Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

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 d'ingénieur en Informatique

Thème

Recherche intelligente des informations dans le coran

Réalisé par :

LAMRAOUI Younes

Encadreur :

ABDERRAHIM Mohamed

Présenté le 04 Juil. .2011 devant la commission d'examination composée de MM.

BENAISSA MOHAMMED	*(président)*
BENMOUNA YOUSSEF	*(Examinateur)*
BENMANSOUR Fazilet	*(Examinateur)*
ABDERRAHIM ALAEDDINE	*(Co-encadreur)*

Année universitaire : 2010-2011

Remerciement

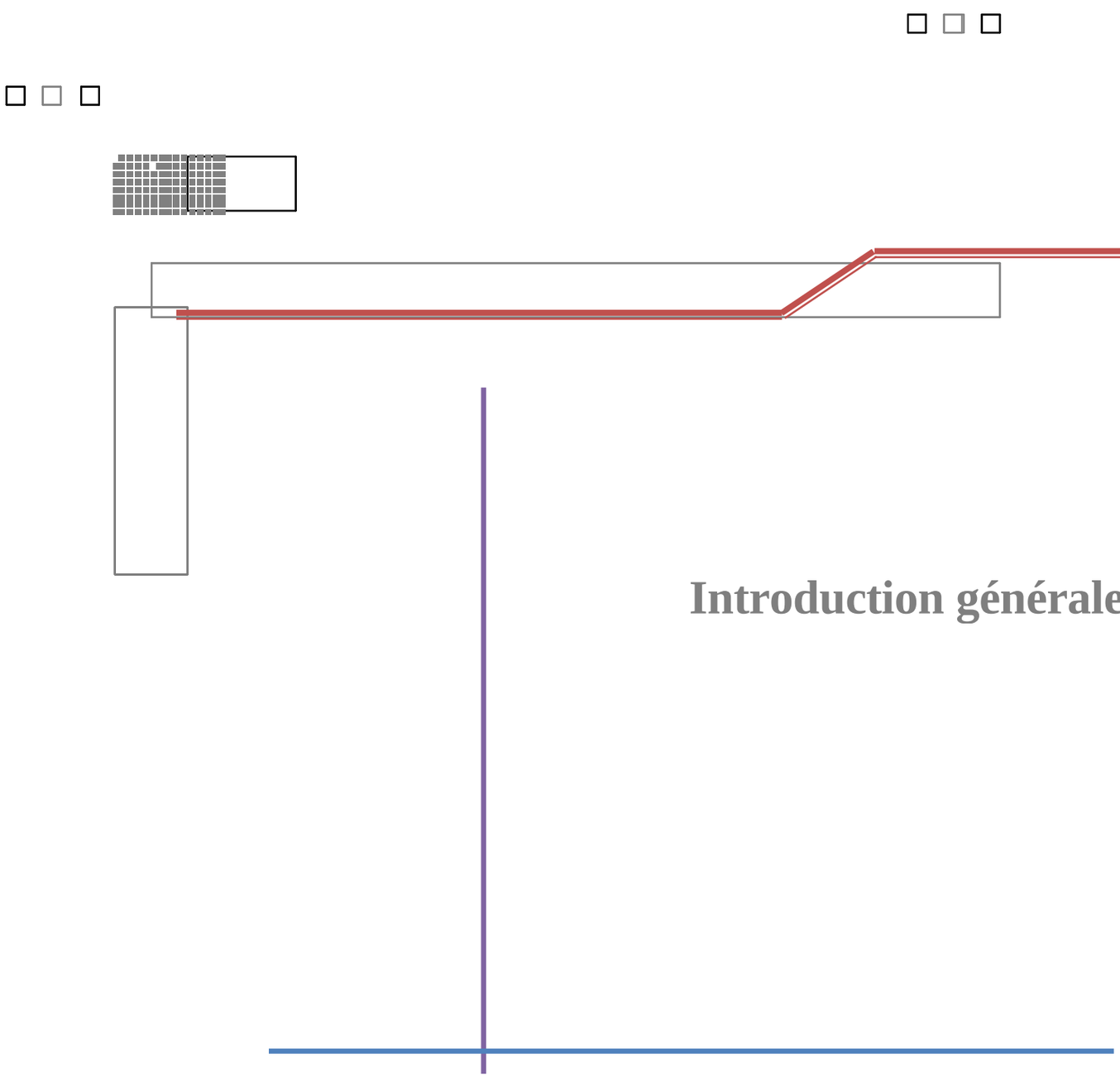
Avec beaucoup de gratitude et de sincérité, je remercie vivement mon encadreur Mr Abderrahim Mohammed El Amine, Maître de Conférences au département d'Informatique université de Tlemcen, pour sa présence scientifique et humaine, et l'honneur qu'il m'a fait en acceptant de m'encadrer.

Toute ma reconnaissance va à mon Co-encadreur Mr Abderrahim alaeddine, pour avoir accepté de diriger ce projet. Son soutien moral, ses orientations, ses précieux conseils, m'ont accompagné tout au long de ce projet. Je le remercie aussi pour ses qualités humaines. Je remercie également Mr BENAISA Mohamed, Professeur au département d'Informatique de l'université de Tlemcen, d'abord pour la formation et pour l'honneur qu'il me fait de présider le jury. Qu'il reçoit l'expression de mon profond respect.

Ma profonde reconnaissance va également à Mr BENMANSOUR, et à BENMOUNA pour avoir accepté d'examiner mon travail. Je n'oublie pas d'adresser mes vifs remerciements à toute ma promotion pour les sens de fraternité, et de concurrence qui nous ont tenus debout pendant toute la formation.et mes chers amis Mr ZAZOUA K. Mohammed ,BOUDAOUD Faycal BENDAOUD Amine et Zaki.

Sommaire

Liste des figures



Introduction générale

Introduction générale

Les Systèmes de Recherche d'Information (SRI), sont conçus à l'origine, pour répondre aux besoins d'automatiser la gestion de l'information. Ces informations ont besoin d'être stockées, organisées et indexées afin de permettre à des utilisateurs en quête d'une information de répondre à leurs besoins. De ce fait, des approches de recherche d'information, regroupant entre autres des techniques d'indexation ainsi que des mécanismes d'appariement et de reformulation ont été développées afin de mieux répondre aux besoins de l'utilisateur.

Les premières approches de recherche d'information, qualifiées de classiques, se basent sur une recherche par mots clés, les documents sont représentés comme des sacs de mots souvent pondérés, et la pertinence d'un document vis-à-vis d'une requête est souvent estimée en S'appuyant sur les fréquences d'apparition des mots de la requête dans ces mêmes documents.

Les SRI sont alors confrontés à un nouveau défi du à ces limites. Ce qui a poussé des chercheurs à marquer un arrêt pour explorer d'autres terrains, notamment celui de la linguistique et de l'intelligence artificielle, pour proposer des améliorations à la solution, qui est généralement préconisée aujourd'hui.

Dans ce contexte, l'utilisation des ontologies aux seins des SRI arabe peut être une solution pour remédier de façon efficace à ces problèmes. D'une part, les ontologies fournissent les ressources généralement sous forme de relations sémantiques permettant un traitement pour "élargir" le champ de recherche pour les requêtes : les connaissances ontologiques permettent de représenter le sens de la requête et d'effectuer des inférences sur les informations décrivant le contenu des ressources (les méta-données), contribuant ainsi à améliorer la qualité de la recherche. D'autre part, elles constituent le cadre partagé (le même vocabulaire) que les différents acteurs peuvent mobiliser [Bachimont, 2000].

Problématique :

La recherche d'information est un processus qui se base essentiellement sur la requête exprimé par l'utilisateur pour répondre à ses besoins. En effet, le résultat d'une recherche ne peut-être pertinent si la requête ne d écrit pas explicitement et clairement les besoins de l'utilisateur ainsi que la bonne représentation des informations apparus dans la base de collection des documents. Cela, issu au mal compréhension du domaine recherché ou à une limitation des connaissances d'utilisateur.

La reformulation de requêtes est une des stratégies qui permet d'améliorer la construction d'une requête. Elle consiste de manière générale à enrichir la requête de l'utilisateur en ajoutant des termes permettant de mieux exprimer son besoin. Une de ces techniques est la reformulation par l'utilisation d'une ontologie. Elle consiste à ajouter des termes proches aux termes de la requête extraite à partir d'une ontologie.

Les travaux décrits dans ce mémoire s'intéressent à la reformulation de requêtes ainsi que l'indexation conceptuelle du coran par l'utilisation d'une ressource externe qui est l'ontologie Coranique.

Objectifs :

Dans le cadre de ce mémoire, on se propose d'apporter une solution aux problèmes de reformulation de pour améliorer les résultats de la recherche d'un SRI.

Notre travail se décompose en plusieurs palais :

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

- Deuxièmement, nous proposons un modèle de recherche en intégrant l'ontologie Coranique au niveau de la requête et de l'indexation afin de prendre en compte le niveau sémantique dans un SRI.

Organisation du mémoire :

Ce mémoire est composé de trois chapitres :

- Le premier chapitre présente les notions de base et les principaux concepts utilisés dans le domaine de la recherche d'information. Il présente l'architecture générale d'un système de recherche d'information (SRI) telle

qu'elle est admise actuellement ainsi qu'un aperçu sur les principaux modèles de recherche existants dans la littérature, il comporte aussi les différentes mesures d'évaluation de pertinence.

- Le deuxième chapitre dresse un état de l'art sur le concept d'ontologie :
- Les circonstances de leur apparition, différentes définitions, les principales ontologies implémentées, un bref aperçu sur les langages de spécification d'ontologie et leurs domaines d'application. Il se termine par une présentation du comment de l'utilisation des ontologies dans le domaine de la recherche d'information.

- Le troisième chapitre est le cœur de notre mémoire. Il concerne l'application du concept d'ontologie d'informations ainsi que leur mode de fonctionnement. Puis, étudier les effets des ontologies aux différents niveaux des SRIs. à la recherche d'information au coran. Nous présentons une implémentation ainsi que l'ontologie et les autres outils utilisés. Différents résultats correspondants à différentes expérimentations sont ensuite rapportés et commentés, pour enfin terminer par une conclusion sur l'apport des ontologies dans la reformulation de requête et les perspectives de leur utilisation pour l'ensemble de la recherche d'information.

Dans la conclusion, nous présenterons les principaux points abordés dans ce mémoire et nos dégageons quelques piste pour la

Les Systèmes de Recherche d'Information

Plan

1. Introduction

La recherche sémantique de l'information est une des principales motivations du Web sémantique Un moteur de recherche sémantique peut être vu comme un outil qui répond à des requêtes – formulées avec les concepts et les relations d'une ontologie de domaine – en les alignant avec des annotations sémantiques des documents cibles.

Dans une vue idéale, ce problème de Recherche d'Information (RI) peut être considéré comme similaire à la RI dans les bases de données relationnelles où les réponses sont des ensembles de tuples satisfaisant la requête de l'utilisateur. Cependant, cette vue ne se concrétise que si les contenus de tous les documents peuvent être représentés par des instances de concepts ou de relations définis dans une ontologie donnée.

Les avancées de la recherche visant à automatiser le peuplement des ontologies et l'annotation des documents sont prometteuses (Popov et al., 2004, Cimiano et al.,2005, Etzioni et al., 2005, Thiam et al., 2009). Cependant, la localisation précise de toutes les instances dans un document reste une tâche difficile. Une certaine imprécision sémantique peut se produire du fait que les métadonnées choisies ne sont pas parfaitement appropriées.

Dans ce chapitre nous allons passer en revue les concepts, les approches et les modèles utilisés dans le domaine des SRI.

2. Concepts clés de la recherche d'information (RI)

La majorité des approches proposées dans la recherche des documents (documents XML) reposent sur des systèmes d'indexation à base clés ou encore sur les termes. Les seules informations utilisées concernant termes sont leurs fréquences d'apparition dans les documents, ou dans les éléments du document (en fonction du niveau de granularité). Ainsi, ces approches ne prennent pas en considération le sens du mot (terme).

L'indexation par des mots clés est généralement imprécise .Cette imprécision est due au fait que les Termes d'indexation présentent une forte ambigüité.

En effet, le sens d'un mot clé peut varier selon le contexte dans lequel il apparait (phénomène de polysémie) [Abderahim]

Aussi, ces approches ne prennent pas en compte la synonymie. Par conséquent, dans ces systèmes,

il est impossible de trouver des parties des documents représentées par un mot M_1 synonyme d'un mot

M_2 représente une requête. Par conséquent, il se peut qu'un système de RI basé sur les mots ne renvoie pas un élément pertinent, c'est-à-dire un élément qui satisfait la requête

Un moyen pour améliorer les performances des systèmes de RI sur les documents est la prise en compte de la sémantique des termes d'indexation.

Ce type d'indexation passe du niveau des mots au niveau des concepts (les sens des mots) pour mieux décrire le contenu du document et de la requête. Ces approches utilisent des ressources sémantiques (thésaurus, ontologies, etc.) dans les phases d'indexation et de recherche.

Plusieurs concepts clés s'articulent autour de cette définition:

Document : le document représente l'unité élémentaire sélectionnée comme réponse d'une requête. Cette information élémentaire, appelée aussi granule de document, peut représenter tout le document ou une partie de lui. Dans le reste de ce rapport, nous utiliserons indifféremment les termes : document ou information pour désigner un granule documentaire

Requête : la requête constitue l'expression du besoin en information de l'utilisateur.

Elle représente l'interface entre le SR I et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature. Une requête peut être écrite à l'aide d'une liste de mots clés.

Thesaurus : Un thesaurus est un ensemble de termes formels, auxquels peuvent être associées des définitions permettant de représenter des connaissances. Ces définitions permettent de poser des contraintes sur l'utilisation des termes. Elles permettent ainsi d'effectuer des vérifications syntaxiques ou sémantiques.

Autrement dit, elles permettent de préciser le contexte d'utilisation du terme, de guider l'association de certains termes avec d'autres termes et d'indiquer également des relations possibles en termes.

2.1. Langage naturel ou quasi naturel

L'utilisateur exprime sa requête en langage libre (langage naturel) sous forme de mots clés. Le Système se charge de traduire (analyser) ces mots clés en une requête de langage de base de Donnée sou une autre forme interne utilisable par le système, cas des systèmes

SMART1 [Salton, 1971], OKAPI [Robertson & al., 99] .

2.2. Langage booléen

Appelé aussi langage de format structuré, ce langage utilise les opérateurs booléen pour formulé la requête (ET, OU, NON), cas des systèmes DIALOG [Bourne & Anderson, 79].

2.3. Langage graphique

L'idée est de concevoir une interface qui peut aider l'utilisateur à formuler sa requête. Cette interface va proposer d es termes représentant le contenu sémantique des mots pour faciliter le choix des termes lors de la construction de la requête. Ce dernier est représenté par un graphe, les nœuds étant les termes du thesaurus et les liens étant les relations sémantiques entre ces termes. L'utilisateur peut identifier le type de relation qu'il souhaite utiliser et sélectionne un terme.

Le projet NEURODOC2 [Lelu & al., 92] **.** est plus adapté à l'utilisation d'un thésaurus volumineux. NEURODOC offre à l'utilisateur un tableau de bord où chaque nœud possède un nom et résume le sous-ensemble de mots et de documents fortement liés. Besoin d'information : On peut définir différents types de besoin d'information correspondant a :

1. **Le premier besoin apparait quand l'utilisateur sait parfaitement ce qu'il cherche et où chercher. on l'appel aussi besoin vérificatif.**
2. **Le deuxième besoin naît quand l'utilisateur sait seulement le domaine de l'information qu'il veut (sujet).**
3. **Le troisième besoin naît quand l'utilisateur ne sait rien sur le domaine de recherche et il veut explorer des nouveaux concepts pour ce domaine.**

3. Architecture générale des systèmes de recherche d'information

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

La partie adroite désigne l'information accessible par le système tel que les collections des documents qui regroupe les documents de même domaine ou de domaine proche. Dans la partie à gauche nous trouvons le besoin en information de l'utilisateur : exprimé par une requête. Enfin, pour combiné les deux cotés ; le système propose un certain nombre de traitement pour trouver les informations pertinentes.

