République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

Pour l'obtention du diplôme de Licence en Informatique

Thème

Moteur de recherche

Sémantique

Réalisé par :

- **Mlle. Hachemi Hadjira.**
- **Mlle. Rimouche Nour El Houda.**

Présenté le *27 Juin 2013 devant la commission d'examination composée de*

- *Mr. Bentaallah Mohammed Amine* *(Encadreur)*
- *Mr. Hadjila Feth Allah* *(Examinateur)*
- *Mr.Benazouz Mortada* *(Examinateur)*

Année universitaire: 2012-2013

Tables de matières

Tables de matières

Tables de matières

Introduction générale

Introduction général

Actuellement, le monde connait une avance technologique considérable dans tous les secteurs et cela grâce à l'informatique qui est une science qui étudie les techniques du traitement automatique de l'information. Elle joue un rôle important dans la société d'information d'aujourd'hui.

La recherche d'information est un domaine qui s'intéresse à la structure, l'analyse, à l'organisation, au stockage, à la recherche, à la découverte de la 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'opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI).

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, de même que Google. La pertinence des réponses est liée à un système de tri de pertinence construit sur la notion de lien existant entre les pages. Ce principe de recherche et d'évaluation est qualifier aujourd'hui de classique, et les approches en RI se sont orientées vers une nouvelle génération de systèmes de recherche basés sur l'accès contextuel et sémantique à l'information.

Ce mémoire s'inscrit dans les domaines de la recherche d'information, du Web sémantique. La recherche d'information sur le Web est actuellement principalement effectuée par les moteurs de recherche tels que Google, Yahoo et Bing.

L'objectif de notre projet présenté dans ce mémoire est la conception et la réalisation d'un logiciel permettant la création d'un moteur de recherche sémantique.

Le premier chapitre présente les concepts de base de la RI. Nous commençons par donner une définition de la RI et nous illustrons également le processus de RI en présentant les étapes, nous décrivons les différents modèles servant de cadre théorique pour la modélisation du processus de RI. Par la suite nous présentons les mesures et les compagnes d'évaluations, en fin nous terminons par les applications de la RI.

Le deuxième chapitre traite un réseau lexical (WordNet) qui couvre la grande majorité des noms, verbes, adjectifs et adverbes de la langue anglaise, dans un premier lieu nous commençons a donné une définition de WordNet, dans la section suivante nous présentons les principes de ce dernier et nous exposons par la suite les relations sémantiques en outre

quelques données statiques par la suite nous présentons les limites du WordNet, après nous décrivons quelques domaines d'applications, en fin nous terminons par ces avantages et ces inconvénients.

Le troisième chapitre présente l'essentiel de notre travaille, commençant par l'introduction, puis les étapes importante pour notre application, ensuite l'illustration des interfaces de l'application, on outre l'exposition de la requête, et finalement la représentation de l'environnement de développement (**NetBeans).**

Chapitre I : Recherche d'information

I. Introduction

Les récents progrès des technologies de l'information de manière générale et des réseaux de communication de manière particulière, ont redonné à l'information de nouveaux contours et davantage de valeur selon divers aspects : scientifique, technique, économique, d'usage, etc. Elle est devenue aujourd'hui un des biens les plus précieux et les plus stratégiques dans notre société guidée par la connaissance. Ainsi, avoir la bonne information au bon moment et en fonction de ses propres besoins est nécessaire pour prendre la bonne décision, mais faudrait-il savoir la localiser et la sélectionner dans ces masses d'information sans cesse croissantes.. Si on prend l'exemple du web, qui représente la plus grande source d'information disponible jusqu'à présent et qui ne cesse d'augmenter, un moteur de recherche populaire produit plus de huit (8) milliards de pages dans son index en juillet 2005 alors qu'elles étaient seulement 320 millions en 1997 et 3.3 milliards en septembre 2002. Le nombre d'utilisateurs est estimé aujourd'hui par des millions. Ces facteurs ont soulevé des défis majeurs pour les tâches de collecte et de gestion de l'information, le stockage, la transmission, et la recherche efficace de l'information.

II. Définitions

Plusieurs définitions de la recherche d'information ont vu le jour dans ces dernières années, nous citons dans ce contexte les trois définitions suivantes:

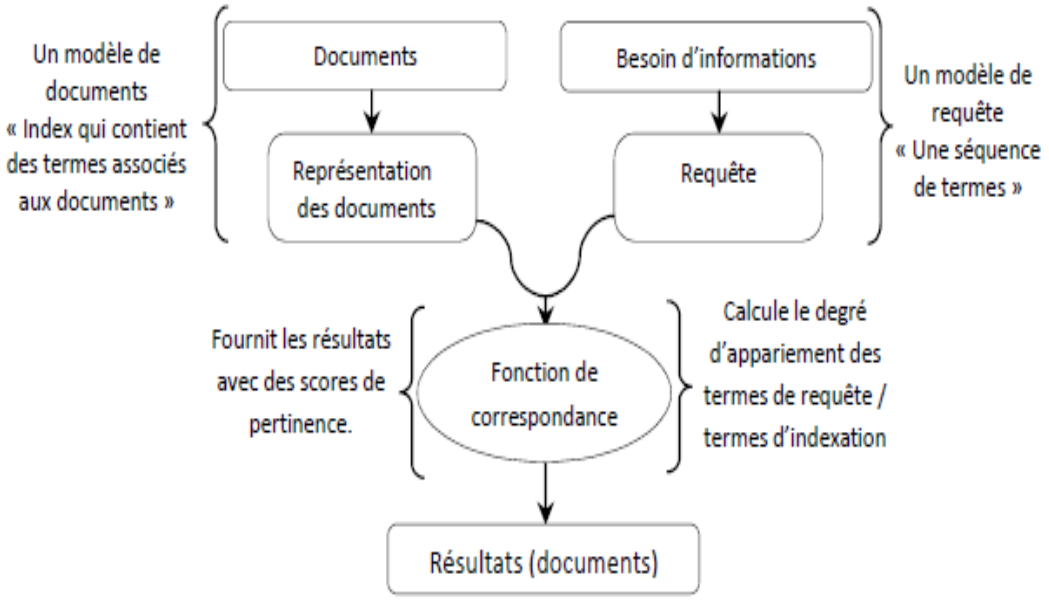
- La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations. [1]
- La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le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

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

III. Les étapes de la RI

Comme montré dans la FigureI.1, les étapes du processus de la RI sont les suivant :

1- la représentation de la requête.
2- La représentation des documents.
3- la fonction de correspondance.



