MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE
LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et de l'Informatique
Département de Mathématiques et d'Informatique
Filière : Informatique

MEMOIRE DE FIN D'ETUDES
Pour l'Obtention du Diplôme de Master en Informatique
Option : **Ingénierie des Systèmes d'Information**

Les Entrepôts textuels et la Recherche d'information

Présenté par :

Aissani Asmaa

Encadré par:

M^elle^ MAGHNI SANDID Zoulikha.

Année Universitaire 2015/ 2016

Remerciements

Avant d'entamer ce rapport, je saisis cette occasion pour exprimer mes vifs remerciements, à Dieu le tout puissant de m'avoir donné courage et patience pour mener a bien ce travail, qu'il soit béni et glorifié, puis à toute personne ayant contribué, de près ou de loin, à la réalisation de ce travail.

Je remercie vivement Madame SANDID MAGHNI qui m' a encadrés avec patience et n'a pas cessé de m'encourager pendant la durée du projet, ainsi pour sa générosité en matière de formation et d'encadrement.et qu'elle m'a apporté lors des différents suivis, et la confiance qu'elle m'a témoigné.

Je tiens à remercier aussi nos professeurs de m'avoir incité à travailler en mettant à ma disposition leurs expériences et leurs compétences.

Je retiens à remercier également le membre de jury d'avoir accepté l'évaluation de ce travail.

Beaucoup m'en exprimer leurs soutiens chaleureux quand j'en ai eu besoin.j'exprimes également ma gratitude à eux, mes amis et ma famille je les remerciés infiniment.

Dédicaces

Avant tout, je dois rendre grâce à dieu de m'avoir donné le courage de terminer ce travail.

Je dédie ce travail à toute la famille

Tout d'abord à mon cher père qui est les yeux de ma vie, A ma chère mère (allah yarhamha), aux près de qui je trouve le réconfort et le

Repos, j'espère être à la hauteur de ses espérances et ne jamais la décevoir.

A tous mes frères et mes sœurs

A toute la famille Aissani et Mehidi

A, Chahinez ma chère sœur qui m'as toujours soutenus et m'aider dans mon parcours d'étude aux bons et mauvais moments, je suis très contente et ravis de toi ma puce,

je te souhaite une vie plein de joie et de bonheur avec ton marie.

.

A, Soumia plus qu'une amie, Imene, Fatiha, Nadjet, Lilia, Hafsa, et tous mes chère amies qui mon aider et m'encouragé.

A mes camardes de l'institut, compagnons de ces années d'études, j'espère qu'on restera camarade.

Asmaa.

Résumé

L'analyse multidimensionnelle basée sur les bases de données multidimensionnelles « BDM » factuelles numériques est une tâche bien maîtrisée de nos jours. Ces BDM sont souvent construites sur des données transactionnelles issues des systèmes d'information (SI) des entreprises « *les entrepôts de données* ». Cependant, seules 20% des données d'un SI sont des données transactionnelles et peuvent être traitées. Les 80% restants, la « paperasserie électronique », restent hors de portée de la technologie OLAP faute d'outils et de méthodes adaptées à la gestion de données textuelles. Ne pas prendre en compte ces données mène inévitablement à l'omission d'informations pertinentes durant un important processus de prise de décision voire l'inclusion de données non pertinentes générant ainsi des analyses approximatives ou erronées.

Dans le contexte de ces travaux, nous nous sommes focalisés sur les deux notions qui sont l'intégration des entrepôts textuels et leur modélisation multidimensionnelle. Pour mettre en place une solution d'intégration de ce type de données, nous avons eu recours aux techniques de recherche d'information (RI) et du traitement automatique du langage naturel (TALN).