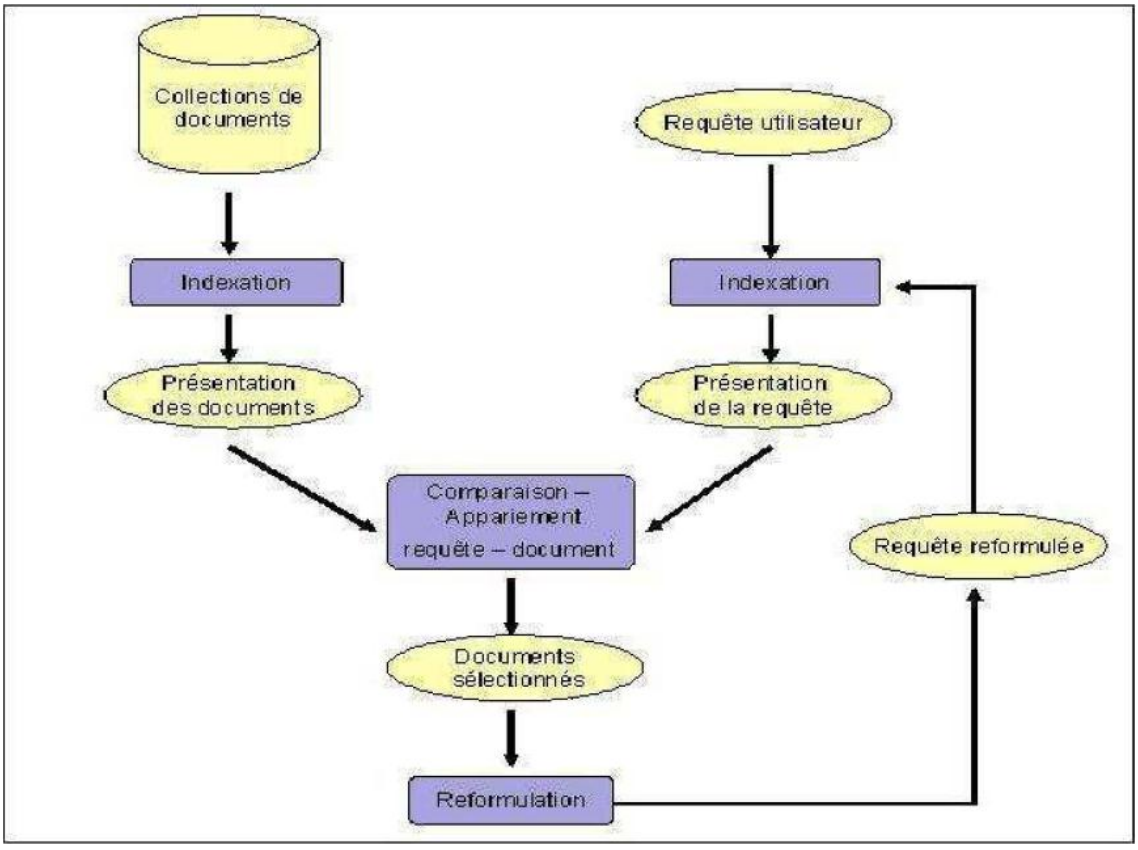


Figure 1.1Le Processus en U de la Recherche d'Information [Hlaoua, 07]

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

3.1. Processus de représentation (indexation)

Cette étape peut être considérée comme une préparation à la recherche. Ce processus a pour rôle l'extraction des bons termes donnant une représentation

détaillée (appelé aussi représentation paramétrique) du contenu sémantique d'une requête ou document. C'est une étape fondamentale dans la conception d'un système de recherche d'information, ce processus permet de passer d'une description brute d' un document d'un concept ou d'une requête vers une description structurée. Ce mécanisme est appelé indexation. Sa qualité dépend en partie de la qualité des réponses du système.

Ce mécanisme est une étape primordiale dans un système de recherche d'information, le résultat obtenu est un descripteur de document ou de requête, qui est une liste de termes significatifs pondérés par des poids donnant l'importance à ces derniers. L'ensemble des index rassemblés dans un dictionnaire, ce dernier est le langage d'indexation qui se devise en deux catégories :

• 1ére catégorie : est le langage contrôlé qui définit un lexique de descripteurs bien spécifié par des experts. L'indexation est alors dirigée le plus souvent de façon manuelle ou parfois semi-automatique. Un risque de confusion a resurgi lorsque les termes des utilisateurs et le vocabulaire des experts s'oppose.

• 2éme catégorie : est le langage libre, cette fois l'extraction des index se fait de manière automatique extrait d'une manière automatique, dans ce cas le taux d'indexation est très élevé et le risque qui apparait maintenant réside dans les descripteurs non significatifs.

L'opération d'indexation peut se dérouler en trois modes différents [Kompaoré, 08] [Tamine, 00] [Nassr, 02] **.**

1. Indexation manuelle (L'indexation humaine)

Ce genre d'indexation est guidé par un spécialiste du domaine. Même s'il existe une différence entre deux spécialistes ou un spécialiste lui même dans le choix de l'index, cette méthode a un résultat toujours fiable. Donc ce type permet d'assurer une meilleure correspondance entre les documents et les descripteurs choisis par les indexeurs. En effet, les spécialistes d'un domaine choisissent les meilleurs termes pour indexer les documents.

L'indexation humaine est une activité fondée sur le jugement de l'être humain. Enfin, il faut noter que l'indexation manuelle est couteuse en temps.

2. Indexation semi-automatique

Cette fois ci le processus d'indexation se déroule en deux phases. Dans la première phase, les index sont extraits automatiquement et sont ensuite transmis aux spécialistes du domaine pour les valider à l'aide d'un thesaurus ou une base terminologique. Cette méthode est appelé aussi l'indexation supervisée.

3. Indexation automatique

L'indexation automatique comporte seulement les procédures automatiques sans avoir l'intervention de l'homme. Cette méthode d'indexation est actuellement la méthode la plus répandue, elle passe par deux étapes : la détermination des index et la pondération de ces derniers.

• Premièrement, l'extraction des index passe par un anti-dictionnaire pour supprimer les mots de liaisons et les mots vides ainsi que la suppression des variations des mots. Toutes ces opérations sont regroupées dans le pilisyntaxique

• Deuxièmement, la détermination des poids va différencier entre les termes dans le document ou la requête.

L'indexation automatique peut se faire selon deux approches : statistique et linguistique.

L'approche statistique se base sur la distribution statistique des termes dans le document.

L'approche linguistique se base sur les techniques de traitement du langage naturel, telles que l'analyse lexicale, syntaxique et sémantique.

l'analyse lexicale, syntaxique et sémantique.

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

Analyse lexicale : L'analyse lexicale (tokenization en anglais) est le processus qui permet de convertir le texte d'un document en un ensemble de termes. Un terme est un groupe de caractères constituant un mot significatif. L'analyse lexicale permet de reconnaitre les espaces de séparation des mots, des chiffres, les ponctuations, etc.

L'élimination des mots vides : Un des problèmes majeurs de l'indexation consiste à extraire les termes significatifs des mots vides (pronoms personnels, prépositions, ...). Les mots vides peuvent aussi être des mots athématiques : les mots qui peuvent se retrouver dans n'importe quel document parce qu' ils exposent le sujet mais ne le traitent pas, comme par exemple

(contenir, appartenir). On distingue deux techniques pour éliminer les mots vides :

L'utilisation d'une liste de mots vides (appelée aussi anti-dictionnaire, stoplist en anglais),
L'élimination des mots dépassant un certain nombre d'occurrences dans le document.

Lemmatisation : Un mot donné peut avoir différentes formes dans un texte, par exemple
أسلم, إسلام, إسلاميون

. Pour indexer ces différentes variations des mots on utilise la méthode de racinisation.

3.2. Pondération des termes

La pondération permet d'attribué un poids au terme d'indexation qui représente l'importance de cet index dans le document respectivement dans la requête et de réduire la taille de l'ensemble des descripteurs de document et des requêtes (nombre d'index). La plupart des techniques de pondération sont basées sur les facteurs TF et IDF [Kompaoré, 08] [Karbasi, 07] **.** :

TF (Term Frequency) : mesure l'importance d'un terme dans un document. Cette mesure est souvent en fonction de la fréquence d'un terme d ans un document ou une requête. Elle peut être utilisée telle quelle ou selon plusieurs déclinaisons (log(TF), présence/absence,...).

IDF (Inverse of Document Frequency) : ce facteur mesure l'importance d'un terme dans toute la collection. Un terme qui apparaitre souvent dans la base documentaire ne doit pas avoir le même impact qu'un terme moins fréquent

La mesure TF * IDF est une bonne approximation de l'importance d'un terme dans un document, elle l'est particulièrement dans les corpus de documents de tailles homogènes, tels que les corpus contenant des résumés. Cette mesure a eu un succès limité dans les corpus de tailles très variables.

3.3. Reformulation de Requêtes

De façon générale les utilisateurs qui font la recherche ne maitrisent pas le stade du domaine et tous les termes accessibles par ce dernier. Le processus de reformulation de requêtes est utilisé lorsqu'un utilisateur est incapable de reformuler sa requête du début afin de donner une information pertinente.

Le principe de reformulation est basé sur l'ajout des termes à la requête initiale ou l'ajustement des poids des index. Nous distinguons principalement deux approches pour la reformulation [Baziz, 02] [Baziz, 05] .:

Reformulation directe : elle consiste à ajouter de nouveaux termes à la requête initiale. Cette modification est réalisée grâce aux liens de cooccurrence entre les termes On parle alors de reformulation de requêtes basée sur les concepts (Concept-based Query Reformulation).

Reformulation indirecte : dans cette approche la requête est modifiée en tenant compte d'une liste de documents déjà jugés sélectionnés. Ce processus est appelé réinjection de la pertinence (relevance feed-back) si le processus est supervisé et de pseudo réinjection de pertinence si le processus est automatique. Cette méthode a un double avantage : une simplicité d'exécution pour l'utilisateur qui ne s'occupe pas desdétails de la reformulation, et un meilleur contrôle du processus de recherche en augmentant le poids des termes importants et en diminuant celui des termes non importants.

Dans notre mémoire on essaye de reformuler la requête automatiquement de manière sémantique par la joindre avec une ontologie coranique.

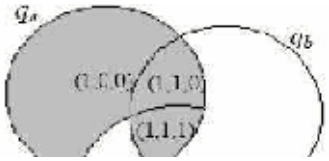
Les modèles booléens se basent sur la théorie des ensembles. En général, la requête est exprimée par une liste de termes et des opérateurs logiques : conjonction (ET), disjonction (OU) et négation (NON). A chaque terme est associé un ensemble de mots ou il apparaît. L'opérateur "ET" restreint le résultat de la requête à l'inter section entre deux ensembles, l'opérateur "OU" fournit l'union et l'opérateur "NON" la différence entre les ensembles. D'autres modèles sont dérivés du modèle booléen, tels que le modèle booléen étendu et le modèle basé sur les ensembles flous.

Finalement, la longueur et le choix des termes de la requête jouent un rôle très important dans le processus de reformulation de la requête.

4. Les modèles de recherche d'information :

L'un des rôles d'un système de recherche d'information est de mesurer la pertinence mot par rapport à une requête. Un modèle de RI fournit une formalisation au processus de recherche d'information. Il doit accomplir plusieurs rôles dont le plus important est fournir un cadre théorique pour la modélisation de cette mesure de pertinence.

Les modèles vectoriels reposent sur la théorie algébrique. La pertinence d'un mot vis-à -

vis d'une requête  est définie par des mesures de distance (ou similarité) dans un espace vectoriel. Plusieurs modèles s'inspirant du modèle vectoriel ont été proposés dans le domaine de la RI : le modèle vectoriel généralisé, le modèle connexionniste et le modèle LSI (Latent Semantic Indexing). Enfin, les modèles probabilistes se basent sur la théorie des probabilités.

La pertinence d'un mot vis-à-vis d'une requête est vue comme une probabilité de pertinence mot/requête. On distingue le modèle BIR (Binary Independence Retrieval), le modèle inférentiel bayésien et le modèle de langage.

Dans ce qui suit, nous détaillons les modèles cités ci-dessus, et quelques modèles dérivés ou inspirés à partir de ces classes.

4.1. Modèles booléens

Différents modèles [Kompaoré, 08][Baziz, 05][Tebri, 04][Nassr, 02][Tamine, 00] de recherche d'informations ont été proposés. Le présent paragraphe a pour objectif d'en présenter les principaux.

4.1.1. Le modèle booléen ou ensembliste (Boolean Model)

Le modèle booléen est le premier modèle inventé par salton [Salton, 71] ., il propose une représentation de la requête sous forme d'une expression logique, tel que les termes de requête reliés par des connecteurs lo gique comme le : «OU», «ET», «NON», etc. Ce modèle se base sur la théorie des ensembles et l'algèbre de boole.

Une requête booléenne représente une définition exacte d'un ensemble de mots. Par exemple : la requête ' مسجد' définit tout simplement le concept indexés avec le terme 'مسجد '. En utilisant les opérateurs de Boole les requêtes et les ensembles de mots correspondants peuvent être combinés pour former de nouveaux ensembles de mots.

La correspondance RSV (Dj, Qk), entre une requête Qk et un document Dj est déterminée comme suit :

RSV (Dj , qi)=1 si qi Dj ; 0 sinon,
RSV (Dj, qi qj)=1 si RSV (Dj, qi)=1 et RSV (Dj , qj)=1 ; 0 sinon,
RSV (Dj, qi qj)=1 si RSV (Dj, qi)=1 ou RSV (Dj, qj)=1 ; 0 sinon,
RSV (Dj, ¬ qi)=1 si RSV (Dj, qi)=0 ; 0 sinon,

Sachant que qi est un terme de la requête Qk ,

Le modèle de recherche Booléen est reconnu par sa force de faire une recherche très restrictive et l'obtention d'une information exacte et spécifique pour un utilisateur expérimenté.

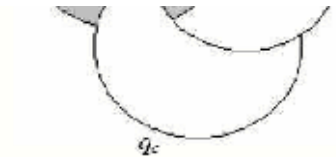


Figure 1-2 : Les 3 composants conjonctifs pour la requête : q = qa (qb ¬qc) [Baziz, 02]

Le modèle booléen présente le principal avantage de simplicité de mise en œuvre. Toutefois, il présente les principaux inconvénients suivants :

Les mots pertinents dont la représentation (Figure 1-2) ne sont pas sélectionnés par la Requête formulée.

- les formules de requêtes sont complexes, non accessibles à un large public,
- la réponse du système dépend de l'ordre de traitement des opérateurs de la requête,
- ce modèle est incapable de trier les documents pertinents,
- les modèles de représentation des requêtes et document ne sont pas uniformes,
- les termes ont la même importance.

Ceci rend le modèle inadapté à une recherche progressive.

4.1.2. Modèle booléen étendu

Le Modèle booléen étendu est appelé aussi modèle P_Norm (tel que l'opérateur Lp-Norm est défini pour la mesure de pertinence requête-document), il a été introduit en 1983 par Salton et al. [Salton & al., 83] **.**. C'est une alternative pour rendre le Modèle booléen plus intéressant et de l'étendre afin de supporter un appariement approché (fonction d'ordre et non exact) en assignant des poids aux termes de la requête et des documents et en mesurant un score de pertinence. Le principe de base du Modèle booléen étendu est de prendre en considération l'importance des termes de recherche par la distribution des poidsaux index et d'interpréter les opérateurs de l'équation de la requête comme des distances entre requêtes et document. Losee [Losee, 98] **.** a montré en utilisant le tri que le modèle Booléen étendu est un cas particulier de la recherche probabiliste.

4.1.3. Le modèle des ensembles flous

Dès l'invention des ensembles flous au milieu des années 60 par lotfi Zadeh leurs utilisations couvrent presque tous les domaines. Ils sont basés sur l'appartenance probable et non certaine d'un élément à un ensemble. Dans la recherche d'information le modèle flou est apparu pour modéliser les notions devague et d'imprécision qui existent à différents niveaux du processus de RI [Bordogna & al., 00][Koczy & al., 98]. et surtout pour réduire l'incomplétude et traiter l'imprécision dans les processus d'indexation et de recherche.