FigureI.1 : les étapes de la RI

III.1. Représentation de la requête

Une requête s'agit d'une expression saisie pour interroger un système de recherche d'information afin de trouver les documents pertinents**.**

Le modèle de représentation de requête s'appuie d'une part sur la distinction de trois concepts : la Requête, la définition de requête et le Résultat de requête et d'autre part, sur la représentation de ces concepts sous forme de classes objets organisées dans des hiérarchies d'héritage. En fait, nous considérons que chaque requête est définie par

deux facettes : l'expression de définition de cette requête et l'ensemble des données résultantes de son évaluation.

III.2. Représentation des documents

Un objectif majeur de la RI et de retrouver des documents traitant du thème exprimé dans une requête utilisateur, la majorité des systèmes existants ont été implémentés de sorte à retrouver les documents ayant les mêmes mots que ceux de la requête. Ces systèmes souvent basé sur la notion de « sac de mots », font implicitement l'hypothèse qu'il y a une correspondance stricte entre les mots et les sens. Ils ne traitent donc pas des problèmes connus des linguistes, notamment la polysémie et la synonymie. Ce décalage entre l'objectif et la méthode a poussé des chercheurs à proposer de nouvelles approches tentant d'utiliser la sémantique en RI. [4]

III.3. Fonction de correspondance

La fonction de correspondance consiste à calculer la mesure entre les documents et les requêtes. Cette fonction est une variante de la fonction cosinus et la fonction sinus, Elle utilise uniquement les poids des termes dans les documents.

IV. Les modèles de RI

Un modèle de RI a pour but de produire une formalisation du processus de RI et une modélisation de la mesure de pertinence. Il existe un grand nombre de modèles de RI, 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. Voilà le vocabulaire d'indexation qu'on a utilisé :

$V = \{t_i\}$, $i \in \{1, ..., n\}$ est constitué de n mots ou racines de mots qui apparaissent dans les documents. Un modèle de RI est défini par un quadruplet

(D, Q, F, R (q,d)) : où

– D est l'ensemble de documents.

– Q est l'ensemble de requêtes.

– F est le schéma du modèle théorique de représentation des documents et des requêtes.

– R (q,d) est la fonction de pertinence du document d à la requête q nous présentons dans la suite les principaux modèles de RI : le modèle booléen, le modèle vectoriel et le modèle probabiliste. [5]

IV.1. Modèle booléen

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

$d = t_1 \wedge t_2 \wedge t_3 ... \wedge t_n$.

Une requête est une expression booléenne dont les termes sont reliés par des opérateurs logiques (OR, AND, NOT) permettant d'effectuer des opérations d'union, d'intersection et de différence entre les ensembles de résultats associés à chaque terme. Un exemple de représentation d'une requête est comme suit :

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

La fonction de correspondance est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document *dj* implique l'expression logique de la requête *q*. Le résultat de cette fonction est donc binaire est décrit comme suit :

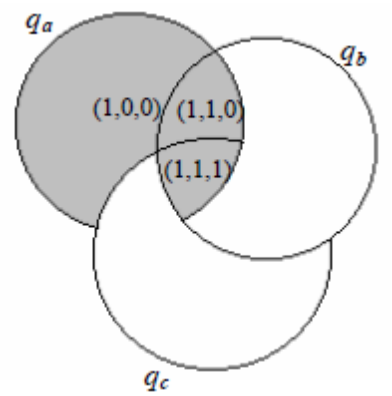
RSV (q, d) = {1,0}. [6]

Le modèle booléen a des avantages qui sont présentés par :

- ✓ Le modèle de recherche booléen est reconnu pour sa force pour faire une recherche très restrictive et obtenir, pour un utilisateur expérimenté, une information exacte et spécifique.
- ✓ La simplicité du modèle le rend aisément compréhensible pour un utilisateur.
- ✓ L'efficacité du modèle est due aux spécialistes qui ont explorés le corpus avec une bonne connaissance du vocabulaire.
- ✓ La formulation des requêtes devient vite laborieuse quand la requête se fait précise (donc longue).

Ce modèle n'a pas seulement d'avantages, il a aussi des inconvénients qui sont les suivants :

- ✓ L'inconvénient majeur de ce modèle comme schématisé dans la FigureI.2, est que les documents pertinents dont la représentation ne correspond qu'approximativement à la requête ne sont pas sélectionnés.



FigureI.2 : Les 3 composants conjonctifs pour la requête : $q = q_a \wedge (q_b \vee \neg q_c)$

- ✓ tous les termes ont la même importance et il est incapable de trier les documents pertinents.
- ✓ L'impossibilité de rendre compte d'une correspondance partielle d'un document à une requête.
- ✓ la pondération binaire des termes du vocabulaire limite la pertinence des résultats et ne permet pas de les ordonner.