Mots clés : *modélisation multidimensionnelle, entrepôt de données, entrepôts textuels, recherche de l'information.*

Sommaire

Liste des figures

Liste de tableaux

Introduction générale

Avec le développement de l'Internet, de plus en plus de documents textuels sont disponibles. Extraire de la connaissance ou analyser et interroger de tels volumes de données est un enjeu important. Ainsi, par exemple, les travaux menés autour de la fouille de textes ont proposé de nouvelles approches pour classer automatiquement des documents, rechercher les nouvelles tendances ou extraire de l'information dans des données textuelles. Plus récemment, de nouvelles approches fondées sur des cubes de textes proposent d'utiliser les technologies OLAP pour analyser et extraire de la connaissance

Dans un processus d'aide à la décision, si nous désirons retrouver les mots caractéristiques de chaque catégorie, nous pouvons utiliser des entrepôts de données. Dans un tel contexte, il est indispensable d'extraire pour chacune des catégories les termes les plus représentatifs en tenant compte du fait qu'il peut exister une hiérarchie entre les différentes catégories (c-à-d. la catégorie "vaccin" peut être divisée en "vaccin en Europe", "vaccin en Asie", "vaccin aux Etats-Unis", etc). Dans cet exemple, nous considérons qu'il existe une hiérarchie disponible. Toutefois une telle connaissance n'est pas forcément aisée à obtenir et sa définition n'est pas toujours caractéristique d'une réalité

Dans ce contexte, un nouveau concept est apparu permettant de gérer cette masse importante d'informations, à savoir : l'entrepôt de documents. C'est « une source d'informations orientées-sujets, filtrées, intégrées, historisées et organisées comme support d'un processus de recherche, d'interrogation ou d'analyse. » [**1**].

Actuellement, si l'intuition reste fondamentale dans le processus de prise de décision, elle ne suffit plus, il faut aussi pouvoir prendre ce que les anglophones appellent « la décision informée » (informed décision). A cette fin, les entreprises disposent aujourd'hui d'importants volumes d'informations qui sont généralement hétérogènes : données structurées (bases de données, fichiers…), données semi-structurées et multimédia (XML, HTML…).

Vu cette hétérogénéité des données, il est nécessaire de les contrôler, les homogénéiser, les organiser, les intégrer et enfin les stocker pour donner à leur utilisateur une vue intégrée et orientée métier ; ces informations factuelles et aussi textuelles constituent un facteur de savoir-faire et de connaissances. Ainsi, face à l'augmentation considérable du nombre de documents gérés par les entreprises, le besoin d'outils pour traiter automatiquement les informations documentaires s'est rapidement fait sentir par les entreprises

Des méthodes particulièrement adaptées à l'analyse du contenu textuel sont venues soutenir les tâches de recherche d'information. Le traitement automatique des langues naturelles (TAL) est l'un des domaines principaux de l'intelligence artificielle (IA). Les applications dans le domaine du traitement automatique des langues naturelles se sont considérablement étendues et cette tendance s'accroît continuellement, certains outils sont apparus tels que les moteurs de recherche, les systèmes d'extraction de l'information (IR : Information Retrieval) essayant d'apporter un accès flexible et aisé à l'information.

l'utilisation d'une conceptualisation externe telle que les ontologies représente une issue intéressante pour satisfaire le besoin en information de l'utilisateur. Il est clair que cette approche doit être spécifique.

ORGANISATION DE MEMOIRE

Ce mémoire est composé de quatre chapitres incluant une introduction et une conclusion générale.

Le chapitre 1 : On expose tout d'abord les notions de la recherche d'information et ses aspects théoriques et fonctionnels, les différentes approches et techniques relatives au fonctionnement et l'analyse en vue d'une indexation terminologique d'un system de recherche d'information ses étape, les modèles et les mesures de performances, ainsi l'amélioration de recherche conceptuels par l'ontologie " WordNet" comme aspect sémantique.

Le chapitre 2 : Est consacré à la description et la présentation des différentes notions des entrepôts, ses caractéristiques, le processus ETL, et la modélisation multidimensionnels En particulier, on décrit les entrepôts de document c'est-à-dire « textuels » pour servir à notre approche.