Salton [Salton, 89]. a dit que ce modèle est une autre extension du modèle booléen. L'idée de base est de traiter les descripteurs des documents et requêtes comme étant des ensembles flous. Cette extension vise également à tenir compte de la pondération des termes dans les documents. Un poids d'u n terme exprime son degré d'appartenance à un en semble.

FIRST est parmi les premiers systèmes de recherche d'information basé sur le modèle des ensembles flous crée par les chercheurs Lucarella & Morara [Lucarella & Morara, 91]. Les auteurs ont proposé l'utilisation d'un réseau où chaque nœud représente un terme de document ou requête et un lien représente la relation sémantique entre les termes. Chaque document Dj est décrit par un ensemble flou comme suit :

$$Dj = \{(t1, dj1), \ldots., (tT,djT)\}$$

Une liaison entre concepts est valorisée de manière directe, ou dérivée par transitivité floue :

$$F(ti , tk) = Min (F(ti, tj) , F(tj , tk))$$

Où F : Fonction de valorisation des liens

L'ensemble flou des documents pertinents à une requête Qk est obtenu co mme suit :

1. Pour chaque terme t de Qk , construire l'ensemble des documents Dt reliés par lien direct ou transitif
2. Pour chaque couple (t, Dt), associer un degré d'appartenance égal à la valeur minimale de tous les liens qui figurent sur le chemin t – Dt.
3. Effectuer sur les ensembles Dt, les opérations d'intersection et union selon l'ordre décrit dans l'expression de Qk relativement aux opérateurs ET et OU respectivement.

4. Ordonner l'ensemble résultat de la précédente opération selon le degré d'appartenance de chaque document à l'ensemble associé à chaque terme.

Chen & Wang [Chen & Wang, 95] . ont étendu ce modèle à l'utilisation d'intervalles de poids

Où :

$$T(x,y)=1-|x- y|$$

tji : Minimum des poids des liens du document Dj au terme ti.

admissibles aux concepts par opposition à l'utilisation de valeurs uniques ainsi qu'à l'utilisation d'une matrice de concepts. La clôture transitive de cette matrice, soit T, est obtenue par multiplications successives de cette même matrice. La valeur de pertinence requête-document est obtenue selon la formule suivante :

Les objectifs pour lesquels les modèles de recherche d'information intègrent les ensembles flous sont :

- réduire l'imperfection et traiter l'imprécision qui caractérise le processus d'indexation,
- contrôler l'imprécision de l'utilisateur dans sa requête,
- traiter les réponses reflétant la pertinence partielle des documents par rapport aux requêtes.

L'inconvénient principal de ce modèle est que le calcul de la valeur de similarité est toujours dominé par les petits poids des termes dans le cas des conjonctions et les grands poids des termes dans le cas des disjonctions. Autrement, ces modèles ne sont pas adaptés au classement (ranking) des documents pertinents, étant donné que les scores de pertinence qu'ils attribuent aux documents sont calculés par des fonctions min ou max qui ne prennent pas nécessairement en compte toutes les valeurs de pertinences des termes de la requête.

Cependant des extensions ont été proposées à ces modèles [Bordogna & al., 00][Loiseau & al, 04] . Notamment pour améliorer l'ordonnancement (ranking) des documents sélectionnés.

4.2. Les modèles vectoriels

4.2.1. Le modèle de base

La première idée de représenter les document et les requêtes sous forme de vecteurs de termes pondérés a été proposée par Luhn [Luhn, 57]. à la fin des années cinquante. Elle a été ensuite développée par Gérard Salton et son équipe [Salton, 71] [Salton, 83]. dans leur projet SMART. L'idée de base du modèle vectoriel est d'utiliser une représentation géométrique pour classerles document par ordre de pertinence par rapport à une requête.

Dans ce modèle le document et les requêtes sont engendrés par les termes d'indexation représentés par des vecteurs [Salton, 83] .. L'espace est de dimension N (N étant le nombre de termes d'indexation de document).

La pertinence d'un document relative à une requête est estimée par la position de ces deux vecteurs dans un espace vectorielle. Autrement, Le degré de pertinence d'un document relativement à une requête est perçu comme le degré de corrélation entre les vecteurs associés. Tel que, on va estimer la coïncidence entre les termes pondérés de la requête et le document. Cette position est mesurée par une distance ou une similarité définie sur cet espace vectoriel.

Les avantages du modèle vectoriel sont liés : d'une part au modèle de pondération qu'il utilise et qui permet d'améliorer les performances de la recherche. D'autre part, le modèle vectoriel permet de retrouver des mots qui répondent partiellement à la requête de l'utilisateur et permet aussi d'établir un classement de mots trouvés par ordre décroisant en fonction de leur degré de similarité avec la requête correspondante qui est un des inconvénients majeur du modèle booléen.

Le modèle vectoriel est similaire au modèle probabiliste dans différents aspects, sauf qu'il manque d'une base théorique saine [Croft & al., 92] . De plus, il trie le document suivant une mesure de similarité avec la requête au lieu d'une probabilité de pertinence des mots par rapport au besoin en information des utilisateurs comme c'est le cas dans le modèle probabiliste. Même si le modèle vectoriel est critiqué comme étant un modèle "ad hoc", il est l'un des modèles de RI classique les plus influents, les plus étudiés et les mieux acceptés.

L'inconvénient de ce modèle est le traitement des termes d'indexation de document de manière disjonctifs.

4.2.2. Le modèle vectoriel généralisé

Afin de résoudre le problème de l'indépendance des termes d'indexation posé par le modèle précédent, Wong [Wong & al, 85] . à proposé une nouvelle représentation de document et requêtes. Dans ce modèle chaque terme est représenté par un vecteur dans un espace vectoriel dont les axes sont

orthogonaux par construction : les axes sont en fait les produits logiques des termes d'indexation. Un document est représenté par la moyenne des vecteurs représentant les termes qu'il contient.

Lors du calcul de pertinence ce modèle combine entre les poids de document et le facteur de corrélation entre les vecteurs.

4.2.3. Le modèle connexionniste

Les systèmes de recherche d'informations basés sur le modèle connexionniste sont fondés sur le réseau de neurones [Kwork, 89], [Boughanem, 92] [Mothe, 94] .Ce réseau est construit à partir des représentations des contenues des requêtes et de document sous forme de couches répartis généralement dans ce sens : requête - termes - document [Kwork, 95]. Ce modèle peut être vu comme un modèle vectoriel récurrent non linéaire. Les neurones formels représentent des objets de la recherche d'information.

Le mécanisme de recherche d'information basé sur ce modèle est fondé sur le principe d'apprentissage ce qui permet aux SRI de devenir adaptatifs. En effet la requête active la couche des termes et cette activation est propagée vers la couche de document à travers les connexions du réseau. Les résultats sont présentés à l'utilisateur selon le niveau d'activation des neurones document. La pertinence de document est alors mesurée grâce à leur niveau d'activation.

Les systèmes de recherche d'information utilisant l'approche connexionniste peuvent être divisés en deux catégories :

1. **Les modèles à auto-organisation**

Dans ce modèle le document subissent un classement à partir de leurs descriptions initiales effectué à l'aide d'algorithmes propres aux modèles connexionnistes. Des modèles d'auto-organisation inspirés par l'organisation corticale des vertébrés ont été proposés dés les années 80, notamment par Kohonen [Kohonen, 89] et Lelu & François [Lelu & François, 92] .

Se basent sur les réseaux à couches [Boughanem, 92] , [Boughanem, 00],[Kwok, 89], [Mothe, 94] . Ils comprennent une couche de termes qui sera activée par une requête initiale d'un utilisateur et une couche document. Les deux couches sont liées par des connexions pondérées. La requête active la couche des termes et cette activation est propagée vers la couche de document. L'activation finale de document donne la réponse du moteur de recherche.

Actuellement, plusieurs modèles basés sur le principe des réseaux de neurones sont

utilisés en recherche d'information [Boughanem & Tamine, 04] .

L'avantage principal de l'utilisation des réseaux de neurones est cette propriété d'apprentissage.

4.3. Modèles probabilistes

Le modèle de recherche probabiliste utilise un modèle mathématique fondé sur la théorie de la probabilité [Salton, 83] . Ce modèle est conçu au début des années 60 par Maron et Kuhns [Maron & Kuhns, 60] . L'idée générale de cette approche est d'implémenter les notions de la théorie de probabilité sur les systèmes de recherche d'information. Le principe de base du modèle probabiliste consiste à présenter les résultats de recherche d'un SRI dans un ordre basé sur la probabilité de pertinence d'un document vis-à-vis d'une requête.

Selon Robertson, deux hypothèses doivent être vérifiées pour garantir l'optimisation d'ordonnancement d'une collection de documents. Premièrement, la pertinence des documents doit être une variable aléatoire binaire prenant l'une des valeurs vrai ou faux et deuxièmement, la pertinence d'un document ne doit pas influencer la pertinence d'un autre.

Le système de recherche d'information utilise une indexation probabiliste basé sur deux types de probabilités conditionnelles :

P (wji / Pert) : Probabilité que le terme ti occure dans le document Dj sachant que ce dernier est pertinent pour la requête.

P (wji/ NonPert) : Probabilité que le terme ti de poids dji occure dans le document Dj sachant que ce dernier n'est pas pertinent pour la requête.

Le modèle probabiliste présente l'intérêt d'unifier les représentations des documents et concepts. Cependant, le modèle repose sur des hypothèses d'indépendance des variables pertinents non toujours vérifiées, ce qui entache les mesures de similitude d'imprécision. En outre, le modèle ne p rend pas en compte les liens de dépendance entre les termes et engendre des calculs de probabilités conditionnelles complexes.

Ce modèle à des avantages et des inconvénients on peut citer :

Avantages : - dépendance entre la représentation des documents et les concepts.

Inconvénients : - les calculs sont complexes,
- l'indépendance des variables n'est pas toujours vérifiée,
- pas de prise en compte de l'indépendance entre les termes d'indexation.

4.3.1. Le modèle BIR (Binary Independence Retrieval)

Le modèle BIR comme dans tous les modèles probabilistes cherche à estimer la

probabilité qu'un document Dj soit pertinent pour une requête Qk. Autrement, ce modèle doit juger la pertinence d'un document par rapport à une requête par le calcule de probabilité qui est basé sur le théorème de Bayes. L'idée de base du modèle BIR est de représenter les termes de document par des valeurs binaires (1 si le terme apparait dans un document, 0 sinon).

4.3.2. Le modèle de réseau infèrentiel bayésien

Un réseau bayésien est un graphe direct acyclique où les nœuds représentent des variables aléatoires et les arcs des relations causales entre nœuds [C allan, 96]**.** Ces derniers sont pondérés par des valeurs de probabilités conditionnelle

Le travail original en recherche d'information basé sur le modèle des réseaux bayésiens est développé par Turtle [Turtle & Crof t, 91] .

Les probabilités conditionnelles de chaque nœud sont calculées par propagation des liens de corrélation entre eux. Ce modèle vise a considérer la dépendance entre les termes mais engendre une complexité de calcul importante.

Dans ce graphe les nœuds représentent des variables propositionnelles ou également des constantes et les arcs représentent des liens de dépendances entre les nœuds. Ainsi, si la proposition représentée par le nœud p cause ou implique la proposition représentée par le nœud q, on trace alors un arc de p vers q.

Dans le domaine de la recherche d'information les nœuds et les arcs sont présentés comme suit :

– Les nœuds : représentent des concepts, des groupes de termes ou des documents.
– Les arcs : représentent les dépendances entre termes et entre termes et documents.

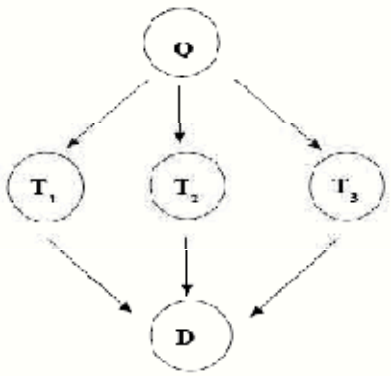


Figure 1.3 : Modèle de réseau bayésien simple [Nassr, 02]

L'avantage de ce modèle réside dans la présentation de la dépendance entre les termes, Mais il rassemble quelque inconvénient : l'estimation de probabilité conditionnelle n'est pas spécifiée, et les approches nécessitent l'utilisation des probabilités bayésiennes à sa place. La probabilité bayésienne d'un évènement est le degré de croyance humaine à cet évènement et donc le calcul des probabilités demande un temps d'ordre exponentiel du nombre de termes de la requête.

4.3.3. Le modèle de langages

Le modèle de langue est emprunté de la linguistique informatique. L'objectif d'un modèle de langue est de capter les régularités linguistiques d'une langue en observant la distribution des mots ou les successions de mots dans la langue donnée. Le modèle de langue désigne une fonction de probabilité qui assigne à chaque séquence de mots une probabilité.

En RI les modèles de langues déterminent la probabilité (notée P(Qj|Dk)) pour que la requête puisse être générée par le modèle de langue du document. Les initiateurs de ce modèle Ponte et Croft [Ponte & al., 98] disent que le calcul de cette probabilité repose sur l'hypothèse suivante : un utilisateur en interaction avec un système de recherche d'information fournit une requête en pensant à un ou plusieurs mots qu'il souhaite retrouver. La requête est alors inférée par l'utilisateur à partir de ces ermest. Un mot n'est pertinent que si la requête utilisateur ressemble à celle inférée par le terme. L'approche proposée par Ponte et Croft est non paramétrique, c'est-à-dire, la probabilité P(Qj|Dk) se base complètement sur les fréquences observées dans le corpus, et il n'est pas nécessaire d'apprendre un paramètre spécifique. Ponte et Croft rapportent que cette approche donne de meilleures performances que le modèle vectoriel.

5. Évaluation des Systèmes de RI

Après la réalisation du système, l'étape d'évaluation intervient pour mesurer la fiabilité et la satisfaction d'un système de recherche. En effet elle permet de caractériser le modèle et de fournir des éléments de comparaison entre les modèles. Les systèmes de recherche d'information qui sont aujourd'hui destinés à des utilisateurs non-spécialistes permettent une grande richesse d'exploration en facilitant la consultation directe

des documents, pour cela une étude d'évaluation est nécessaire dans le sens où elle permet de contrôler et d'évaluer les opérations et la performance du système.

Selon [Tebri, 04] [Karbasi, 07]. l'évaluation d'un système de recherche d'information peut être appréhendée selon deux aspects : un aspect efficience qui est lié au rendement (rapidité et/ou quantité de ressources utilisées), il dépend de l'évaluation cognitive de l'utilisateur, tels que la facilité d'utilisation du système, rapidité d'accès, temps de réponse à une requête, présentation des résultats, nombre d'entrée-sortie, etc. Et un aspect efficacité qui est lié à la qualité du résultat, et qui concerne la capacité du système à sélectionner le maximum de documents pertinents et un minimum de documents non pertinents.

6. Conclusion

Nous avons présenté dans ce chapitre les principales notions et concepts de la recherche d'information. Nous avons décrit l'architecture des SR I, ainsi que la reformulation de la requête qui permet de bien représenter cette dernière.

Après, nous avons décrit les modèles de recherche et de représentation d'information de manière particulière par les différentes méthodes d'indexation.

L'estimation de pertinence ainsi que l'évaluation des performances des systèmes de recherche d'information et les mesures de similarité sémantique qui sont des étapes primordiales dans les SRI.

Actuellement, un effort particulièrement considérable est fourni dans la recherche de systèmes capables de retrouver des mots demandés par des utilisateurs ne connaissant pas nécessairement le vocabulaire qu'ils cherchent.

L'assistance de l'utilisateur par amélioration de sa requête en utilisant les « ontologies » est un de ses axes.

Nous développerons ci-après le concept d' « ontologie » dont l'utilisation est récente notamment dans le domaine de la recherche d'information.

Plan

1. Introduction

Ces derniers années un nouveau concept avais apparu et fait un grand succès au domaine d'informatique plus précisément à l'intelligence artificiel, c'est l'ontologie. En philosophie, l'ontologie est une branche fondamentale de la Métaphysique qui s'intéresse à la notion d'existence, aux catégories fondamentales de l'existant et étudie les propriétés les plus générales de l'être. L'ontologie a donc un rapport direct avec notre conception de la réalité.