IV.2.Modèle vectoriel

La pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Le modèle vectoriel représente les documents et les requêtes par des vecteurs d'un espace à *n* dimensions, les dimensions étant constituées par les termes du vocabulaire d'indexation. L'index d'un document d_j est

le vecteur $\vec{}$= ($w_{1,j}$,$w_{2,j}$,$w_{3,j}$, ..,$w_{n,j}$), où $w_{k,j} \in [0\ 1]$ dénote le poids du terme t_k dans le document *dj*. Une requête est également représentée par un vecteur V=($w_{1,q}$,$w_{2,q}$,$w_{3,k}$,..,$w_{n,q}$),où $w_{k,q}$ est le poids du terme t_k dans la requête q. La fonction de correspondance mesure la similarité entre le vecteur requête et les vecteurs documents. Une mesure classique utilisée dans le modèle vectoriel est le cosinus de l'angle formé par les deux vecteurs :

Cosinus: $B_{index} \times \{\bar{Q}\} \rightarrow R_+$

$$(\bar{d}_i, \bar{Q}) \rightarrow \frac{\sum_{t_j \in \bar{d}_i \cap Q} w_{ji}\, q_j}{(\ \sum_{t_j \in \bar{d}_i} w_{ji}^2)^{\frac{1}{2}} (\sum_{t_j \in \bar{Q}} q_j^2)^{\frac{1}{2}}}$$

RSV(q,d) = cos($\vec{q}$,$\vec{d}$).

Plus les vecteurs sont similaires, plus l'angle formé est petit, et plus le cosinus de cet angle est grand. A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante.

Le modèle vectoriel a des avantages qui sont présentés par :

- ✓ Le modèle vectoriel est relativement simple à appréhender (algèbre linéaire) et est facile à implémenter.
- ✓ Il permet de retrouver assez efficacement des documents dans un corpus non structuré et cela dépend de la qualité de la représentation.

Ce modèle n'a pas seulement d'avantages, il a aussi des inconvénients qui sont les suivants :

- ✓ La représentation vectorielle permet une mise en correspondance des documents avec une requête imparfaite.
- ✓ Il comporte également plusieurs limitations qui furent, pour certaines, corrigées par des affinements du modèle.
- ✓ L'indépendance des termes représentatifs supposée par le modèle.
- ✓ Dans un texte l'ordre des mots, les synonymes, la morphologie des contenus ne sont pas pris en compte.

IV.3.Modèle probabiliste

Le modèle de recherche probabiliste utilise un modèle mathématique fondé sur la théorie de la probabilité. Le processus de recherche se traduit par calcul de proche en proche, du degré ou probabilité de pertinence d'un document relativement à une requête. Pour ce faire, le processus de décision complète le procédé d'indexation probabiliste en utilisant deux probabilités conditionnelles :

P(w$_{ji}$/pert) : Probabilité que le terme t_i occurre dans le document D_j sachant que ce dernier est pertinent pour la requête.

P (w$_{ji}$/Nonpert) : Probabilité que le terme t_i de poids d_{ji} occurre dans le document D_j sachant que ce dernier n'est pas pertinent pour la requête.

Si on suppose l'indépendance des variables documents « pertinents » et « non pertinents », la fonction de recherche peut être obtenue en utilisant la formule deBayes.[7]

Soit D_j *(t_1, t_2, …, t_N).*

Où

t_i= 1 si t_i indexe le document D_j.

0 sinon

$$P(\text{pert}, \text{Dj}) = \frac{\text{p(Dj /pert)-} * \text{p(pert)}}{\text{p(Dj)}}$$

Et

$$P(\text{Nonpert}, \text{Dj}) = \frac{\text{p(Dj /nonpert)-} * \text{p(nonpert)}}{\text{p(Dj)}}$$

Avec :

P(pert/D_j) est la probabilité de pertinence du document *Dj* sachant sa description.

*P(D_j)=p (D_j/pert) * p(pert) + p(D_j/Nonpert) * p(Nonpert)*

p (D_j/pert) (respectivement *P (D_j/Nonpert)* est la probabilité d'observer le document D_j sachant qu'il est pertinent (respectivement non pertinent)

Le modèle probabiliste a des avantages qui sont présentés par :

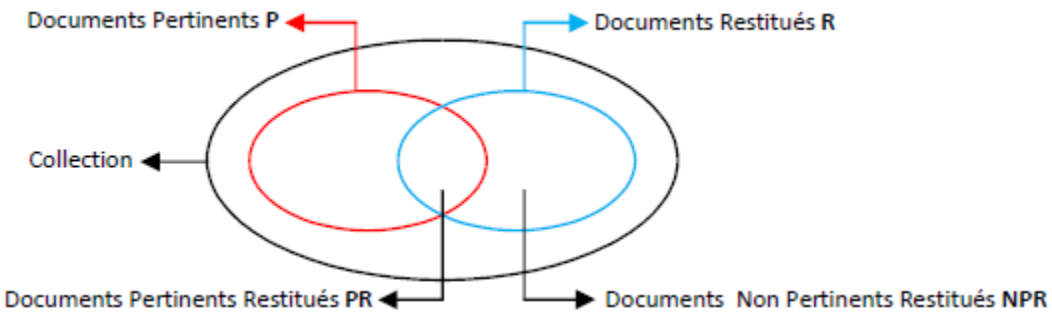
- ✓ Il a une base théorique saine et il est indépendant du domaine d'application.
- ✓ Pour des raisons de simplicité, l'hypothèse de l'indépendance des termes est utilisée en pratique pour implémenter ces modèles.

V. Mesures d'évaluation

Le but du processus de recherche d'information est de minimiser la distance entre la pertinence système et la pertinence utilisateur. Plusieurs mesures standards en RI ont été proposées pour évaluer les performances des SRI. [8]

V .1. Rappel et Précision

Le rappel et la précision sont deux mesures de base pour évaluer les performances des systèmes de recherche d'information. Le principe de ces deux mesures est basé sur la connaissance a-priori des documents pertinents de la collection d'une part, et d'autre part la partition de l'ensemble des documents restitués par le SRI en deux catégories : documents pertinents et documents non pertinents. La figureI.3, illustre la partition de la collection de tests pour une requête.