Le chapitre 3 : On poursuit avec un troisième chapitre qui porte sur la conception de notre application. Pour cela, on va faire une description de traitement de notre processus de recherche d'information d'après un corpus (texte) afin de dégager les mots-clés de recherche et mesurée la similarité entre eux, ensuite on montre la conception de notre ontologie.

On abords ensuite le modèle en étoile de notre entrepôt textuels selon notre corpus de démarche. Après on va concevoir le système de recherche pour la manipulation multidimensionnels de notre entrepôt textuels.

Le chapitre 4 : ce dernier chapitre est une vue sur intégration et implémentation d'un système de recherche d'information à base d'ontologie pour un entrepôt textuels. Enfin, on relèvera les points clés de réalisation du projet et on trace les perspectives de notre travail.

Chapitre 1 : La recherche d'information

1. Introduction

Une information est une donnée structurée dont un individu a besoin pour résoudre un problème particulier.il exprime donc son besoin sous forme de requête, ce besoin d'information peut s'exprimé sous deux types : [**1**]

- Type fermé (extraction d'Information, réponses aux questions);
- Type ouvert (recherche ad-hoc, documents non-structurés)

2. Recherche d'information(RI)

Le nom de « Recherche d'Information » fut donné par Calvin N. Mooers en conférence dédiée à ce thème « International Conference on Scientifc Information » qui s'est tenue en 1958 à Washington.

La recherche d'information abrégée en RI ou IR (Information Retrieval), est la science qui étudie la manière de répondre *pertinemment* à une *requête* en retrouvant de l'information dans un *corpus*. **[2]**

Dans cette définition, il y a trois notions clés : pertinence, requête, corpus.

- **La pertinence** : on dit un document *pertinent* c'est-à-dire celui qui doit contenir l'information ou l'utilisateur doit trouver les informations dont il a besoin. c'est sur cette notion que les systèmes de recherche d'information sont jugés (une notion complexe qui dépend de l'utilisateur, la requête). **[3]**

- **Requête :** Une requête exprime le besoin d'information d'un utilisateur

- **Le corpus** : est composé de documents (*texte*, des sons, des vidéos, pages web, des images) d'une ou de plusieurs bases de données, qui sont décrits par un contenu ou les métadonnées associées il peut se présenter sous un format brut ou semi-structuré. **[4]**

Il existe deux approches pour retrouver un document dans un corpus :

- **Approche naïve :** considère une requête comme une chaîne de caractères, et la compare avec tous les documents (recherche séquentielle).
- **Approche basée sur une indexation** : on construit une structure d'index qui permet de retrouver rapidement les documents incluant des mots demandés (le cas de notre étude). **[5]**

2.1. Processus de recherche d'information (SRI)

Le processus de recherche d'information a pour but la mise en relation des informations disponibles d'une part, et les besoins de l'utilisateur d'autre part. Il s'articule autour de deux étapes essentielles : les phases d'indexation et de recherche. **[6]** Le processus complet est représenté en Figure 1.

Ce processus est composé de trois fonctions principales **[7]** :

- L'indexation des documents et des requêtes.
- L'appariement document-requête, qui permet de comparer les documents et les requêtes.
- La fonction de modification, qui intervient en réponse aux résultats obtenus.

Il existe plusieurs SRI open source. Ils sont libres et distribués sous différentes licences : GPL (General Public License), apache, MPL (Mozilla Public License). D'ailleurs, ils possèdent une structure technique similaire telle que chacun d'eux est formé par un ensemble de packages relatifs au document, recherche, indexation et requête. Cependant, le processus de RI est différent d'un système à un autre en terme d'interaction entre les blocs constitutifs, structure d'index et modèle de recherche. **[9]**

Parmi les SRIs open source qui satisfaisant le besoin d'utilisateur : Zettair[1], Lucene[2], Indri[3], MG4J[4] et Terrier[5].