Ses défis on peut les voir dans plusieurs champs d'informatique, par exemple l'ingénierie des connaissances, le traitement du langage naturel (NLP), les systèmes d'information coopératifs, l'intégration intelligente d'information, la gestion des connaissances et la recherche d'information qui est le domaine qui nous intéresse dans ce mémoire.

La manipulation des connaissances partagées par les ontologies rend son utilisation nécessaires dans les systèmes de recherche d'informations afin de doté ces derniers par un peu de sémantique et diminue la divergence entre le besoin de l'utilisateur et les réponses système.

2. Les ontologies

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

2.1. Définitions des ontologies

Définition1 : «Une ontologie peut prendre différentes formes, mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification. Cette dernière inclut des définitions et une indication de la façon dont les concepts sont reliés entre eux, les liens imposent collectivement une structure sur le domaine et contraignent les interprétations possibles des termes.» [Baziz, 02]

Définition2 : « une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire ».

Cette définition explique la façon d'élaborer u ne ontologie en nous offrant des directives relativement floues : repérer les termes de base et les relations entre les termes, identifier les règles servant à les combiner, fournir des définitions de ces termes et de ces relations. Notons que d'après cette définition, une ontologie inclut non

seulement les termes qui y sont explicitement définis, mais aussi les termes qui peuvent être créés par déduction en utilisant les règles. En 93, Gruber [Gruber, 93] **.** formule la définition suivante :

De manière générale, on identifie les catégories suivantes : Les ontologies de représentation des connaissances, les méta-ontologies, les ontologies de domaine, les ontologies de tâches, les ontologies de domaine-tâche, les ontologies d'application ainsi que les ontologies interactives :

Définition3 : « une ontologie est une spécification explicite d'une conceptualisation ».

Cette définition est la plus citée et restera la plus juste. Elle signifie que la construction d'ontologie intervient après une étape de conceptualisation.

La définition de Guarino et Giaretta, «Une ontologie est une spécification rendant partiellement compte d'une conceptualisation». La conceptualisation étant spécifiée parfois de manière très précise, une théorie logique ne peut pas toujours en rendre compte de façon exacte : elle ne peut assumer la richesse interprétative du domaine conceptualisé dans une ontologie et ne le fait donc que partiellement.

Pour Borst, en modifiant légèrement la définition de Gruber : « l'ontologie est une spécification formelle d'une conceptualisation partagée ».

Définition4 : « une ontologie est un ensemble de termes structurés de façon hiérarchique, conçu afin de décrire un domaine et qui peut servir de charpente à une base de connaissances ».

Cette définition se base sur le fait qu'ils construisent des ontologies de connaissances spécifiques à des domaines d'expertise en identifiant les termes significatifs d'un certain domaine de l'ontologie Sensus (qui inclut plus de 50 000 termes). Bernaras et ses collègues construisent une ontologie différemment, en partant d'une base de connaissances qui sera raffinée et enrichie de nouvelles définitions si de nouvelles applications sont crées. Une définition a été proposée dans ce sens par [Gomez, 99] :

Définition5 : « une ontologie fournit les moyens de d écrire de façon explicite la conceptualisation des connaissances représentées dans une base de connaissances. »

La Linguistique est aussi concernée par la question des ontologies dans la mesure où les données dont on dispose pour élaborer les ontologies consistent en des expressions linguistiques de connaissances.

La caractérisation du sens de ces expressions conduit à déterminer des signifiés contextuels, dépendants des contextes (documents) où les expressions apparaissent.

Ces signifiés contextuels doivent alors être normés, ce qui revient à fixer une signification pour un contexte de référence, celui de la tâche (application) pour laquelle l'ontologie est élaborée [Bachimont, 00] .

2.2. Les principaux types d'ontologies

Le domaine des ontologies est très vaste et ses utilisations comprennent plusieurs champs. On peut trouver une ontologie différente selon chaque cas d'utilisation.

Selon [Baziz, 02] :

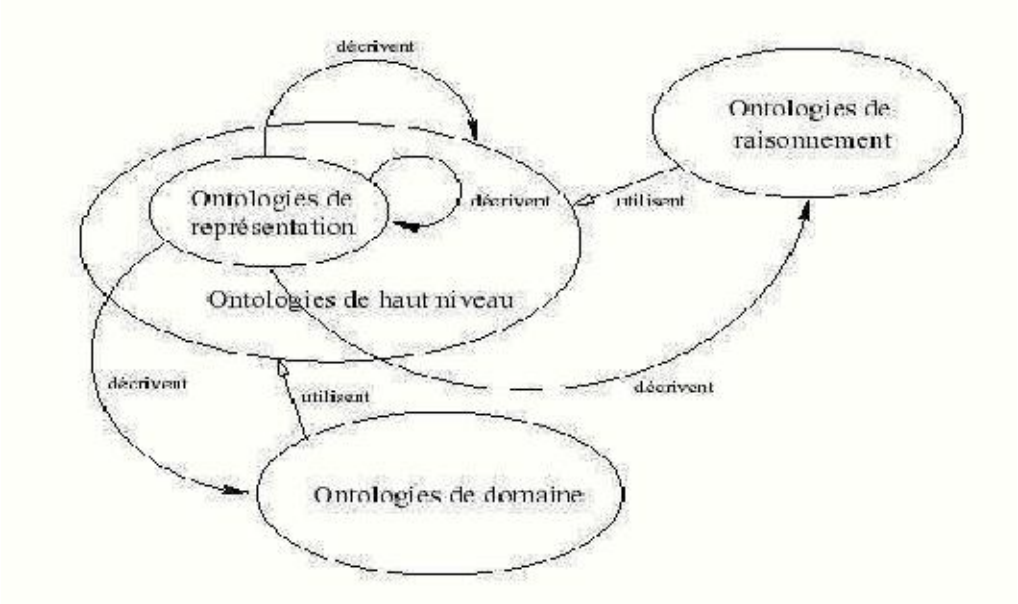


Figure 2.1 : les différents types d'ontologies [Fürst, 04]

Les ontologies de représentation de connaissance : regroupent les primitives de représentation utilisées afin de formaliser les connaissances. L'exemple le plus représentatif de ce type d'ontologie est la Frame-Ontology [Gruber, 93] , qui rassemble les primitives de représentation (classes, instances, cases, facettes, etc.) utilisées dans les langages à base de frame.

Les ontologies générales/communes : incluent le vocabulaire lié aux objets, aux événements,au temps, à l'espace, à la causalité, au comportement et à la fonction.

Les méta-ontologies : également appelées ontologies génériques ou noyaux également appelées ontologies génériques ou noyaux d'ontologies, spécifiant les processus de raisonnement appliqués aux connaissances. Résolution automatique de problèmes. Seules sont décrites les connaissances portant sur la façon d'utiliser d'autres connaissances. Ces ontologies sont réutilisables dans différents domaines. L'exemple le plus représentatif serait une ontologie méréologique, qui inclurait le terme partie de.

Les ontologies de domaine : sont réutilisables dans un domaine donné. Elles fournissent le vocabulaire des concepts d'un domaine (par exemple scalpel, scanner dans un domaine médical) et les relations entre ces derniers, les activités de ce domaine (par exemple anesthésie, accoucher) ainsi que les théories et les principes de base de ce domaine.

Les ontologies de tache : fournissent un vocabulaire systématisé des termes utilisés pour résoudre les problèmes associés à des taches qui peuvent appartenir ou non à un même domaine. 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 (par exemple plan, objectif, contrainte), des verbes génériques (par exemple assigner, classer, sélectionner), des adjectifs génériques (comme assigné) et d'autre mots qui relèvent de l'établissement d'échéances.

Les ontologies de domaine-tâche : Se sont des ontologies de tâches réutilisables dans un domaine donné, mais pas dans différents domaines. Par exemple une ontologie domaine-tâche dans le domaine médical, pourrait inclure les termes liés au timing d'une intervention chirurgicale : planifier-intervention chirurgicale.

Les ontologies d'application: Contiennent suffisamment de connaissances pour structurer un domaine particulier.

2.3. Les ontologies les plus connues

Cette section introduit les ontologies les plus connues en prenant en compte la typologie d'ontologies énoncée ci-dessus.

Actuellement avec explosion du web et sa facilité on communication transforme à un support pour plusieurs ontologies, telles que Ontolingua sur le serveur Ontolingua [Farquhar & al., 97] **.** et l'ontologie coranique sont disponibles gratuitement sur le web. D'autres ontologies, telles que les ontologies de Cyc [Lenat & al., 90] **.** sont partiellement disponibles gratuitement sur le Web. La majorité d'entre elles sont propres à des entreprises et leurs utilisations n'est pas gratuite.

Voici quelques ontologies les plus connues [Baziz, 02]:

-L'ontologie Page (également connue sous le nom de Top) et « (Onto) 2Agent » qui est un moteur de recherche sur Internet basé sur une ontologie et qui aide à sélectionner des ontologies.

- Les ontologies de haut niveau : fournissent des concepts généraux permettent de définir tous les termes des ontologies existantes. Parmi elles, Sowa, le Penman Upper Level et Cyc.

- L'ontologie méréologique : pourrait être l'exemple typique d'une méta-ontologie. Cette ontologie d éfinit la relation partie-de et ses propriétés. Cette relation permet d'exprimer que des instruments sont formés de composants, qui peuvent eux même être constitués d'éléments plus petits.

- L'ontologie Cyc: est une ontologie de sens commun qui fournit une grande quantité de savoir humain élémentaire. Elle consiste en un ensemble de termes et d'affirmations liées à ces termes. Elle se décompose par ailleurs, en différentes micro-théories. Chaque micro-théorie rend compte seulement d'un point de vue important d'un domaine de connaissances. Certains domaines peuvent traiter plusieurs micro-théories qui représentent différentes perspectives et affirmations et divers niveaux de granularité et de distinction. Les ontologies Cyc sont implémentées dans un langage appelé Cycl.

WordNet est une base de données lexicale pour l'anglais fondée sur des principes psycholin guistiques. Ses informations sont ventilées en unités appelées « synsets » en anglais, qui sont des jeux de synon ymes interchangeables dans un contexte particulier utilisés pour représenter différents sen s. Dans notre mémoire on utilise Wordnet arabe qui sera abordée avec plus de détails dans le chapitre suivant.

Sensus est une ontologie basée sur le langage naturel, qui a pour fonction de fournir une vaste structure conceptuelle aux travaux menés en matière de traduction automatique. Il elle a été mise au point en rassemblant et en extrayant des données de ressources électroniques telles que : Penman Upper Model, Ontos, WordNet et des dictionnaires électroniques de langages naturels. Elle compte plus de 50 000 notions.
Dans le domaine des ontologies d'ingénierie, les ontologies EngMath et PhysSys

sont les plus connues. EngMath est une ontologie Ontolingua mise au point pour la modélisation mathématique en ingénierie. PhysSys est une ontologie d'ingénierie destinée à modéliser, simuler et concevoir des systèmes physiques : Représentation du système, comportement de processus ph ysique, et relations mathématiques descriptives.

Les ontologies qui représentent le mieux les ontologies dédiées à la modélisation d'entreprises sont l'Enterprise Ontologie et la Tove Ontologie. L'Enterprise Ontology est un ensemble de termes et de définitions pertinent pour les entreprises commerciales et inclut des connaissances sur les activités et les processus, les organisations, les stratégies, le marketing, etc.

Les ontologies élaborées dans le cadre du projet Tove1 sont l'ontologie de conception d'entreprises, l'ontologie des projets, l'ontologie-agenda, ou encore l'ontologie des services.

L'ontologie (KA) 2 constitue une référence pour les ontologies dédiées à la gestion des connaissances, qui sera utilisée par le Knowledge Annotation Initiative de la communauté d'acquisition des connaissances. Cette ontologie servira de base pour annoter les documents sur Internet de la communauté d'acquisition des connaissances de façon à fournir un accès intelligent à ces documents. Des spécialistes situés dans des zones géographiques différentes travaillent ensemble à la mise au point de cette ontologie.

3. Ontologies et recherche d'information

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

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

3.1. Quelles ontologies choisir

Une première solution vise à construire une ontologie à partir du ou des corpus sur lesquels les tâches de RI vont être réalisées. Cette solution assure a priori l'adéquation entre l'ontologie construite, le corpus et la tâche à réaliser. Cette solution

n'est pas toujours adaptée: elle est coûteuse et ne prend pas en compte l'existence de ressources qui pourraient être réutilisées. Maintenant avec l'avènement de domaine des ontologies, elles sont devenues des standards à réutiliser. Une autre solution très utilisée par la majorité des approches de R I visant à intégrer ces ontologies dans ces approches [Baziz, 05] .

Généralement, l'unique caractéristique prise en compte dans le choix de l'ontologie est le domaine de connaissance représentée dans l'ontologie qui doit couvrir le domaine traité. C'est le cas par exemple du système Cat-a-cone qui repose sur la hiérarchie de concepts du domaine de la médecine MESH [Hearst, 97] (in Abderahim). pour explorer une collection documentaire du même domaine, ou bien des travaux présentés dans [Baziz, 05] . qui repose sur l'ontologie générale WordNet pour une tâche de RI ad-hoc sur une collection de TREC.

Dans notre mémoire on a choisit l'ontologie coranique qui nous sert bien à effectuer une recherche Sémantique, et cette dernière elle sera bien décrite dans le chapitre suivant.

Cependant, nous pensons que ce choix n'est pas aussi simple. Les ontologies considérées doivent être adaptées aux tâches de RI visées et surtout trouver une connaissance pertinente par rapport à l'information présente dans les collections. Ceci est également l'explication donnée dans de nombreux travaux pour justifier le fait que l'indexation conceptuelle ait besoin d'être couplée à une indexation classique afin d'améliorer les performances des systèmes [Vallet, 05] [Mihalcea, 00] .

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

3.2. Principe d'utilisation des ontologies par un Système de Recherche d'Informations

L'ontologie peut intervenir au différents phases dans le processus de recherche d'information à l'interface entre la requête de l'utilisateur et le module chargé de la recherche effective des documents dans la base.

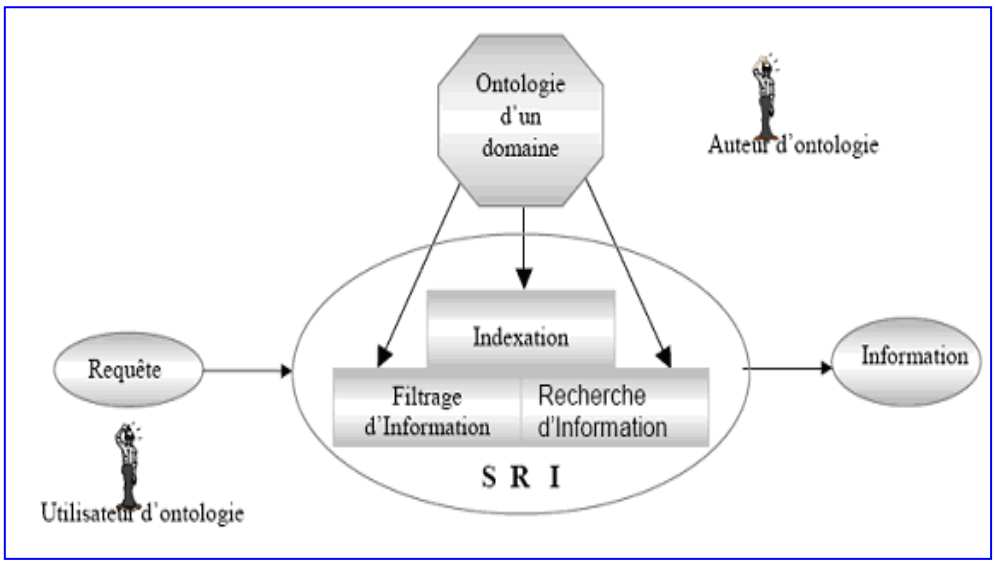


Figure 2.2: Principe d'intégration d'une ontologie au processus de RI [Baziz, 02]

3.2.1. L'ontologie et la reformulation de la requête

Les utilisateurs recherché ou ils expriment ces besoins difficilement à l'aide d'une requête mal écrite. A ce point, l'ontologie intervient pour aider l'utilisateur a formulé leur requête afin d'exploiter efficacement la collection de document.

