes de jquery.

Jsoup implémente le HTML5 spécification, analyse et HTML à la même DOM comme le font les navigateurs modernes.

- Gratter et analyser HTML à partir d'une URL, un fichier ou chaîne
- Trouver et extraire des données, en utilisant traversée DOM ou les sélecteurs CSS
- Manipuler les éléments HTML, les attributs, et le texte propre contenu soumis par les utilisateurs contre un blanc-liste de sécurité, pour empêcher les attaques XSS sortie HTML tidy.

Jsoup est conçu pour faire face à toutes les variétés de HTML trouvés dans la nature de l'impeccable et la validation, au invalide tag soupe, jsoup créera un arbre d'analyse sensible.[36]

3.3.3. Jena

Jena est un ensemble d'outils dédiés à la construction d'applications orientées Web sémantique.

Parmi ces outils, on trouve notamment une API Java open-source permettant de manipuler de nombreux langages tels que OWL, RDF/RDFS, SPARQL ou encore N3 et de raisonner sur des modèles ontologiques à l'aide de moteurs d'inférences inclus dans Jena ou externes.

3.3.4. Applet

Un applet est un programme conçu à l'aide du langage de programmation Java, utilisé à l'aide d'un navigateur Internet. Ils utilisent la machine virtuelle Java (Java Virtuel Machine) pour s'exécuter. L'avantage des applets java est donc de pouvoir être utilisé dans des pages Internet, permettant ainsi plus d'interaction avec le visiteur. Les applets java sont multi-plateformes, c'est-à-dire qu'ils fonctionnent de la même manière quel que soit le système d'exploitation : Windows, Linux, Mac OS , Unix..

4. Implémentation

4.1. L'interface graphique du notre moteur de recherche :

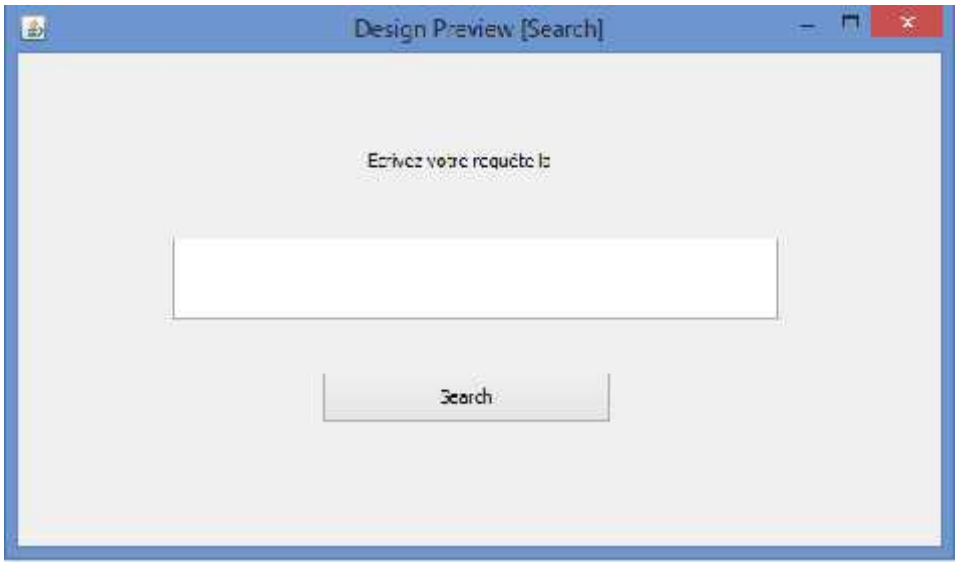

Figure 3.5 La fenêtre principale

5. Conclusion

Nous avons vu dans ce chapitre les différents outils et technologies nécessaires pour la réalisation de notre système tels que l'environnement de développement, la langage de programmations, les bibliothèques open source utilisées…etc.

Nous avons montré l'implémentation de chaque composant de notre système proposé, afin d'assurer une présentation claire et détaillée de notre outil. De plus, on a cité et expliqué les caractéristiques de notre application. Finalement, nous avons donné quelques résultats obtenus on utilisant des captures d'écran.

CONCLUSION GÉNÉRALE

Dans notre travail, nous avons développé un moteur de recherche sémantique qui est relié à une ontologie qui représente des connaissances avec des relations sémantiques, pour obtenir des résultats plus pertinents, en utilisant l'API Jena et des requêtes SPARQL sous l'environnement de programmation Netbeans.

L'objectif que nous avons fixé, au début de ce mémoire, est du développement d'un moteur de recherche sémantique relié à une ontologie générale bien définie qui contient des concepts dans des divers domaine avec des relations sémantiques mais malheureusement on n' a pas trouvé une telle ontologie combinant la majorité des concepts avec ses relations. Pour cela nous avons fusionné un ensemble d'ontologie dans une avec des contraintes d'intégrité bien sûr

Pour les perspectives on propose :

- Notre ontologie n'est pas générale, elle est simple donc on propose d'utiliser une ontologie plus riche.
- L'ajout d'un analyseur syntaxique aux étapes de développement d'un moteur de recherche pour des résultats plus pertinents.

BIBLIOGRAPHIE

[1] A.B.V.Gomez Pérez ,Overview of Knowledge Sharing and Reuse Components Ontologies and problem-Solving Methods. Dans Proceeding of the IJCAI-99 workshop on Ontologies and problem-Solving Methods (KRR5) (pp. 1.1-1.15). Stockholm (Suède),1999.

[2] A.Bouramoul, Recherche d'Information Contextuelle et Sémantique Sur Le Web, Thèse Doctorat en Science, Université MENTOURI de Constantine, 2011.