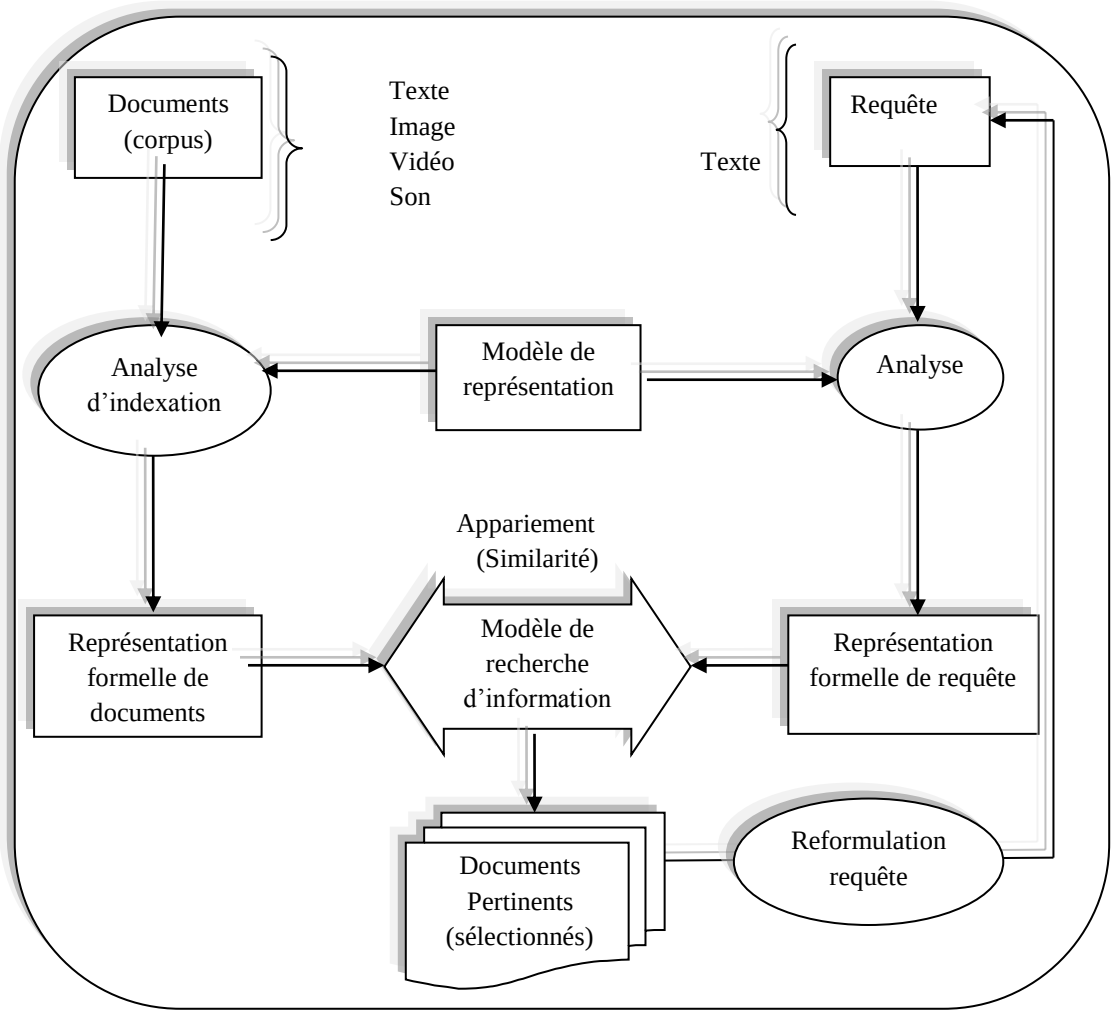


Figure 2.Processus en U d'un système de recherche d'information

[1] http ://www.seg.rmit.edu.au/zettair
[2] http ://lucene.apache.org/core/
[3] http ://ciir.cs.umass.edu/ strohman/indri
[4] http ://mg4j.dsi.unimi.it/man/manual/index.html
[5] http ://terrier.org/