Il existe deux types d'expansion de requête dans la littérature : La première consiste à utiliser des ressources, internes ou externes, comme par exemple un dictionnaire [Moldovan & al., 99] ou bien WordNet [Voorhes, 94] , basée sur l'extension des requêtes par l'ajout de nouveaux termes en relation avec les termes de la requête.

La deuxième solution réinjection de pertinence reposant sur l'analyse des termes contenus dans le document jugés pertinents pour la requête initiale. L'idée est que l'ajout de termes liés aux termes initiaux de la requête peut permettre de retrouver les mots qui ne sont pas restitué auparavant. Cette approche, ne faisant pas intervenir d'ontologies, ne fait pas partie de notre étude.

Un autre intérêt des ontologies est de permettre la désambiguïsation des termes de la requête.

Dans [Guha & al., 2003] . la désambiguïsation se fait selon trois méthodes. La première consiste à choisir le concept dont les labels les plus fréquents dans le document. La seconde approche consiste à réaliser un profil utilisateur et à choisir le concept le plus proche de son profil. Finalement, la troisième prend en compte le contexte de la recherche et les motsrecherchés par l'utilisateur.

Dans (Ka)2 [Benjamins & al., 99] , les pages Web sont annotées manuellement par des concepts d'une ontologie. Tous les concepts liés aux termes d'une requête donnée sont inférés et ajoutés à cette dernière. L'utilisateur est assisté dans la formulation ou le raffinement de sa requête à l'aide d'une interface proposé par ce système. L'utilisateur a la possibilité de naviguer dans l'ontologie et de centrer la visualisation sur la représentation des concepts qui l'intéresse comme il a été fait dans WebBrain3.

[Köhler & al., 06] . améliore la désambigüisation des sens des mots en utilisant la lemmatisation des mots. De plus, ils proposent une méthode pour améliorer le rappel sans modifier la précision par l'utilisation des sous-concepts et super-concepts dans les différentes relations en respectant une certaine limite sur la profondeur des relations de subsomption.

[Aufaure & al., 07] . adapte le model vectoriel par la substitution des termes de la requête par des concepts de l'ontologie et classifie par service les résultats d'une requête en utilisant une ontologie de services permettant de spécifier les services liés à un domaine spécifique : acteurs, activités ou tâches. L'enrichissement de requêtes utilisateurs se fait par analyse morphologique et sémantique en utilisant les concepts et les relations entre l'ontologie de domaine et WordNet. L'utilisateur peut aussi utiliser l'ontologie de domaine pour désigner les concepts à utiliser dans sa requête.

Dans [Tomassen & al., 06] , l'enrichissement de la requête utilisateur se fait par substitution des concepts de la requête par les vecteurs caractéristiques des concepts correspondants dans l'ontologie. Cette méthode associée a chaque concept de l'ontologie de domaine un vecteur caractéristique décrivant la similarité sémantique du concept avec les termes et concepts auxquels il est en relation (Synonyme, conjugaison, etc.) par rapport aux contenus de document. Le but de ce système est

de rapprocher les requêtes au contexte d'utilisateur et aux caractéristique de document utilisant l'ontologie.

Dans [Kim & al., 07] , la recherche d'information se déroule en deux phases. Premièrement, la requête utilisateur est reformulée à l'aide des concepts de l'ontologie qui correspondent aux mots clés de la requête. Deuxièmement, le système réalise la recherche d'objet contenant ces concepts.

Harman [Harman, 92] . a prouvé que la reformulation de requêtes a des effets positifs en RI. L'objectif de la reformulation est soit de limiter le silence soit de réduire les risques de bruit.

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

L'utilisation d'ontologies pour la spécification du besoin utilisateur et la recherche des mots correspondant présente plusieurs avantages [Hernandez, 05] : .

* Elle permet tout d'abord la reformulation de requêtes envoyées à des systèmes

traditionnels à partir de l'ajout ou de la désambiguïsation de termes de l'ontologie.

* Les ontologies lourdes peuvent également mener à la mise en place de *mécanismes d'inférences permettant la recherche des éléments de l'ontologie répondant à une requête à partir de langages d'interrogation sophistiqués.*

3.2.2. Appariement à partir d'ontologies

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

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

Le calcul de similarité s'appuie sur trois intuitions [Hernandez & al., 08] :

La première intuition est que les mots lié au concept généralisant ou spécifiant le concept utilisé dans larequête peuvent intéresser l'utilisateur.

Lecalcul de la similarité prend donc en compte la distance séparant les deux concepts par la relation de subsomption.La similarité revient à prendre le nombre d'arcs séparant les deux concepts par le chemin le plus court à partir de la relation de subsomption.

La deuxième intuition est que deux concepts ayant un concept les généralisant (ou subsumeur) commun sont plus similaires. Afin d'appliquer cette intuition, chaque concept est représenté par un ensemble flou à partir des concepts le généralisant. La similarité entre concepts est alors calculée à partir des éléments faisant partie l'intersection entre les

descriptions des concepts.

La troisième intuition est que la similarité entre concepts doit prendre en compte les relations autres que les relations de subsomption. L'ensemble des concepts généralisant les deux concepts est alors considéré.
Un sous-graphe de l'ontologie est construit à partir des concepts de cet ensemble pouvant être reliés dans l'ontologie par n'importe quel type de relation. La similarité est calculée par rapport aux nombres de nœuds .

3.2.3. L'ontologie et la représentation du document :
une autre utilisation des ontologies dans les SRI au niveau de représentation des requêtes et document. L'indexation de document et requêtes à l'aide des mots neutre Prouve qu'elle est insuffisante et ne donne pas des bons résultats [Khan, 00] . . Les chercheurs ont pensé d'ajouter un peut de sémantique dans les termes choisis comme des index à travers une ontologie avec leur capacité de manipuler les connaissances d'un domaine. Ce type d'indexation s'appel l'indexation sémantique.

L'indexation sémantique n'est possible que par l'existence et l'utilisation de ressources décrivant explicitement l'information correspondant aux objets. Deux types de démarches peuvent être distinguées : la démarche issue de la R I et la démarche issue du Web sémantique. La démarche issue du domaine de la RI consiste à choisir comme langage de représentation du document, l'ensemble des concepts et instances de l'ontologie. L'utilisation d'ontologies sous forme de hiérarchies de concepts, ontologies légères2 ou lourdes est le prolongement de l'utilisation dans le cadre de la RI des ressources terminologiques [Haav & Lubi, 01] . Le document est alors indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambigus [Aussenac & Mothe, 04] . L'ontologie utilisée dans ce cas reflétant le ou les domaines de connaissance étudiés à au document. Il est en effet nécessaire de retrouver dans l'ontologie les concepts présents dans le document pour indexer ce dernier à partir de toutes les thématiques abordées.

L'indexation sémantique est un type d'indexation qui s'inscrit également dans la démarche orientée Web Sémantique. Les précurseurs de cette nouvelle version du Web considèrent que les ressources participant au Web Sémantique seront toutes reliées entre elles par des relations sémantiques. Plus précisément, les données présentes sur le Web Sémantique seront modélisées sous forme d'ontologies où chaque ressource apparaît comme un élément de ces ontologies au même titre que la connaissance qui les décrit. L'objectif est donc d'ajouter au contenu du Web une structure formelle et de la sémantique (à travers des méta-données et de la connaissance) dans le but de permettre une meilleure gestion et un meilleur accès aux informations. Cette démarche repose sur des ontologies modélisant les objets du monde à travers les acteurs et entités que les documents constituent et comportent [Guha & al., 03]

.Elles peuvent être vues comme une représentation des méta-données explicitement ou implicitement présentes dans les documents. La phase d'indexation est aussi appelée annotation de documents. L'annotation de documents a pour but de représenter les informations relatives au média (date de création, taille, format d'encodage), les métadonnées présentes dans les documents (auteurs, date de production), les index (les descripteurs du contenu du document), l'identifiant du document par le système (emplacement) et une vue sur

Indexation sémantique (Sense Based Indexing) : l'indexation sémantique est une approche d'indexation basée sur le sens des mots [Mihalcea & al, 00]. Elle repose sur des algorithmes de désambiguïsation de mots (WSD) pour indexer les documents et les requêtes avec le sens des mots (mots-sens) plutôt qu'avec des motssimples. Une manière d'indexer serait par exemple, d'associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens.

D'autres approches de désambiguïsation plus élaborées, utilisent des représentations hiérarchiques pour calculer la distance sémantique ou similarité sémantique entre les mots à

comparer. Cette notion de distance sémantique est plus générale et peut aussi être utilisée, dans l'indexation conceptuelle pour calculer la proximité sémantique entre les concepts d'une ontologie. Pour Sanderson, une désambiguïsation "performante" permet d'améliorer les performances des SRI, notamment dans le cas des requêtes courtes.

[Hernandez, 05]

Indexation conceptuelle : l'indexation conceptuelle se base sur des concepts tirés d'ontologies et de taxonomies pour indexer le document contrairement aux listes de mots simples.
L'appariement des concepts (Concept Mapping) peut être utilisé dans un système spécifique en rapport à un domaine fermé, comme le domaine du religion, le domaine légal, le domaine médical (dans le système Textpresso et MetaMap (UMLS)), ou encore général, comme c'est le cas dans le système FERRET, les travaux de Mauldin et Woods utilisent un dictionnaire de langue, pour l'apprentissage des relations entre termes pour le premier (en utilisant un algorithme génétique) et pour construire avec les termes des documents des hiérarchies de concepts (taxonomie) pour le deuxième. Aggarwal et ses collègues, construisent des "chaînes de mots conceptuelles" (conceptual word-chains) à partir de la taxonomie de concepts de *Yahoo!* ***et utilisent cette représentation comme alternative à la représentation classique par fichier inverse.***

Notre mémoire s'inscrit donc dans le cadre général de l'utilisation des ontologies et autres ressources conceptuelles pour la représentation de l'information et plus particulièrement, nous nous intéressons au processus d'indexation conceptuelle guidée par ontologie dans un SRI.

L'indexation du contenu du document à partir d'une ontologie présente de plus les avantages suivants :

* Aider l'utilisateur à formuler sa requête. En présentant l'ontologie à l'utilisateur, il est possible de le guider dans le choix des termes de sa requête.

* Faciliter la RI : dans le contexte de la recherche d'information, une ontologie n'est généralement pas représentée logiquement. Le formalisme utilisé sert habituellement à faciliter la gestion des concepts en tant qu'objets, leur classification, la comparaison de leurs propriétés et la navigation au sein de l'ontologie en accédant à un concept et à ceux qui lui sont reliés. Son utilisation est donc réduite à l'accès à la connaissance pour permettre une meilleure indexation ou stockage des informations et faciliter la

recherche

4. Conclusion

Durant ce chapitre on a vu les déférentes définitions d'ontologies ainsi que les différent types d'ontologies et les plus connues de ces derniers. Le choix des ontologies est une tache cruciale dans ce domaine afin d 'être utiliser efficacement par les différents modules du système de recherche d'information. Après on a montré l'intérêt des ontologies; dans la reformulation de la requête, l'appariement ontologique, la représentation du document, et l'indexation conceptuel.

Dans ce qui suit, nous présenterons notre outil de recherche d'information basé sur L'ontologie coranique Par la suite nous proposons une première évaluation de l'apport de cette démarche dans l'augmentation des performances du système ainsi construit.

Plan

1. Introduction
2. L'ontologie coranique
3. Problématique
4. Le système de RI proposé
5. Conception de la base de l'ontologie coranique
6. L'architecture de notre système
7. Les outils utilisés
8. Implémentation de notre système
9. Les resultats obtenus
10. Conclusion

1. Introduction :

Face a la richesse et la puissance de la langue arabe au niveau syntaxique et sémantique les outils de recherche actuels sont très pauvre et ne proposent que des solutions de dépannage et la recherche d 'information en arabe est devenus un véritable casse-tête d'ou des résultats de mouvais qualités et un sentiment pour l'utilisateur qui en utilisant un moteur censé lui faciliter les recherches découvre un maelstrom de listes de liens sans rapport évident avec ces besoins. De notre point de vue la solution vient de l'utilisation des techniques de traitement sémantique et syntaxique de la langue arabe qui constitue un grand pas vers son intégration dans la technologie de l'information. La prise en charge de la richesse de la langue arabe lors de la recherche d'information est très nécessaire pour avoir des résultats de recherche plus satisfaisants et plus riche.

Les ontologies sont un sujet de recherche populaire dans diverses communautés notamment l'ingénierie des connaissances, la recherche d'information et le traitement du langage naturel (NLP), les systèmes d'information coopératifs, l'intégration intelligente d'information et la gestion des connaissances. On peut d'abord se demander si l'utilisation des ontologies dans la recherche d'information est un phénomène récent ou pas. Et la recherche d'information électronique tel qu'il est connu actuellement en utilisant des SRI, comment une ontologie peut- elle être associée au processus de recherche d'information ?

De manière générale, ce qui est attendu d'une ontologie, est qu'elle assure la réutilisation de connaissances. En recherche d'information, son apport est ciblé. C'est également l'objectif de notre travail tel que dans ce dernier chapitre nous allons essayer de développer un système de recherche d'information spéciale pour le coran . On essaye de résoudre le problème de variation conceptuel par l'intégration de l'ontologie coranique dans les différentes étapes de SRI, en vue d'obtenir des résultats satisfaisants.

2. L'ontologie coranique

L'ontologie coranique clés utilise la représentation des connaissances pour défini les concepts dans le Coran, et montre les relations entre ces concepts en utilisant la logique des prédicats. Les concepts fondamentaux dans l'ontologie sont basés sur les connaissances contenues dans les sources traditionnelles d'analyse du Coran. y compris le hadith du prophète Mahomet le tafsir(exégèse coranique) de ibn Kathir. entités nommées dans les versets, tels que les noms de personnages historiques et lieux mentionnés dans le Coran, sont liées à des concepts dans l'ontologie dans le cadre de l'étiquetage des entités nommées. [Kais Dukes]

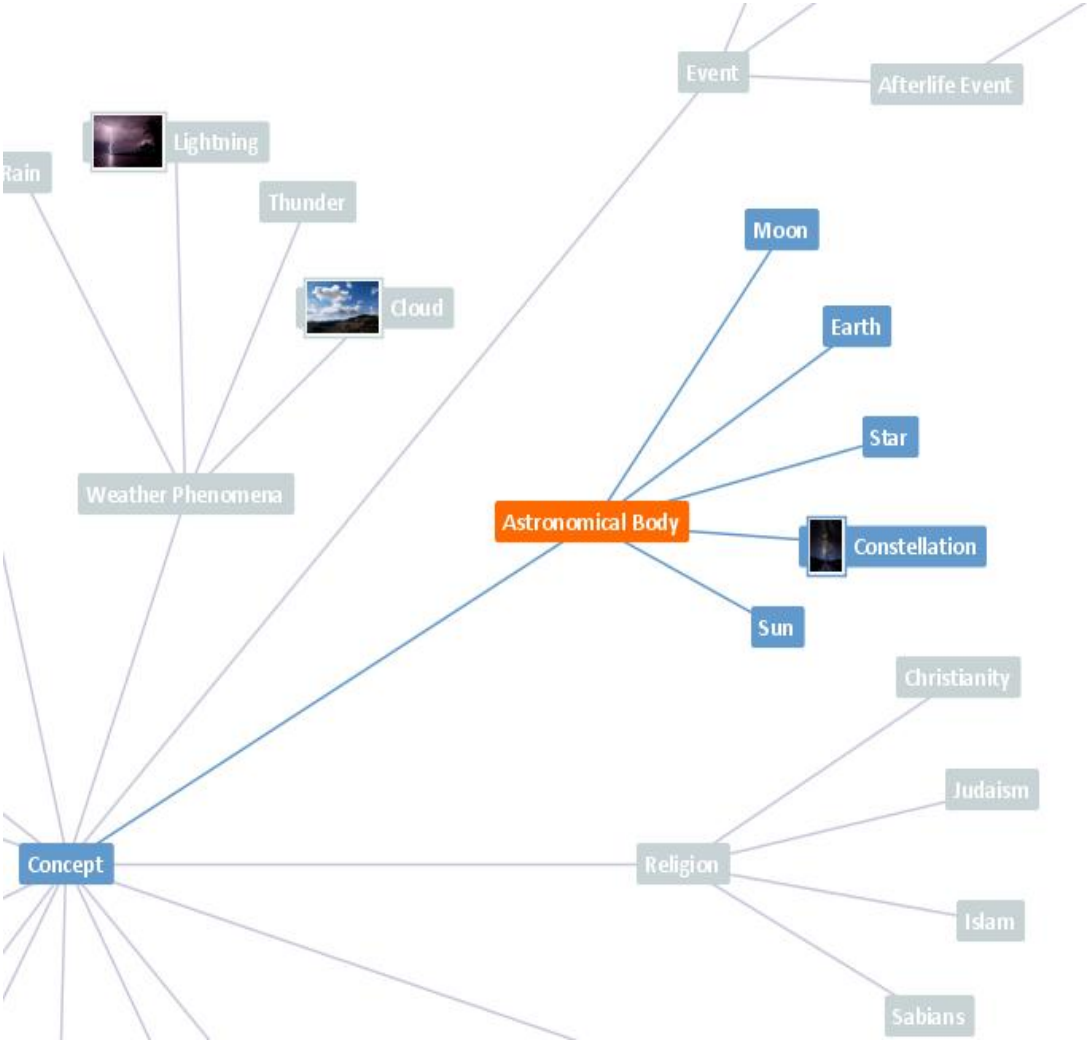


Figure 3.1 Le diagramme montrant une représentation visuelle de l'ontologie. Le graphique est un réseau de 300 concepts liés avec 350 relations. [Kais Dukes]