FigureI.3 : partition de la collection pour une requête [9].

- **Le Rappel**

Cette mesure peut être vue comme une mesure de couverture du système. Elle calcule la capacité du SRI à retrouver les documents pertinents de la collection. Le rappel indique le pourcentage de documents pertinents qui ont été retrouvés par le SRI par rapport à l'ensemble des documents pertinents de la collection.

$$Rappel_i = \frac{documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ appartenant\ à\ la\ classe\ i}$$

- **La Précision**

Cette mesure calcule la capacité du SRI à retrouver uniquement les documents pertinents. La précision permet de mesurer la fraction des documents pertinents parmi ceux qui ont été retrouvés par le système.

$$Précision_i = \frac{documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ attribués\ à\ la\ classe\ i}$$

Un SRI idéal est un système qui restitue tous les documents pertinents (Rappel = 1), et tous les documents qu'il retrouve sont pertinents (précision =1) pour la requête de l'utilisateur. En pratique, cet idéal n'est jamais atteint puisque ces deux quantités évoluent en sens inverse.

Intuitivement, si on augmente le rappel en retrouvant plus de documents pertinents, et si on diminue la précision en retrouvant aussi plus de documents non pertinents. Inversement, une plus grande précision risque de rejeter des documents pertinents diminuant ainsi le rappel. [10]

V.2.F-mesure

Une mesure populaire qui combine la précision et le rappel est leur pondération, nommée F-mesure ou F-score :

$$F = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$

Ceci est connu comme mesureF_1, car précision et rappel sont pondérés de façon égale. Il s'agit d'un cas particulier de la mesure générale F_β (pour des valeurs réelles positives de β):

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{précision} \cdot \text{rappel})}{(\beta^2 \cdot \text{précision} + \text{rappel})}$$

VI. Les Campagnes d'évaluations

Les campagnes d'évaluation en RI permettent d'évaluer sur des collections différentes plusieurs SRI, afin de valider les différents modèles mis en œuvre, et comparer les systèmes.

Les objectifs essentiels des campagnes sont les suivants :

- Encourager la RI sur de grandes collections fermées.
- Développer la communication entre l'industrie, l'académie et le gouvernement en mettant en place un forum ouvert pour faciliter les échanges d'idées sur la recherche.
- Augmenter la vitesse de transfert de la technologie du laboratoire de recherche aux enseignes commerciales.
- Rendre disponible et accessible des techniques d'évaluations appropriées pour les industriels et les académiciens. [11]

Chaque campagne est constituée d'un certain nombre de tâches fournissant des résultats, et un protocole d'évaluation pour chaque tâche. Les campagnes de TREC (TextREtrievalConference) qui ont vu le jour en 1992 avec 25 participants issus du monde académique et industriel sont devenues la référence en ce qui concerne l'évaluation des systèmes de recherche d'information. [12]

VII. Les applications de la RI

Le domaine de la RI a donné naissance à plusieurs applications :

VII.1.Moteur de recherche

Un moteur de recherche est une application web permettant de retrouver des ressources (pages web, articles de forums Usenet, images, vidéo, fichiers, etc.) associées à des mots quelconques. Certains sites web offrent un moteur de recherche comme principale fonctionnalité ; on appelle alors moteur de recherche le site lui-même (Google Vidéo par exemple est un moteur de recherche vidéo).

Les sites 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.

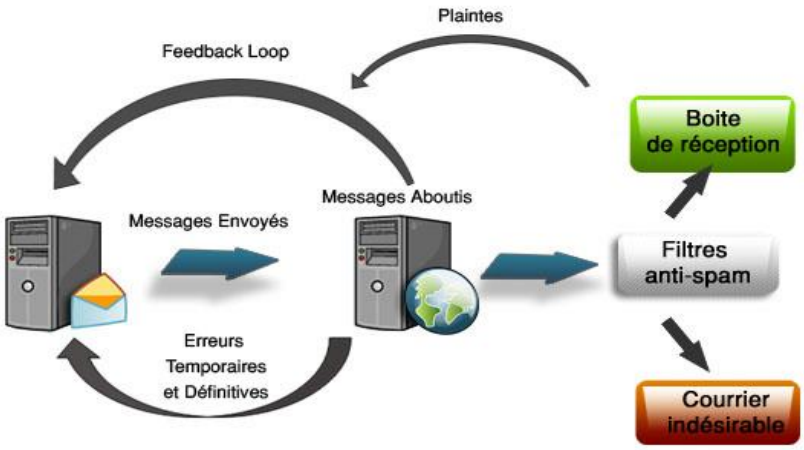
Les moteurs de recherche ne s'appliquent pas qu'à Internet : certains moteurs sont des logiciels installés sur un ordinateur personnel.

On trouve également des métas moteurs, c'est-à-dire des sites web où une même recherche est lancée simultanément sur plusieurs moteurs de recherche (les résultats étant ensuite fusionnés pour être présentés à l'internaute).

VII.2.filtrage des e-mails (spam)

Un spam est un courrier électronique envoyé en grand nombre, de façon anonyme ou sous une fausse identité, à des destinataires qui ne l'ont pas sollicité et que l'on maintient dans l'incapacité de s'opposer à cette diffusion.

La **lutte anti-spam** schématisé dans la figureI.4 est un ensemble de comportements, de systèmes et de moyens techniques et juridiques permettant de combattre le spam, courriers électroniques publicitaires non sollicités.



FigureI.4 : Filtrage des e-mails.

VII.3.Application Biomédicale

Dans le domaine biomédical plus encore que dans d'autres domaines, l'emploi de termes spécialisés est la clef de l'accès à l'information.