[3] A.F. Smeaton, Information retrieval and natural language processing, In proceedings of a conference jointly sponsored by ASLIB, University of York, page 2, march 1989.

[4] B.Bachimont, Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle, Compiègne, Mémoire d'Habilitation à Diriger des Recherches, Université de Technologie de Compiègne,2004

[5] C. Tambellini. Un système de recherche d'information adapté aux données incertaines: adaptation du modèle de langue. Thèse de doctorat en informatique, Université de Nice-Sophia Antipolis-UFR sciences, 2007.

[6] C.Roussey . Une méthode d'indexation sémantique adaptée aux corpus multilingues. Thèse informatique, l'institut national des sciences appliquées de lyon,2001

[7] F. Picarougne. Recherche d'information sur Internet par algorithmes évolutionnaires. Thèse de doctorat en informatique, Université François Rabelais Tours, page 29, Novembre 2004.

[8] F.Furst. L'ingénierie ontologique : Rapport de recherche, Magister,université MENTOURI DE CONSTANTINE, 2002.

[9] F.Moussaoui , conception et développement d'un outil de recherche sur le web à base d'agent, master, universite kasdi merbah ouargla,2013.

[10] G. Van Heijst, A. S. Using explicit ontologies in kbs development. International Journal of Human and Computer Studies. Knowledge, Acquisition: 183–292 pages.1997

[11] I. Chibane. “Impact des liens hypertextes sur la précision en recherche d’information. Conception d’un système de recherche d’information adapté au Web”, 2008

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment, J. ACM, 604–632, 1999.

[13] K. AMROUCHE. Passage à l’échelle en Recherche d’Information : Méthode d’élagage pour la réduction de l’espace de recherche, Thèse doctorat en informatique, Institut National de formation en Informatique (I.N.I) Oued-Smar Alger , 2008.

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

[15] K.Bal. Recherche d’information dans les documents XML: Approche par agrégation partielle les sources de pertinence.2010

[16] L. Maisonnasse. Les supports de vocabulaires pour les systèmes de recherche d’information orientés précision : application aux graphes pour la recherche d’information médicale. Thèse de doctorat en informatique, Université Joseph Fourier- Grenoble I, France, 2008.

[17] M. Charhad, Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l’Indexation et la Recherche par le Contenu Sémantique , pages 24-25, Novembre 2005.

[18] M.F.Porter, The Porter Stemming Algorithm,2006.

[19] M.K.Khelif, Web sémantique et mémoire d’expériences pour l’analyse du transcriptome, Thèse de doctorat en informatique, Université de Nice-Sophia Antipolis-UFR sciences, pages 7-16, Avril 2006.

[20] M.Marchal, Les moteurs de recherche Comment indexent-ils l'information, et comment la restituent-ils ?,EFREI.

[21] M.Nahrang , D.Delhomme, les moteurs de recherche, comment ça marche?, DESS IIR Réseaux , 2003-2004.

[22] Media Trend, http://www.themediatrend.com/wordpress/2008/11/20/les-moteurs-de-recherche-semantique-un-pas-dans-le-web-30/ ,consulté le 15/04/2015.

[23] N.D.Y. Kompaoré. "Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes: vers un processus de RI adaptatif». Thèse de doctorat en informatique, Université Paul Sabatier de Toulouse, 2008.

[24] N.F.McGuinness, Ontology Development 101: A Guide to CreatingStanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880,2001.

[25] N.Guarino , Understanding, building, and using ontologies, Dans I. J. Studies,1997. of a conference jointly sponsored by ASLIB, University of York, page 2, march 1989.

[26] 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.

[27] R. LEKHCHINE,Construction d'une ontologie pour le domaine de la sécurité : Application aux agents mobiles,Magister,2008/2009.

[28] 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, 16- 36, 1991.

[29] S. Brin, L. Page, The anatomy of a large-scale hypertextualWeb search engine, Computer Networks and ISDN Systems, 30(1–7)107–117, 1998.

[30] S. Staab, A. M., Axioms are objects too: Ontology engineering beyond the modeling of concepts and relations,Research report 399, Karlsruhe, Institute AIFB.2000

[31] T. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition, 199–220, 1993.

[32] T.D.Cao, Exploitation du web sémantique pour la veille technologique , thèse de doctorat en informatique Université de Nice-Sophia Antipolis-UFR sciences, 2006.

[33] W.N.Borst, Construction of engineering ontologies, University of Tweenty Centre for Telematica and Information Technology, Enschede,1997.

[34] Y.ASKANE,Y.EL-OUCHI ,S.EL MESSBAHI ,S.EL JADIDI ALAOUI, rapport l'onologie, Université Abdelmalek Essadi ,2009/2010.

[35] M.Roaissat, mise au point d'un robot d'indexage web, master, université de m'sila, 2012.

[36] Jsoup, http://www.jsoup.org ,consulté le 20/03/2015.

ملخص

قمنا في هذا العمل بتطوير محرك بحث دلالي يستند على أنطولوجيا عامة، التي تمثل معارف مع روابط دلالية للحصول على نتائج أكثر ارتباطا بالموضوع.

الكلمات المفتاحية:

البحث عن المعلومات،محرك البحث،الويب الدلالي ، أنطولوجيا.

Summary

In this work we've developped a search engine based on a general ontology that represents acquaintances with semantic relations, to obtain some more relevant results.

Keywords:

Information Search, Search Engine, Semantic Web, Ontology

Résumé

Dans ce travail on a développé un moteur de recherche sémantique basé sur une ontologie générale qui représente des connaissances avec des relations sémantiques, pour obtenir des résultats plus pertinents.

Mots clés :

Recherche d'Information, Moteur de Recherche,Web sémantique, Ontologie.