L'ontologie définit une liste des principaux concepts dans le Coran et un ensemble de relation sémantiques entre ces concepts. La relation la plus importante est la relation d'adhésion fixée "par exemple"dans laquelle un concept est défini comme une instance ou d'un membre individuel d'un autre groupe. Par exemple la relation «Satan est un djinn» dans l'ontologie représenterait la connaissance contenue dans le Coran que l'individu connu sous le nom de Satan appartient à l'ensemble des créations sensibles nommé les djinns. D'autres concepts dans l'ontologie et regroupés en catégories logiques, en fonction des propriétés qu'ils partagent. Par exemple, islam, christianisme, et le judaïsme sont classées sous la rubrique « Religion » **[Kais Dukes]**

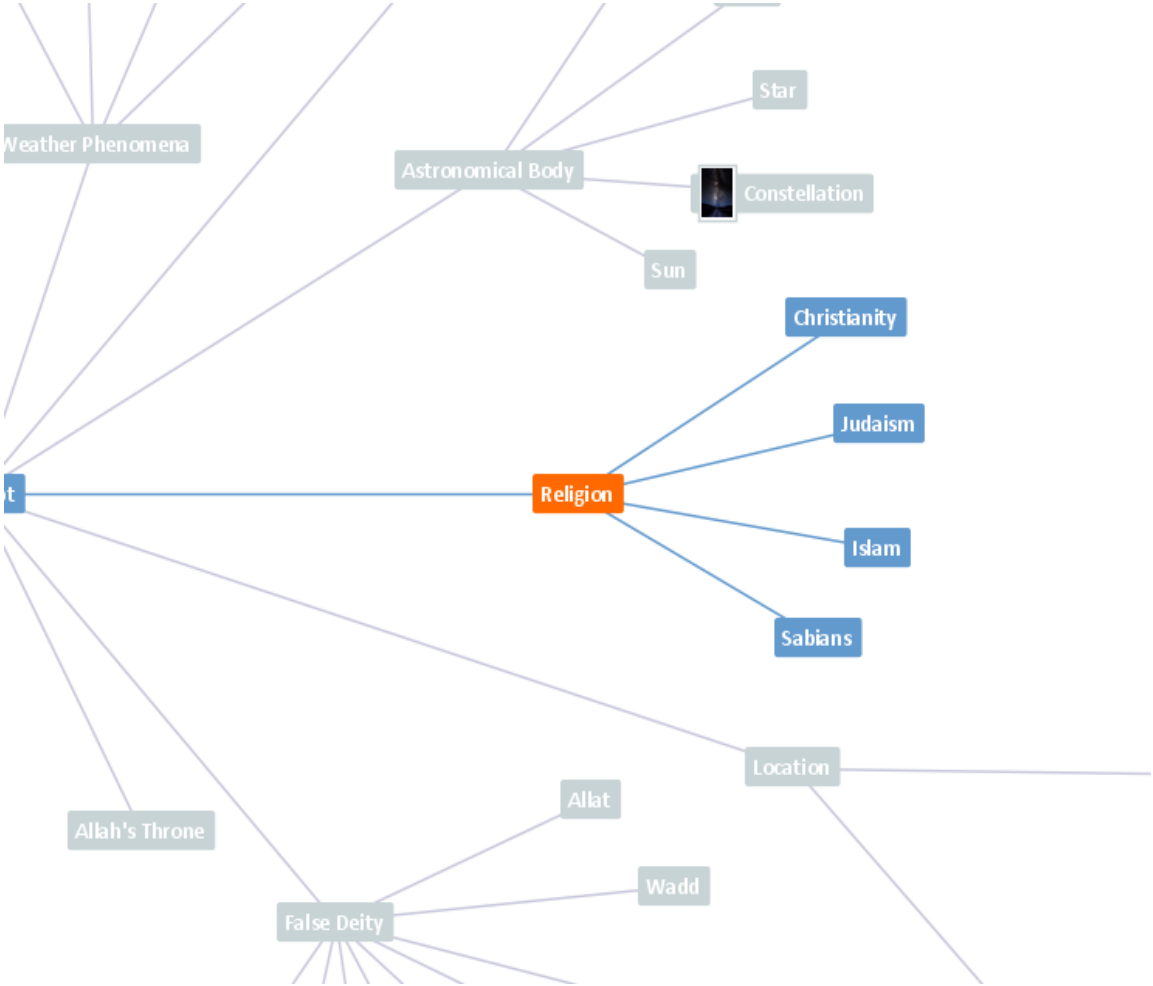


Figure 3.2 Concept Plan pour la religion. [Kais Dukes**]**

3. Problématique

La clé du succès dans l'obtention d'un bon résultat de recherche, réside dans la qualité de requêtes, alors à présent, le problème de la recherche d'information ne réside pas tant dans le volume de la connaissance que dans sa vitesse de croissance mais plutôt dans l'aptitude à retrouver la bonne information. Une recherche qui s'effectue en utilisant un vocabulaire standard peut se révéler peu fine.

Les problèmes des Systèmes de recherches actuelles résident soit au niveau de la requête d'utilisateur tel que cet utilisateur ne sait pas comment formulé son besoin, soit le document où l'utilisateur fait son recherche, est mal représentée par le système de recherche utilisé.

La variation morphologiques et lexicales dégrade l'efficacité des systèmes de RI en termes de rappel. D'un autre coté, les variations sémantiques et syntaxiques touchent la précision [Baziz, 02].

Dans le cadre de la recherche d'information, la récupération de mots clé de la requête d'utilisateur est jugée insuffisante, car les termes utilisés dans cette requête peuvent présenter par rapport au document de la base, des différences sur plusieurs plans, par exemple :

- des variations morphologiques comme dans « مسلم» et « مسلمون»
- des variations lexicales (on utilise pour le même sens des mots différents) comme dans le cas dans « حية » et « ثعبان» ;
- des variations sémantiques comme dans le cas de « البر:مرادف الطاعة و البر هو القمح

L'utilisation des ontologies au niveau de la requête d'utilisateur peut constituer une solution (parmi d'autres) pour résoudre le problème des variations sémantiques. Par ailleurs l'utilisation d'un analyseur morphologique peut suffire pour résoudre les deux premiers cas de variations (morphologiques ( ) et lexicales .

Le tableau suivant (Tableau 3.1) montre des exemples de requêtes enrichies par utilisation d'une ontologie (ontologie coranique dans notre cas).

N requête	Requête	Requête Enrichie
1	دين	إسلام ;مسيحية ; يهودية
2	إمدار	إسراف ; تبذير
3	إبداع	ابتكار; تكوين ; خلق

Tableau 3.3 : quelques mots de recherche.

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

La solution proposée dans ce mémoire consiste à intégrer l'ontologie uecoraniq aux modules de système de recherche. Plus précisément nous allons utiliser l'ontologie

coranique pour indexer le coran et la requête d'utilisateur.
Rappelons que, les ontologies contiennent une représentation formelle des concepts de Domaine (coran) , et leur composante terminologique (lorsqu'elle existe) permet d'accéder aux différents termes qui désignent les concepts. On peut situer l'utilisation des ontologies pour la RI à plusieurs étapes du processus qui va du traitement de la requête jusqu'à l' afichage des résultats:

4. Le système de RI proposé

Les systèmes de recherche d'information classiques essayent d'optimiser les temps de recherche en identifiant les mots selon des critères d'appariement entre les mots contenus dans les requêtes utilisateurs (et uniquement ceux-là) et ceux des concepts (les indexes).
La solution proposée dans ce mémoire consiste à intégrer l'ontologie coranique aux modules de système de recherche. Plus précisément nous allons utiliser l'ontologie coranique pour indexer le coran et la requête d'utilisateur.
Rappelons que, les ontologies contiennent une représentation formelle des concepts de Domaine (coran) , et leur composante terminologique (lorsqu'elle existe) permet d'accéder aux différents termes qui désignent les concepts. On peut situer l'utilisation des ontologies pour la RI à plusieurs étapes du processus qui va du traitement de la requête jusqu'à l' afichage des résultats:

- Pour la reformulation de la requête : où les utilisateurs verront leur requêtes coran, améliorées et enrichi avec des termes résultant de concepts reliés sémantiquement à ceux de leur requêtes.
- Pour l'indexation (la représentation) de coran : où des concepts de l'ontologie au lieu de mots isolés seront reliés aux concept après désambiguïsation.
- Pour le filtrage : des ontologies peuvent être utilisées pour créer à partir de coran des groupes de mots constituant ainsi différents profils utilisateurs.

Dans notre cas, nous intéressons à la première et la deuxième étape : l'idée est donc d'exploiter le contenu d'une ontologie générale pour indexer la requête et le coran de manière à retrouver plus précisément les bons mots. Pour expérimenter cette approche, nous avons exploité l'ontologie coranique sous forme de base de donnée quand a construit.

5. Conception de la base de l'ontologie coranique :

5.1 Diagramme de classe :

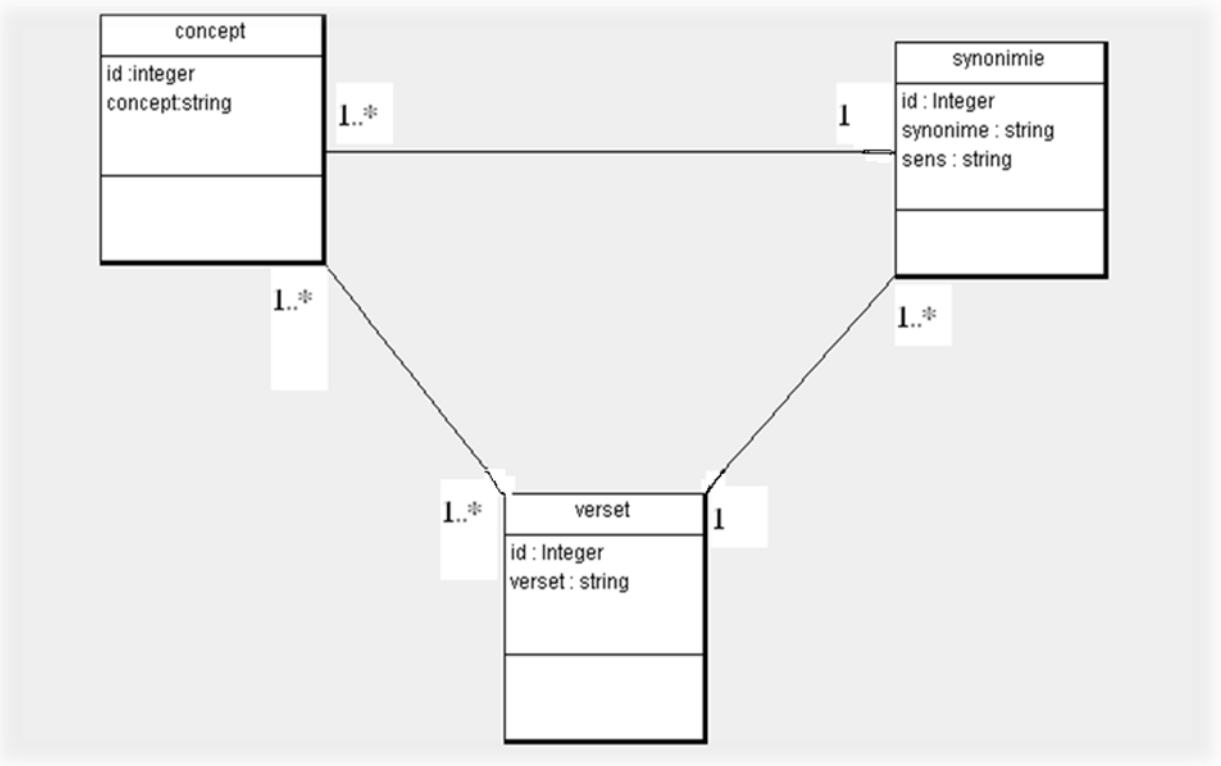


Figure 3. 4 : Diagramme de classe

6. L'architecture de notre système

La figure (voir figure 3.5) présente l'architecture générale de notre système de recherche d'informations arabe:

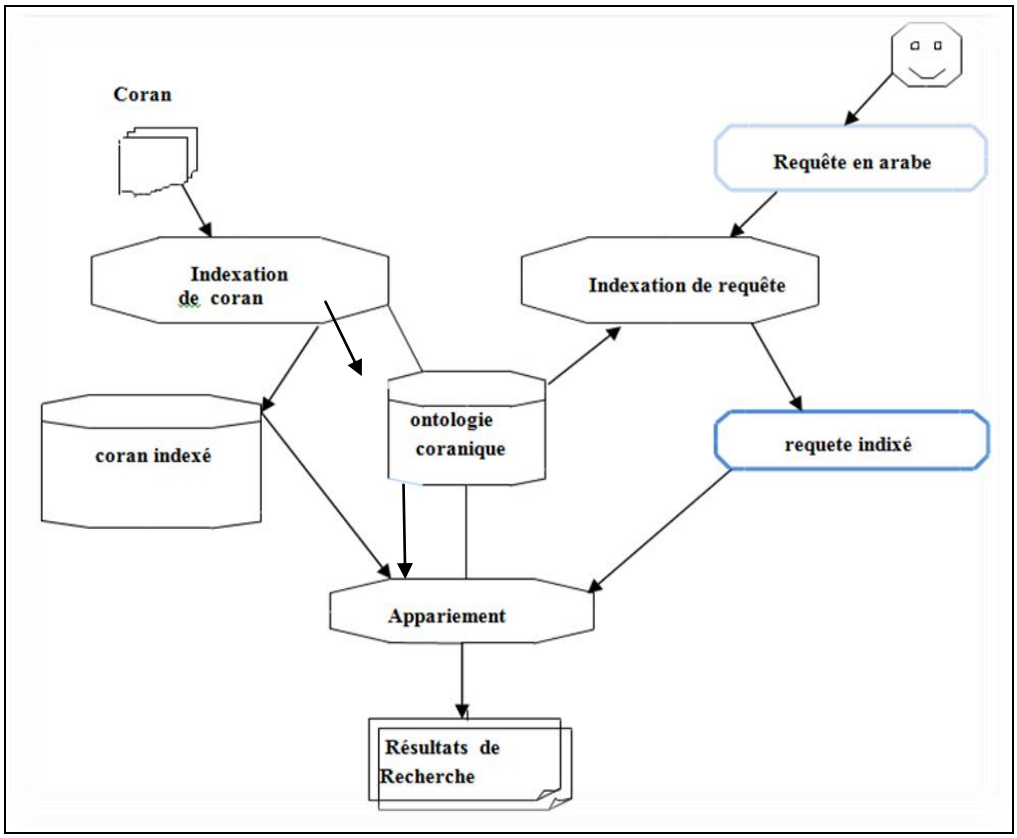


Figure 3.5: L'architecture de notre système.