Pour faciliter l'accès à l'information de ce domaine, plusieurs terminologies ont été développées pour une indexation contrôlée des documents dans les portails de santé.

VII.4.La classification des textes

La classification automatique de textes est un domaine où la fouille de textes et les techniques statistiques produisent des résultats à partir des calculs de fréquence d'occurrence de termes extraits.

L'analyse syntaxico-sémantique était considérée, jusqu'à présent, comme pénalisante en raison des limitations des analyseurs eux-mêmes. Elleest peu sensible à la qualité des corpus d'entraînement puisqu'elle se sert de ressources stables (des dictionnaires non variants), alors que les méthodes statistiques y sont très sensibles. [13]

VIII. Conclusion

La recherche d'information peut être définie comme une discipline de recherche qui intègre des modèles et des techniques dont le but est de facilité l'accès à l'information. L'objectif principal du domaine RI est de fournir des modèles, technique et systèmes pour stocker et organiser des masses d'information.

Nous avons présenté dans ce chapitre les principales notions et concepts de la recherche d'information. Dans une première partie nous étions intéressés à la définition de la RI ainsi que les différentes étapes qui sont la représentation des documents, la représentation des requêtes et la fonction de correspondance, puis nous avons présenté les modèles de la RI dont l'objectif est de produire une formalisation du processus de RI et une modélisation de la mesure de pertinence.

Ce chapitre a mis en évidence les mesures d'évaluations qui ont été proposées pour évaluer les performances des SRI, puis les compagnes d'évaluation qui permettent d'évaluer sur des collections différentes plusieurs SRI, afin de valider les différents modèles mis en œuvre, et comparer les systèmes. Enfin nous avons proposé les applications de la RI(moteur de recherche, filtrage des spam, application biomédicale, classification des textes).

Chapitre II : Le WordNet

I. Introduction

Des programmes issus du monde de l'Intelligence Artificielle ont également établi des passerelles avec WordNet. Le WordNet est utilisable librement, y compris pour un usage commercial, ce qui en a favorisé une diffusion très large. Plusieurs autres ressources linguistiques ont été constituées (manuellement ou automatiquement) à partir de, en extension à, ou en complément à WordNet. L'ensemble de ces ressources linguistiques constitue un système complet couvrant des aspects lexicaux, syntaxiques et sémantiques. Combinées, ces ressources fournissent un point de départ intéressant pour des développements sémantiques dans le cadre du Web sémantique, tels que la recherche d'information, l'inférence pour la compréhension automatique de textes, la désambiguïsation lexicale ou la résolution d'anaphores.

II. Définition

WordNet est une base de données lexicale développée depuis 1985 par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. C'est un réseau sémantique de la langue anglaise, qui se fonde sur une théorie psychologique du langage. La première version diffusée remonte à juin 1991. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour.

WordNet est distribué sous une licence libre, permettant de l'utiliser commercialement ou à des fins de recherche.

La dernière version distribuée en avril 2013 est la 3.1. Cette version est par ailleurs consultable en ligne.[14]

III. Principe

WordNet est donc un réseau lexical où :

- Les synsets sont les nœuds.
- Les relations sémantiques entre synsets sont les arcs.

Chapitre II : Le WordNet

III.1. Les synsets

La composante atomique sur laquelle repose le système entier est le synset (*synonyme)* c'est un groupe de mots interchangeables, dénotant un sens ou un usage particulier. La version 2.0 de WordNet définit ainsi le nom commun anglais' *car'* à l'aide de cinq synsets comme il est montré dans la figure II.1.

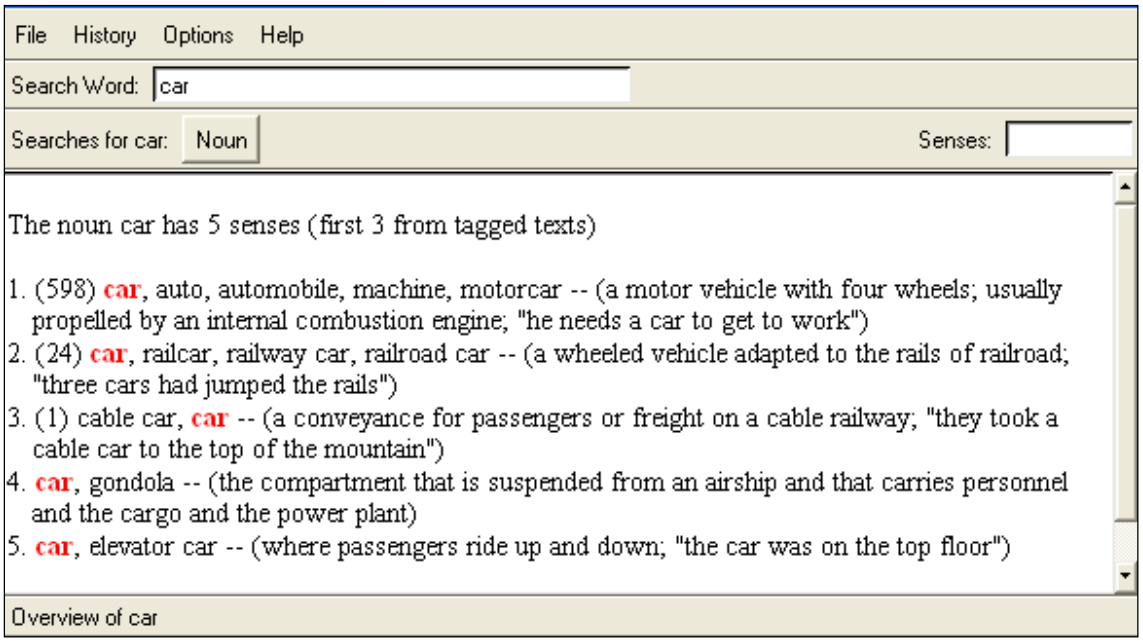


Figure II.1 : les différents sens de mot car.