2.2. Les difficultés de la recherche d'information

Le langage *"naturel"* pose plusieurs défis pour tout analyseur automatique ce qui mène a une problématique pour les SRIs car la différence des langages artificiels cause : **[10]**

Implicite

2.2.1. Implicite

Tout n'est pas dit dans les textes et leur compréhension requiert une importance connaissance sur le contexte et sur le monde : **[9]**

- Connaissance du langage et des **conventions langagières**
- Connaissance du **contexte**
- Connaissance du **monde**

2.2.2. Redondance

La langue offre de nombreuses façons de formuler le même contenu **[11]** :

- Au niveau lexical

Synonymie : vélo et bicyclette.

Hyperonymie et hyponymie : *véhicule* / *vélo* / *VTT*

Métonymie et homonymie : *pédale* / *pédalier* / *vélo.*

- Abréviations et sigles : s'il vous plaît et SVP, VTT et Vélo Tout Terrain…etc.

2.2.3. Ambiguïté

Un même énoncé peut souvent être interprété de différentes façons. La recherche d'informations est encore compliquée par le fait que **[11]** :

- Les mots peuvent jouer des rôles différents dans les textes.
- Les atomes de sens peuvent être des mots ou des groupes de mots (termes).

2.3. Indexation

L'étape d'indexation permet de réaliser le passage d'un document textuel (ou une requête) à une représentation exploitable par un modèle de RI par la construction de mots clés appelé langage d'indexation "*transformer des documents en substituts capables de représenter le contenu de ces documents*". **[9]**

Document textuel (ou requête) ⟶ Représentation exploitable par le SRI

Indexation

Le processus d'indexation peut se faire de trois façons, manuelle, automatique ou semi-automatique.

- ***Manuelle (contrôlé) :*** chaque document est analysé par un spécialiste du domaine ou par un documentaire. Cette indexation permet d'assurer une meilleure pertinence dans les réponses apportées par le SRI.

- ***Semi-manuelle (contrôlé) :*** le choix final revient au documentaliste, qui intervient souvent pour choisir d'autres termes plus significatifs.

- ***Automatique (libre):*** le processus d'indexation est entièrement informatisé.

Nous nous intéressant au plus au troisième type « l'indexation semi-manuelle » pour notre SRI, on présente par suite les étapes de cette approche

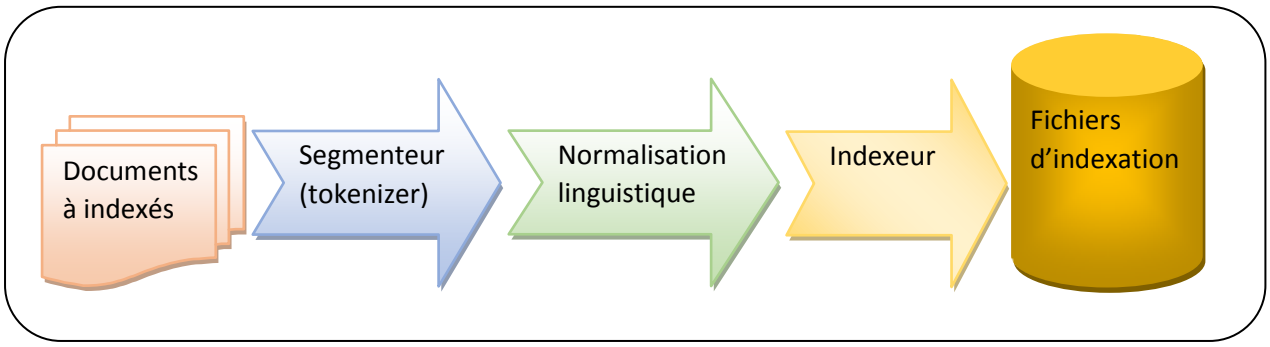


Figure 2. Schéma des étapes d'indexation

2.3.1. Les étapes d'indexation

2.3.1.1. Segmentation

Identification des unités élémentaires (phonèmes, morphèmes, mots, etc.). Pour l'écrit, des mots et des phrases. C'est l'étape initiale indispensable pour tout travail sur le texte c'est au niveau de découpage (Tokénisation). On obtient des *mots*, ou des ***termes***, ou des ***tokens*** **[9]**