Notre système de RI se compose de trois modules importants (voir la figu3.5au dessus) : le module d'indexation de coran : il de coranet de ontologie coranique une base de données d'indexes.

- Le module d'indexation de la requête : Il permet d'indexer la requête en utilisant l'ontologie coranique .
- Le module d'appariement Il permet de trouver et de filtrer les mots requête mots qui correspondent à la requête :

Notre système commence par l'indexation conceptuelle (ou sémantique) de coran en utilisant une ressource sémantique (ontologie coranique). Il génère comme résultat une base d'indexes. La requête de l'utilisateur est aussi indexée par interrogation de l'ontologie coranique afin de récupérer une liste des concepts reliés aux termes de la requête par des relations de synonymie, généralisation et spécialisation.

.

7. Les outils utilisés :

EasyPHP fut le premier package **WAMP** à voir le jour (1999). Il s'agit d'une plateforme de développement Web, permettant de faire fonctionner localement (sans se connecter à un serveur externe) des scripts PHP. EasyPHP n'est pas en soi un logiciel, mais un environnement comprenant deux serveursun serveur web Apache et un serveur de bases de données MySQL), un interpréteur descript (PHP), ainsi qu'une administration SQL phpMyAdmin. Il dispose d'une interface d'administra-tion permettant de gérer les alias (dossiers virtuels disponibles sous Apache),et le démarrage/arrêt desserveurs. Il permet donc d'installer en une seule fois tout le nécessaire au développement local du PHP. Par défaut, le serveur Apache crée un nom de domaine virtuel (en local) `127.0.0.1` ou localhost. Ainsi, quand on choisit « Web local » dans le menu d'Easy.

NET BEANS : NetBeans est à l'origine un EDI Java. NetBeans fut développé à l'origine par une équipe d'étudiants à Prague, racheté ensuite par Sun Microsystems. Quelque part en 2002 .

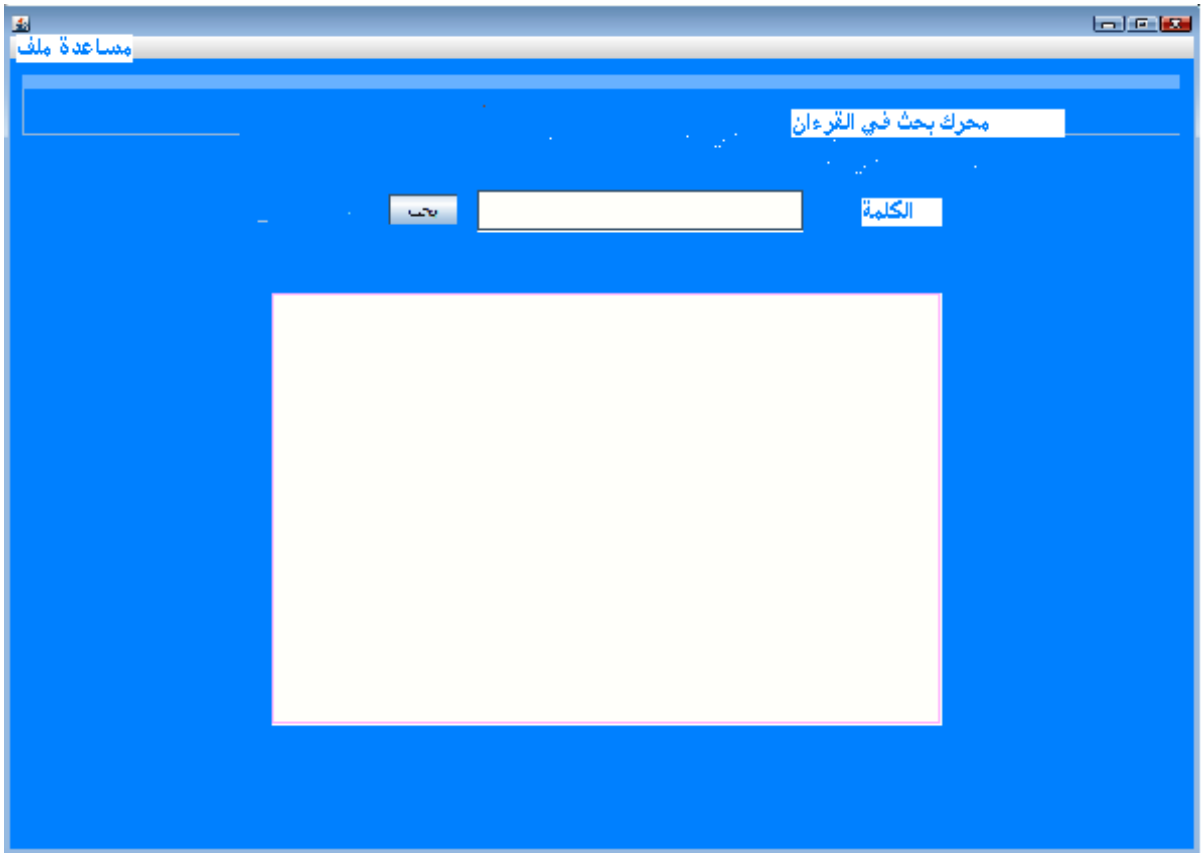
Sun a décidé de rendre NetBeans open-source. Mais NetBeans n'est pas uniquement un EDI Java. C'est également une plateforme, il permet donc de développer des applicati-ons. La licence de NetBeans permet de l'utiliser gratuitement à des fins commerciales ou non.

8. Implémentation de notre système

La réalisation de notre système reste une tache très difficile à cause de l'absence de l'extension de la base de donnée de l'ontologie coranique car elle est très volumineux. Pou cela on a essayé de donner quelques exemples pour éclaircir un peu l'idée pratique de notre travail. En utilisant EASY PHP MY SQL pour la création de la base de donnée et afficher quelque exemples et NETBEANS pour la création de l'interface de notre application, en reliant les deux (la base de données et l'interface) par le port (JDBC, ODBC).
Dans notre système l'utilisateur lance sa requête (le mot qu'il veuille chercher au coran) Le résultat c'est l'affichage de tous les termes en rapport avec la requête, avec leurs sens et leurs versets.

La figure suivante (voire **figure 3.6**) propose le menu principal de notre application :

Figure 3.6 : Le menu principal de notre moteur de recherche



9. les résultats obtenus :

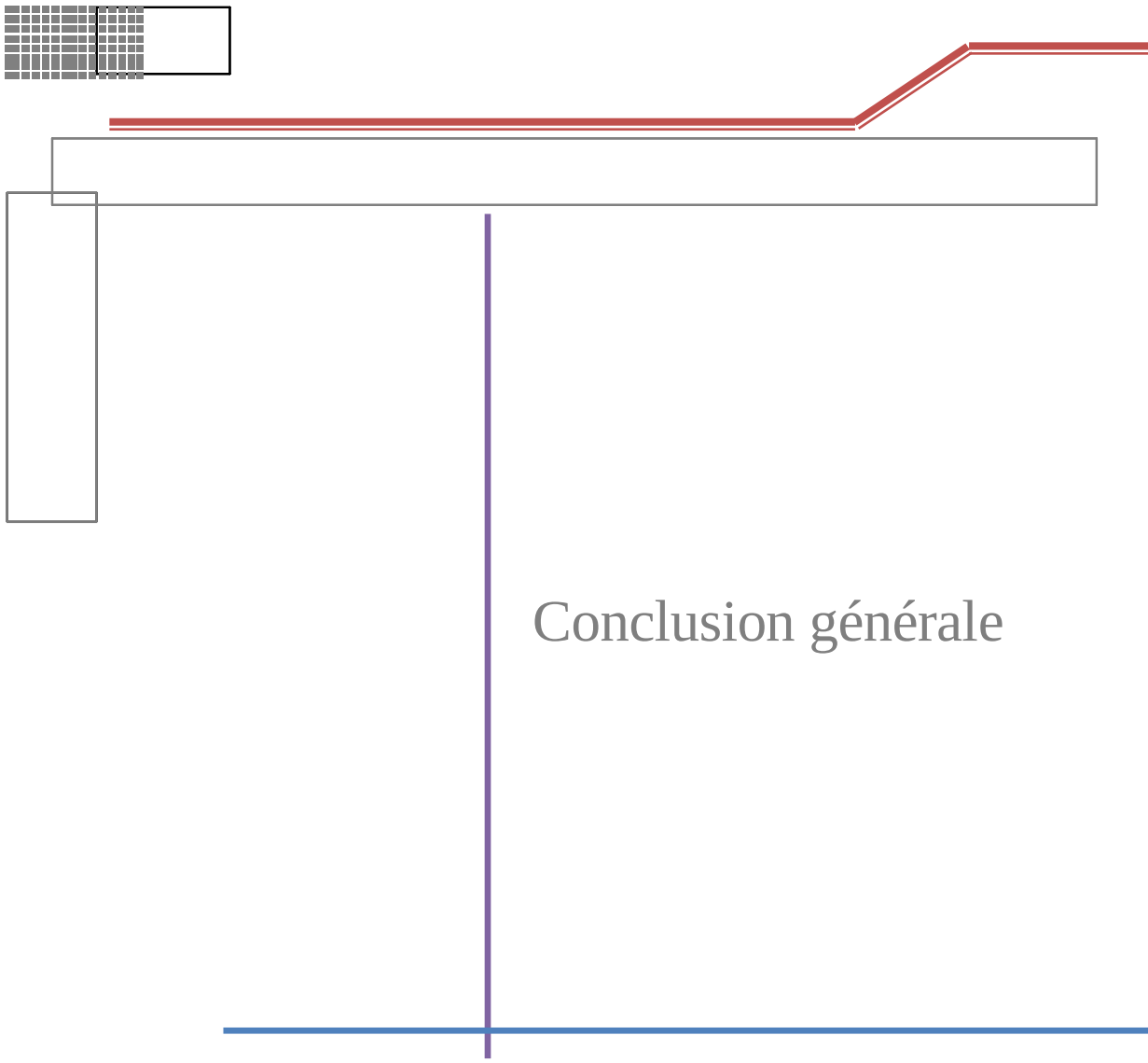
requête (mots) recherché	Mots trouvés	Versets de coran	Sens
يونس	يونس	فَلَوْلَا كَانَتْ قَرْيَةٌ ءَامَنَتْ فَنَفَعَهَآ إِيمَٰنُهَآ إِلَّا قَوْمَ يُونُسَ لَمَّآ ءَامَنُوا۟ كَشَفْنَا عَنْهُمْ عَذَابَ ٱلْخِزْيِ فِى ٱلْحَيَوٰةِ ٱلدُّنْيَا وَمَتَّعْنَٰهُمْ إِلَىٰ حِينٍ' يونس 97 "	
	ذَا ٱلنُّونِ	"وَذَا ٱلنُّونِ إِذ ذَّهَبَ مُغَٰضِبًا فَظَنَّ أَن لَّن نَّقْدِرَ عَلَيْهِ" الانبياء 86	صاحب الحوت
معدن	فضة	قَوَارِيرَا۠ مِن فِضَّةٍ قَدَّرُوهَا تَقْدِيرً" 15 "	
	حديد	وَأَنزَلْنَا ٱلْحَدِيدَ فِيهِ بَأْسٌ شَدِيد " الحديد 24	
	قطر	قَالَ ءَاتُونِىٓ أُفْرِغْ عَلَيْهِ قِطْرًا ″ الكهف 95	نحاس

Tableau 3.7 : les résultats de notre application

10. Conclusion :

Au terme de ce dernier chapitre nous avons proposé la réalisation d'un SRI basé sur l'utilisation d'une ressource conceptuelle (ontologie coranique).

Les résultats confirment le fait que la recherche conceptuelle des mots permettent d'élargir le champ de recherche qui permet d'enrichir les informations de l'utilisateur, pour cela on a l'appeler recherche intelligente.



Conclusion générale

Conclusion générale

Suite aux grandes quantités d'informations diffusées en coran, nous sommes obligés à développer des systèmes de recherche plus intelligents pour répondre aux besoins des utilisateurs. L'avènement des technologies de traitement automatique de la langue et le développement des ressources sémantiques facilite ces taches en grande partie.

Notre travail s'inscrit dans le cadre d'intégration des ontologies aux différents niveaux des systèmes de recherche d'informations arabe, autrement dit, l'indexation de coran et des requêtes utilisateurs en utilisant l'ontologie coranique.

Ce mémoire nous a permet de travailler sur une nouvelle ressource sémantique (récemment développée) qui est l'ontologie coranique. Il nous a permet, entres autres d'étudier ces caractéristiques et ces contenus et par la suite exploiter la sémantique derrière cette ressource afin d'estimer l'apport de son utilisation dans un SR I coranique.

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

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

.

Bibliographie

[Abderrahim,]

[Abderrahim, 10] Abderrahim M. A. : Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. 2ème Conférence Internationale sur l'Informatique et ses Applications CIIA'2009, Saida - Algérie, (2009)

[Ambroziak, 97] J. Ambroziak. Conceptually assisted Web browsing. In Sixth International World Wide Web conference, Santa Clara, CA. full paper available online at http://www.scope.gmd.de/info/www6/posters/702/guide2.html.(1997)

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

[Aufaure & al., 07] Aufaure M. A., Soussi R., Baazaoui H., « SIRO: On-line semantic information retrieval using ontologies ». 2nd International Conference on Digital Information Management, ICDIM'07, p. 321- 326, (2007)

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

[Bachimont, 00] B. Bachimont Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. Charlet et al. (eds), Ingénierie des Connaissances; Evolutions récentes et nouveaux défis, Eyrolles, pp. 305-323. (2000)

[Baziz & al., 05] M. Baziz, Boughanem M., Aussenac-Gilles N., Chrisment C., Semantic Cores for Representing Documents in IR, Proceedings of the 20th ACM Symposium on Applied Computing, pp 1020-1026, ACM Press ISBN: 1-58113-964-0, (2005).

[Baziz, 05] M. Baziz, Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de doctorat, université Paul Sabatier (2005)

[Baziz, 02] M. Baziz, Application des Ontologies pour l'Expansion de Requêtes dans un Système de Recherche d'Informations. Rapport de DEA Informatique de l'Image et du Langage (2IL). Université Paul Sabatier & Institut National Polytechnique de Toulouse, (2002)

[Benjamins & al., 99] Benjamins R., Fensel D., Decker D., Gomez Perez A., (K A)2, building ontologies for the internet : amid-term report, Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure, pp 1-24, (1999).

[Bordogna & al., 00] Flexible Querying of Structured Documents. Proceedings of the Fourth

International Conference on Flexible Query Answering Systems(FQAS),(2000)

[Boughanem & al., 04] Boughanem M., Kraaij W., Nie J-Y., Modèles de langage pour la recherche d'information. Les systèmes de recherche d'informations, pp 163-182, Hermès, 2004.

[Boughanem & Tamine, 04] Boughanem, M. and Tamine, L. (2004). Connexionisme et génétique pour la recherche d'information. Dans : Les systèmes de recherche d'informations, (Eds.), Hermes-Lavoisier, Lavoisier.

[Boughanem, 00] M. Boughanem : Contribution à la Formalisation et à la Spécification des Systèmes de Recherche et de Filtrage d'Information. Habilitation à Diriger les Recherches, Université Paul Sabatier de Toulouse. (2000)

[Bourne & Anderson, 79] C. Bourne, B.Anderson : DIALOG LabWorkbook, second edition, Lockheed Information Systems, PaloAlto, Californie (USA), (1979)

[Callan, 96] J. P. Callan, Document filtering with inference networks. Proceedings of ACM SIGIR 96, pages : 262-269, (1996).

[Chen & Wang, 95] Chen S. & J.Y Yang: Document Retrieval Using Knowledge Based Fuzzy Information Retrieval Technique. IEE Transactions on Systems, Man and Cybernetics, 793-803. (1995)

[Croft & al., 92] James P. Callan, W. Bruce Croft, Stephen M. Harding: The INQUERY Retrieval System. DEXA 1992: 78-83.