Chaque synset dénote une acception différente du mot *car*, décrite par une courte définition. Une occurrence particulière de ce mot dénotant par exemple le premier sens (le plus courant), dans le contexte d'une phrase ou d'un énoncé, serait ainsi caractérisée par le fait qu'on pourrait remplacer le mot polysémique par l'un ou l'autre des mots du synset sans altérer la signification de l'ensemble.

III.2. Les relations sémantiques

III.2.1. L'hyperonymie

L'hyperonymie est la relation *sémantique* hiérarchique d'un lexème à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus

spécifique. Le premier terme est dit hyperonyme de l'autre, ou super ordonné par rapport à l'autre. C'est le contraire de l'hyponymie. [15]

III.2.2. L'hyponymie

L'hyponymie est une Relation d'inclusion entre deux mots dont l'un est l'hyponyme de l'autre. La relation d'hyponymie est l'expression linguistique de la relation logique d'inclusion d'une classe dans une autre (La Linguistique, Paris, Denoël, 1969, p. 193).

On peut aussi définir les hyponymie comme la relation sémantique d'un lexème à un autre selon laquelle l'extension du premier est incluse dans l'extension du second. Le premier terme est dit hyponyme de l'autre. C'est le contraire de l'hyperonymie.

III.2.3. La méronymie

La méronymie est une relation sémantique entre mots d'une même langue. Des termes liés par méronymie sont des méronymes. La méronymie est une relation partitive hiérarchisée : une relation de partie à tout. Un méronyme X d'un mot Y est un mot dont le signifié désigne une sous-partie du signifié de Y. La relation inverse est l'holonymie. WordNet inclus trois types de méronymie :

- X est un composante de Y.
- X est un élément de Y.
- X est le matériau dont Y est constitué.[17]

III.2.4. L'holonymie

L'Holonymie est une relation sémantique entre mots d'une même langue. Des termes liés par holonymie sont des holonomes. L'holonymie est une relation partitive hiérarchisée : un holonyme A d'un mot B est un mot dont le signifié désigne un ensemble comprenant le signifié de B. La relation inverse est la méronymie.[18]

III.2.5.La Synonymie

La synonymie est un rapport de similarité sémantique entre des mots ou des expressions d'une même langue. La similarité sémantique indique qu'ils ont des significations très semblables. Des termes liés par synonymie sont des synonymes.

Il existe des bases de données de synonymes, présentées comme des dictionnaires, librement téléchargeables. On en trouve aussi vendues ou consultables sous la forme de livres, de logiciels, ou de web, ou des jeux.[19]

III.2.6. L'antonymies

Deux items lexicaux sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe. La symétrie peut se décliner de différentes manières, selon la nature de son support. On distingue plusieurs supports qui sont autant de type d'antonymie :

- ✓ Les antonymes complémentaires.
- ✓ Les antonymes scalaires.
- ✓ Les antonymes duals. [20]

III.2.7 La Troponymie

La troponymie est une relation sémantique entre deux verbes, l'un décrivant de manière plus précise l'action de l'autre. Le premier verbe est dit troponyme du second. [21]

IV. Quelques données statistiques

Dans cette section, nous présentons, de manière quantitative, le contenu de WordNet. La table II.1 montre la structure de WordNet en nombre de mots, nombre de synsets et nombre de sens globalement et par catégorie grammaticale. Du nombre total de formes décrites, la plupart sont des noms (74.6%), le reste étant constitué par des adjectif(14.6%),des verbes(7.6%) et des adverbes(3.2%). La polysémie (nombre de sens par mot) se manifeste dans WordNet par le fait qu'il y a des mots qui peuvent appartenir à plusieurs synsets (146350 formes traitées /111223 synsets).

Partie de discours	Nombre de mots	Nombre de synsets	Nombre de sens
Noms	109195	75804	134716
Verbes	11088	13214	24169
Adjectifs	21460	18576	31184
Adverbes	4607	3629	5748
Total	146350	111223	195817

Table II.1 : Nombre de mots, synsets, et sens dans le WordNet.

Chapitre II : Le WordNet

V. Les limites du WordNet

V.1. Informations manquantes

WordNet ne précise pas l'étymologie, la prononciation, les formes de verbes irréguliers et ne contient que des informations limitées sur l'usage des mots.

V.2. Profusion de sens pour un mot donné

La contrepartie de son importante couverture est que WordNet est très précis dans le sens des définitions. On a une granularité très fine des sens. Par exemple, le verbe *to give* (« donner ») n'a pas moins de 44 sens. Une telle profusion ne facilite pas une tâche de désambiguïsation lexicale.

V.3. Absence de relations pragmatiques

WordNet ne matérialise pas d'une façon formelle tout le sens contenu dans les définitions des termes. Par exemple, l'information qu'un chat ne rugit pas figure dans la définition, mais ne se retrouve formalisée dans aucune relation. De même, des relations pragmatiques telles que savon / bain (SOAP#1 / BATH#2) sont absentes de WordNet.

VI. Quelque domaine d'application

L'utilisation de WordNet en recherche d'informations :

- Pour représenter les documents.
- Pour étendre la requête de l'utilisateur (ajout de synonymes, par exemple pour augmenter le rappel, c'est-à-dire la proportion de documents pertinents rapportés).
  - Acquisition de relations sémantiques.
  - Désambiguïsation sémantique.
  - Pour l'étiquetage sémantique de corpus.
  - Pour la structuration et catégorisation des documents.

Ressource lexicale permet d'incorporer certaines connaissances lexico-sémantiques au traitement des textes :

- Pour indexer les documents.

- Pour réduire les différences de vocabulaire, entre les textes et les questions sur ces textes.

En général WordNet est utilisé :

- Pour la recherche d'informations.
- Pour l'extraction d'informations.
- Pour les systèmes de questions/réponses.
- Pour enrichir la représentation avec des synonymes, hyperonymes,….

VII. Avantages et inconvénient

VII.1 Avantages

- Une des rares ressources pour la langue générale (anglais) disponible en ligne.
- Organisation de la base lexicale assez riche grâce aux diverses relations représentées.

VII.2. Inconvénients

- Distinctions de sens très (trop) fines, sans méthodologie précise pour les découper, sans repérer des processus lexicaux type métonymie, métaphores…
- Pas de définition des mots
- Aucune information syntaxique, morphologique, de dérivation,…

VIII. Conclusion

Dans ce chapitre nous avons présenté en détail WordNet et son principe de fonctionnement qui est basé sur la notion de synset et de relation sémantique. Â la fin du chapitre nous avons présenté les domaines d'application de cette base de données lexicale, ainsi que ses avantages et ses inconvénients.

Chapitre III : Recherche d'information sémantique

I. Introduction

Un Système de Recherche d'Information (SRI) permet de retrouver les documents pertinents par rapport aux besoins exprimés par l'utilisateur à travers une requête, à partir d'une base de documents.

Après avoir établi une étude conceptuelle de notre système, nous passons à l'implémentation de l'application définis et détaillée aux deux chapitres précédents tout en présentant les outils utilisés et en expliquant les étapes de notre projet .cette implémentation est basée sur la recherche d'information sémantique et réalisée avec le langage de programmation Java.

II. Les étapes de la présentation de l'application

Comme n'importe quel système de recherche d'information, notre système passe par les étapes suivantes :

II.1.Prétraitement

Avant d'utiliser les documents textuels, il est utile d'appliquer certains prétraitements afin de rendre les documents compréhensibles par la machine.

Voici un document sur lequel on va appliquer les étapes de la représentation de l'application :

GIRL?.girl !.house, 01 !! doctor ?? ..25 PEOPLE ..people ?? are 866 she

Ces prétraitements sont les suivants:

- **Tokenisation**

Dans cette étape, il s'agit d'enlever toute la ponctuation. Voici la liste des ponctuations qu'on a utilisé :{+-*/ ;:.() !, ?<>0123456789}

----------Le texte sans punctuation----------

GIRL girl house doctor PEOPLE people are she.

- **Elimination des majuscules**

En effet le mot "GIRL" et le mot "girl" vont être considérer différent alors qu'ils ont le même sens donc on transforme les majuscules en minuscule.

----------Le texte sans majuscule ----------

girlgirl house doctor peoplepeople are she.

II.2 Représentation

A la fin de l'étape précédente chaque document sera présenté par un vecteur dont chaque composant est un mot.

Exemple :

Girl	Girl	House	Doctor	people	People	Are	She

- **Elimination des mots vides**

Les mots vides sont les mots qui se répètentfréquemment danstous lesdocuments et qui n'ont aucun pouvoir discriminant lors du processus de la catégorisation de texte.

La listes des mots vides contient les pronoms personnels, les prépositions, les articles….etc. exemple {he, she, help, and, also,is…..}

Si le mot apparuest un mot vide alors le système va le supprimer.

Donc le tableau précédent devient :

Girl	Girl	House	doctor	People	People

- **Fréquence des mots dans les documents**

Le système calcule le nombre de chaque mot dans chaque document pour obtenir la matrice.

Exemple : si les documents contiennent les textes suivants :

Document 1 (girlgirl house doctor people).

Document 2 (girl house house doctor house house doctor doctor people).

Document 3(house doctor people doctor).

	Girl	House	Doctor	People
Document1	2	1	1	1
Document2	1	4	3	1
Document3	0	1	2	1

- **Mapping des mots en sens**

Cette étape consiste a remplacé le mot par ces sens ,en effet chaque mot peut appartenir à plusieurs sens.

Exemple: le mot people devient :{ a group of human beings ,citizenry, multitude , number of family line….}

- **Les interfaces de l'application**

Figure. III.1 :l'interface de l'application.

- **La sélection des sens**

L'utilisateur doit sélectionner le sens « sence » adéquat, pour cela il doit cocher

la case «all sences » en suivant le chemin File→ sence→all sences, le système prend en considération tous les sens du mot.

Dans le cas où la case « all sences » n'est pas cochée, le système va prendre en considération seulement le premier sens.

Figure.III.2 : La sélection du sens.

- **Extraction des relations sémantiques**

Dans cette étape l'utilisateur peut choisir les différentes relations sémantiques.

Notre système offre la possibilité d'enrichir la représentation en prenant en considération les relations suivantes en suivant le chemin :

file→Rel. Sem→Hypernyms/Hyponyms/Meronyms/Holonyms.

-Exemples de relations d'hyperonymie et d'hyponymie

Par exemple partant du sens le plus général du mot CAT (le <chat>félin),on obtient une liste ordonnée d'ancêtres et de descendants ,permettant de déterminer qu'un chat est un carnivore ,mammifère ,un animal, ect.

-exemple de relations d'holonymie et meronymie

Grâce à ces relations on peut déterminer qu'un chat a des pattes, un pelage, une queue….

Figure.III.3 : Sélection des relations sémantiques.

Comme montrer dans la Figure III.4 l'utilisateur doit donner une valeur précise pour la profondeur des relations sémantiques.

Pour l'exemple CAT :si on prend la profondeur égale à 2 on aura :

-Relation d'hperonymie

chat→félin→carnivore.

-Relation d'hyponymie

chat→chat domestique→minou.

-Relation de méronymie

cat→patte→doigt.

-Relation d'holonymie

cat→félidés→lion.

Figure.III.4 : La profondeur des relations sémantiques.