[Deerwester & al, 90] S. Deerwester, S. Dumais, S. Furnas, G. Landauer & R. Harshman : Indexing by Latent Semantic Analysis : Journal of the American Society for Information Science, 391-407

[Dumais, 94] S. Dumais : Latent Semantic Indexing (LSI), TREC3 report. In Proceedings of the 3rd Conference on Text Retrieval Conference. (1994)

[Dumais, 95] S.Dumais Latent Semantic Indexing (LSI), TREC-3 Report In proceedings of TREC-3, pages : 319-230, 1995

[Euzenat, 02] Desmontils E., Jaquin C., Indexing a Web site with a terminology oriented ontology, The Emerging Semantic Web, I. Cruz S. Decker, J. Euzenat, D.L. McGuinness (Eds.), IOS Press, ISBN 1-58603-255-0, pp 181-197, (2002).

[Farquhar et al., 97] A. Farquhar, R. Fikes, and J. Rice : The ontoloingua server : A tool for collaborative ontology construction. Journal of Human-Computer Studies, 46 :707-728, (1997).

[Fürst, 02] F. FÜRST, Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation. Thèse de doctorat, École Polytechnique de l'Université de Nantes (EPUN) (2004)

[Gargouri & al., 05] Gargouri Y., Lefebvre B., Meunier J. G., « Domain and Competences

Ontologies and Their Maintenance for an Intelligent Dissemination of Documents ». Book Series Lecture Notes in Computer Science. Book MICAI 2005: Advances in Artificial Intelligence ISBN 978-3-540-29896- 0. Category Knowledge Representation and Management, p. 90-97, (2005).

[Gligorov & al., 07] Gligorov R., ven K ate W., Aleksovski Z., van Harmelen F., Using

Google Distance to Weight Approximate Ontology Matches. Proceedings of the 16th international conference on World Wide Web, pp 767 - 776, (2007).

[Gomez, 99] A.Gomez, « Développements récents en matière de conception, de

maintenance et d'utilisation d'ontologies ». in 3èmes rencontres Terminologie et intelligence artificielle TIA (1999).

[Gruber, 93] T. R. Gruber, "Toward Principles for the design of Ontologies used for

Knowledge Sharing," in Proc of International Workshop on Formal Ontology, Padova, Italy, March (1993).

[Guarino et al., 99] Guarino, N., C. Masolo, and G. Vetere, OntoSeek: Using Large

Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs,. National Research Council, LADSEBCNR: Padavo, Italy.(1999)

[Guha & al., 03] Guha R. V., McCool R., Miller E. : Semantic search, Proceedings of the

12th International World Wide Web Conference, pp 700-709, (2003)

[Haav & Lubi, 01] Haav H. M., Lubi T.L., A Survey of Concept-based Information

Retrieval Tools on the Web, Proceedings of the 5th East-European Conference ADBIS, Vol 2, pp 29-41, (2001).

[Haines & al., 93] David Haines, W. Bruce Croft: Relevance Feedback and Inference

Networks. SIGIR : 2-11. (1993)

[Hearst & Karadi, 97] Hearst M.A, Karadi C., Cat-a-Cone: an interactive interface for

specifying searches and viewing retrieval results using a large category hierarchy, Proceedings of the 20th International conference on Research and Development in Information Retrieval, SIGIR, pp 246-257, (1997).

[Hearst, 97] M.A. Hearst, C. Karadi, Cat-a-Cone: an interactive interface for

specifying searches and viewing retrieval results using a large category hierarchy, In Proceedings of the 20th International conference on Research and Development in Information Retrieval, SIGIR, pp 246-257, (1997).

[Hernandez & al., 08] Hernandez N., Hubert G., Mothe J., Ralalason B. : RI et Ontologies –

Etat de l'art 2008, RAPPORT INTERNE N° IRIT/RR—2008-14—FR

JUILLET (2008)

[Hernandez, 05] Hernandez N, Ontologies de Domaine pour la Modélisation du Contexte

en Recherche d'Information. Thèse de doctorat, université Paul Sabatier (2005)

[Hlaoua, 07] L. Hlaoua : Reformulation de Requêtes par Réinjection de Pertinence

dans les Documents Semi-Structurés. Thèse de doctorat, université Paul Sabatier (2007)

[Hofman, 99] T. Hofman, Probabilistic Latent Semantic Indexing : In the Proceedings

of the 22nd Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval, August, Buckley USA (1999)

[Horacio & al.,06] Horacio, Rodríguez, Sabri, Elkateb, William, Black, Piek, V ossen,

Adam, Pease, Christiane, Fellbaum: Building a WordNet for Arabic, http://www.adampease.org/Articulate/publications/LREC.pdf (2006)

[Karbasi, 07] S. Karbasi : Pondération des termes en Recherche d'Information:

Modèle de pondération basé sur le rang des termes dans les documents. Thèse de doctorat, université Paul Sabatier (2007)

[Kiryakov, 04] Kiryakov A., Popov B., Terziev I., Manov D., Ognyanoff D., Semantic

annotation, indexing, and retrieval, Journal of Web Semantics, 2(1), pp 49-79, (2004).

[Kim & al., 07] Kim H., Park C. S., Park J. Y., Jung B., Lee Y. J., « A Multimedia

Content Management and Retrieval System Based on Metadata and Ontologies ». IEEE International Conference on Multimedia and Expo, p. 556 – 559, (2007).

[Koczy & al., 98] Baranyi, P.; Gedeon, T.D.; Koczy, L.T.; Intelligent information retrieval

using fuzzy approach. Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on Volume 2, 11-14 Oct. Page(s):1984 - 1989 vol.2 Digital Object Identifier 10.1109/ICSMC.1998.728188. (1998)

[Köhler & al., 06] Köhler J., Philippi S., Specht M., Rüegg A., « Ontology based text

indexing and querying for the semantic web ». Knowledge-Based Systems, Vol 19, Issue 8, December 2006, p. 744 – 754, (2006)

[Kohonen, 89] T.Kohonen, Self-Organisation and Associative Memory 3rd Edition,

Springer V erlag, Berlin, pages : 80-89, (1989).

[Kompaoré, 08] Y. Kompaoré, Fusion de systèmes et analyse des caractéristiques

linguistiques des requêtes : vers un processus de RI adaptatif. Thèse de doctorat, université Paul Sabatier (2005)

[Kraaij, 04] Kraaij, W., Variations on Language Modeling for Information Retrieval.

[Kais Dukes,] http://corpus.quran.com/ontology.jsp

Phd thesis, University of Twente. (2004)

[Lancaster, 68] F.W. Lancaster, Evaluation of the MEDLARS Demand Search Service.
Bethesda, Md.: The National Library of Medicine (1968)

[Lelu & François, 92] Lelu, A. and François, C. Information retrieval based on neural
unsupervised extraction of thematic fuzzy clusters. In Fifth International Conference, Neural Networks and their Applications: NEURO NIMES, pages 93–104. (1992)

[Lenat et al., 90] D.B. Lena and R.V. Guha :Building large knowledge-based
systems.Representation and inference in the Cyc project, Addison-Wesley, Reading,Massachusetts, USA, 1990.

[Losee, 98] Losee, R. M., Text Retrieval and Filtering: Analytic Models of
Performance. K luwer, Boston. (1998)

[Lucarella & Morara, 91] D. Lucarella & R. Morara : FIRST Fuzzy Information Retrieval
Systems. Journal of Information Science, 81-91. (1991)

[Luhn, 57] Luhn, H., A statistical approach to mechanized encoding and searching
of literary information. IBM, 1(4):309–317. (1957)

[Maron & Kuhns, 60] Maron, M. and Kuhns, J. On relevance, probabilistic indexing and
information retrieval. Journal of the Aassociation for Computing Machinery, 7 :216–244. (1960)

[Mihalcea, 00] R. Mihalcea, D.I. Moldovan, Semantic Indexing using WordNet Senses,
In Proceedings of ACL Workshop on IR & NLP, (2000)

[Miller, 02] L. Miller, A. Seaborne, A. Reggiori, Three implementations of squishql,
a simple rdf query language, In Proceedings of the International Semantic Web Conference, pp 423-435, (2002).

[Miller, 90] MILLER G.A., BECKWITH R., FELLBAUM C., GROSS D. &
MILLER K.: " Introduction to WordNet : An on-line lexical database ", Journal of Lexicography, n°3, pp.235-244, (1990).

[Moldovan & al., 99] Moldovan D., Harabagiu S., Pasca M., Mihal-cea R., Goodrum R., Girju
R., Rus V., LASSO: A tool for surfing the answer net. Proceedings of the 8th Text Retrieval Conference (TREU-8), (1999).

[Mothe, 94] J. Mothe : Modèle Connexionniste pour la Recherche d'Information,
Expansion dirigée de requêtes et apprentissage. Thèse de Doctorat, université Paul Sabatier. (1994)

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

[Ponte & al., 98] Ponte, J. and Croft, W., A language modelling approach to information
retrieval. In Proceedings of the 21st Annual International ACM SIGIR

Conference on Research and Development in Information Retrieval, pages 40–48. (1998).

[Resnik, 95] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995, pp. 448–453, (1995).

[Robertson & al., 99] S.E. Robertson, S. Walker, M. Beaulieu: OKAPI at TREC 7: Automatic Adhoc Filtering, VLC and Interactive Track, In Proceedings of the 7th Text Retrieval Conference TREC7, (1999)

[Robertson, 77] S. E. Robertson: The probability ranking principle in IR. Journal of Documentation, 33 (4), 294-304. (1977)

[Sabri & al.,06] Sabri, Elkateb, William, Black, Piek, Vossen, David, Farwell, Adam, Pease, Christiane, Fellbaum.: Arabic WordNet and the Challenges of Arabic. http://www.mt-archive.info/BCS-2006-Elkateb.pdf (2006)

[Salton & al., 83] Salton, G., Fox, E., and Wu, H., Extended Boolean information retrieval. Communications of the ACM, 26(11):1022–1036. (1983)

[Salton, 71] G. Salton : The Smart Retrieval System : Experiments in Automatic Document Processing, G. Salton Editor, Prentice Hall Inc., Englewood Cliffs, New Jersey,(1971)

[Salton, 83] Salton, G., Introduction to modern information retrieval. New York, McGraw-Hill. (1983)

[Salton, 89] Salton, G., Automatic text processing : The transformation, analysis and retrieval of information by computer. Addison-Wesley publishing, MA. (1989)

[Shaw & al., 97] W.M. Shaw, R. Burgin & P. Howell : Performances Standards an Evaluation in IR test Collections: Cluster based Retrieval Models. Information Processing and Management, 1-14, (1997)

[Swartout & al., 97] Swartout (B.), Patil (R.), Knight (K.) and Russ (T.): Towards Distributed Use of Large-Scale Ontologies. Spring Symposium Series on Ontological Engineering, Stanford University, CA, p. 138-148.(1997)

[Tamine, 00] L. Tamine, Optimisation de requêtes dans un Système de Recherche d'Information. Thèse de doctorat, université Paul Sabatier (2000)

[Tebri, 04] H. Tebri : Formalisation et spécification d'un système de filtrage incrémental d'information. Thèse de doctorat, université Paul Sabatier (2004)

[Tomassen & al., 06] Tomassen S. L., Gulla J. A., Strasunskas D., « Document Space Adapted Ontology: Application in Query Enrichment ». 11th International Conference on Applications of Natural Language to Information Systems. Springer, Klagenfurt, Austria, (2006).

[Turtle & Croft, 91] H. Turtle & W.B Croft : Evaluation of an Inference Network Based Retrieval Model. ACM Transactions on Information Systems July (1991)

[Turtle et al., 92] Howard R. Turtle, W. Bruce Croft: A Comparison of Text Retrieval Models. Comput. J. 35(3): 279-290 (1992)

[Vallet 2005] D. Vallet, M. Fernández, P. Castells, An Ontology-Based Information Retrieval Model, In Proceedings of the 2nd European Semantic Web Conference, pp 455-470, (2005).

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

[William & al.,06] William, BLACK, Sabri, ELKATEB, Horacio, RODRIGUEZ, Musa, ALKHALIFA, Piek, VOSSEN, Adam, PEASE, Christiane, FELLBAUM: Introducing the Arabic WordNet Project. http://www.globalwordnet.org/AWN/meetings/GWApaper.pdf (2006)

[William & al.,04] William, J., Black, Sabri, El-Kateb: A Prototype English-Arabic Dictionary Based on WordNet. http://www.fi.muni.cz/gwc2004/proc/95.pdf (2004)

[Wong & al, 85] S.K.M. Wong, W. Ziarko & P.C. N. Wong: Generalized Vector Space Model in Information Retrieval. In Proc of the 8th ACM SIGIR Conference on Research and Development, New-York USA, (1985)

[Xiaomeng & Atle, 06] Xiaomeng S., Atle J. G., « An information retrieval approach to ontology mapping ». Data & Knowledge Engineering, Vol. 58 Issue 1, p. 47-69, (2006).

[Zhao & al., 07] Zhao Y., Zhang J., Guan B., Hu J., Wang W., « The Development of Intelligent Retrieval Algorithm Ontology-based And Its Application in Bearing production Information System ». 11th International Conference on Computer Supported Cooperative Work in Design, 2007. CSCWD'07, p. 722 – 727, (2007).

Résumé :

Les approches de Recherche d' Information (RI) classique traitent les documents et les requêtes comme des « sacs de mots » sans syntaxe et sans sémantique. Elles traduisent l'existence des mots de la requête dans les documents.
La richesse de la langue arabe du coté syntaxique et sémantique rend les systèmes de recherche qui utilise cette langue plus au moins inefficace à la qualité des résultats obtenues par ces systèmes.
Ces dernières années, l'informatique a vu le développent d'un nouveau concept appelé ontologie, il concerne la prise en compte des aspects sémantique et conceptuelle utilisés pour résoudre les problèmes lexicales et sémantiques des langues afin d'améliorer les performances des systèmes de recherche d'informations.
L'idée de ce mémoire est d'exploiter une ontologie (ressource lexicale) pour indexer le coran et la requête de l'utilisateur afin d'améliorer les résultats de la recherche d'un système de RI. Les résultats confirment le fait que l'ontologie permet d'élargir le champ de recherche dans un SRI, ce qui permet d'enrichir les informations de l'utilisateur.

Mots Clés : TALN Arabe, Recherche d'Information Arabe, ontologie coranique , indexation conceptuelle.

Abstract:

Classical Information Retrieval (IR) approaches suffer from "bag of words "representation of documents and queries without any syntactic or semantic information. They evaluate the relevance of a document to a query by considering the frequency of query word within the document.

The richness of the Arabic in his side syntactic and semantic search makes the systems using that Language more or less ineffective in the quality of results obtained by these systems.
In recent years, the computer science show the developing of a new concept called ontology port behind them the semantic aspect used to resolve the problems of lexical and semantic languages to improve the performance of information retrieval systems.
The idea of this study is to exploit ontology (lexical resource) to indexing the Quran and the user query in order to improve the retrieval results. To test this approach, we propose the searching mechanism using Arabic Quran ontology. The results of experimentation confirm the fact that the ontology's improves the efficiency of the IR system.

Keywords: Arabic NLP, Arabic Retrieval Information, Quran ontology., Conceptual indexing.

ملخص:

إن نظم البحث عن المعلومات تتعامل مع الملفات و الطلبيات على أنها وعاء من الكلمات بدون معنى ولا نحو .فهي تعتمد على تعداد كلمات الطلبية في الملفات.

ان ثراء اللغة العربية من الناحية النحوية و البيانية يجعل انظمة البحث ذات فعالية محدودة من حيث النتائج .

لقد أدى التطور الحالي في المجال الدلالة إلى ظهور و استعمال ما يسمى الانطولوجيا في جميع الميادين، فهرسة القران و موضوع البحث بهدف تطوير نتائج محركات البحث .

الفكرة من وراء هذه الدراسة هو تطوير محرك بحث في القرءان بالاعتماد على الانطولوجيا القرآنية.

النتائج المتحصل عليها تؤكد تحسين مردودية محركات البحث التي تعتمد هذه الفكرة .

الكلمات المفتاحية :المعالجة الآلية للغة العربية، البحث عن المعلومة العربية ،الانطولوجيا القرآنية ، **فهرسة**.