III. Requête

- **L'expression du besoin de l'utilisateur (requête)**

Dans cette étape l'utilisateur doit entrer une requête pour laquelle il veut avoir un résultat.

Figure .III.5 : Remplissage de la requête.

Cette requête entrée par l'utilisateur va passer par les mêmes prétraitements du document:

(tokenisation, élimination des majuscules, présentation, élimination des mots vides,

fréquence des mots dans les documents, mapping des mots en sens ,la sélection des sens, extraction des relations sémantiques).

- **La mesure de similarité**

Dans cette étape le système va répondre au besoin de l'utilisateur en lui donnant le résultat (document pertinent) après le calcul de la distance en utilisant la méthode de produit scalaire présenté par **:**

Dist[j]=dist[j]+req[i]*tf[j][i].

Exemple:

Req:

1	3	0	5	2

Tf:

	Sens1	Sens2	Sens3	Sens4	Sens5
Document 1	0	2	1	4	3

Dist= (0*1) + (2*3)+(1*0)+(4*5)+(3*2)

Alors **dist=32**

- **Trie et affichage du document**

Une fois la mesure de similarité a été calculé le système va afficher les documents trié

Plus la valeur de la mesure de similarité est grande plus le document est plus pertinent a la requête.

Figure.III.6 : Tri et affichage du document.

IV. Environnement de développement

Le langage utilisé pour notre application est JAVA, c'est un langage de programmation orienté objet, développé par Sun Microsystems, inspiré de C++.il permet de crée des logiciels compatible avec de nombreux systèmes d'exploitations

(Windows, Linux, Macintosh, Solaris).java donne la possibilité de développer des programme pour téléphone portable et assistants personnels.

Enfin, ce langage peut être utilisé sur internet pour des petites applications intégrés à la page Web ou encore comme langage serveur (jsp).

V. NetBeans

NetBeansest un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Devlopment and Distribution License).en plus de JAVA, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).Conçu en java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et open VMS.

En 1977, NetBeans nait d'Elfi, un projet d'étudiants dirigé par la Faculté de mathématiques et de physique de 1'Université Charles de Pragues. Plus tard, une société se forme autour du projet et édite des versions commerciales de 1'EDI NetBeans, jusqu'à ce qu'il soit acheté par Sun en 1999.Sun place le projet sous double licence CDDL et GPL v2 en juin de l'année suivante.

VI. Conclusion

Ce chapitre présente l'essentiel de notre travail, commençant par l'introduction, puis l'explication des étapes importantes pour notre application : prétraitement (tokenisation, élimination des majuscule), après la représentation (représentation de document sous forme de vecteur, élimination de mots vides, fréquence des mots, fréquences des mots dans les documents, mapping des mots en sens, sélection des sens, extraction des relations sémantiques), ensuite nous avons illustré les interfaces de l'application.

Dans ce chapitre nous avons présenté la requête en commençant par « le besoin de l'utilisateur »,puis la réponse de système, ensuite nous avons présenté l'affichage des documents trié selon la distance, on outre nous avons illustré l'affichage du contenu des documents et enfin nous avons défini le programme avec lequel nous avons travaillé.

Conclusion général

La recherche sémantique a pour objectif d'améliorer la précision de recherche par la compréhension de l'objectif de recherche et la signification contextuelle des termes tels qu'ils apparaissent dans l'espace de données recherché, afin de générer des résultats plus pertinents.

Dans ce projet nous avons présenté les principales notions de la recherche d'information, et ceux des outils de recherche sur le web.

A travers les différentes sections que nous avons présentées, nous ne 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. 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.

L'utilisateur utilise le moteur de recherche comme outil de navigation pour trouver le document ciblé. La Recherche sémantique n'est pas applicable aux recherches par navigation. Dans la recherche sémantique, l'utilisateur fournit au moteur de recherche une phrase qui est destinée à désigner un objet sur lequel l'utilisateur tente de recueillir de l'information et de recherche. Il n'y a pas de document particulier que l'utilisateur connait à ce sujet. Au contraire, l'utilisateur tente de localiser un certain nombre de documents qui, ensemble, lui donner les informations qu'il essaie de trouver

L'ensemble de chapitres composant ce mémoire sont organisés en trois grandes parties : La première partie est la recherche d'information ou nous avons présenté les étapes et les modèles de la RI en outre les mesures et les compagnes d'évaluations et par la suite nous avons exposé les applications de la RI.la deuxième partie est le WordNet ou nous avons présenté les principes et les relation sémantiques de WordNet ,par la suite nous avons décrit quelques domaines statiques et quelques domaines d'applications, en fin nous avons terminé par les avantages et les inconvénients de WordNet. Le troisième partie présente l'essentiel de notre travaille, et bien sur les étapes importante pour notre application, et enfin l'exposition de l'environnement de développement intégré (NetBeans).

Résumé

Dans le but de présenté notre projet de licence nous avons réalisé un moteur de recherche sémantique afin de trouver les documents pertinent pour l'utilisateur pour cella on a utilisé la base de donnée lexical WordNet

Comme principe afin d'apporter les sacs de mots.

L'implémentation et la conception sont faites à l'aide du langage java en utilisant l'IDE NetBeans6.8.

Mots-clés : JAVA ,WordNet

Abstract

In order to present our project of licence we conducted a semantic search engine to find the relevant documents for the user that's why we used the lexical database WordNet as a principle to provide bags of words.

The implementation and design are made with the java language using the IDE NetBeans6.8.

Keyword: JAVA,WordNet

ملخص

في إطار تقديم مشروع نهاية الدراسة أجرينا محراكا للبحث الدلالي للعثور على وثائق ذات صلة للمستخدم لذلك قمنا باستخدام قاعدة البيانات المعجمةWordNet كمبدأ لتوفير أكياس من الكلمات. تم التصميم و التنفيذ عن طريق لغة Java وباستخدام NetBeans