2.3.1.2. Normalisation

Cette phase peut contenir plusieurs étapes, dans ce qui suit nous allons expliquer les étapes les plus importantes et les plus utilisées, elle traites plusieurs niveaux :

Niveau morphologique **:** reconnaissance du mot, la morphologie interprète comment les mots sont structurés et quels sont leurs rôles dans la phrase. Cette analyse consiste à une segmentation du texte en unités élémentaires auxquelles sont attachées des connaissances dans le système : une fois cette segmentation effectuée, ce n'est plus le texte qui est manipulé, mais une liste ordonnée d'unités. **[13]**

Niveau lexical : réduction du mot à sa forme canonique « lemmatisation ».

Ces deux niveaux se résument en trois fonctions principales [3][4]:

- ✓ Le lemme : (La lemmatisation) : Dans cette phase on substitue les différents termes par leur racine, le passage à la forme canonique n'est pas obligatoire, mais permet à l'utilisateur de ne pas avoir à respecter le pluriel ou les formes conjuguée. Elle permet aussi d'éviter d'avoir à indexer un mot sous plusieurs formes (connaitre, connaissances, connaissance …).

- ✓ La racine : (La racinisation, « stemming » en anglais) : La racine s'obtient par une dérivation (paradigme dérivationnel); elle consiste à rechercher la forme tronquée d'un mot à partir de laquelle peuvent être reconstruites ses différentes variantes morphologiques. Cette opération peut être réalisée assez simplement, en utilisant un algorithme comme l'algorithme de Porter, pour l'anglais. Par contre, elle peut entraîner une perte de sens, car la racine extraite peut être commune à des mots se rapportant à des concepts différents : Les mots ***port***, ***port***es (ouverture) et ***port***era ont la même racine ***port*** mais se rapportent à trois concepts différents, la racinisation permet d'augmenter le rappel mais peut diminuer la précision.

- ✓ Extraction de mot composé : Mots non obligatoirement successifs qui doivent être reconnus comme formant une seule entité. Il est important de reconnaître les mots composés car ce sont des unités de sens. Par exemple : « ***arbre à cames*** » ou « ***pomme de terre*** ».

Niveau syntaxique : C'est une partie de la grammaire qui traite la manière dont les mots peuvent se combiner pour former des propositions et de l'enchaînement des propositions entre elles. Cela consiste à associer, à la chaîne découpée en unités, une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités " l'analyse syntaxique entretient d'étroites relations avec la logique"**. [12]**

Niveau sémantique : niveau de la reconnaissance des concepts, le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les niveaux précédemment énoncés.

Ces deux niveaux concernent des traitements sur des règles dépondant de la langue, affectent les opérations suivantes**[3][4]** :

- ✓ *Elimination des mots vides* ***:*** Les mots vides sont des mots qui permettent de lier entre eux les mots d'une phrase pour la structurer (les articles, les conjonctions de coordination, les verbes auxiliaires, etc.). Ces mots ne portent pas de sens, ils ne peuvent pas constituer des index il faut donc les éliminer, cependant il ne faut pas oublier de tenir compte de certain mots vides qui auraient pour homographes des mots significatifs (ex : or pour l'anglais).
- ✓ *Extraction des entités nommées :* ce sont des mots ou des groupes de mots qui désignent des personnes, des organisations, des dates, des lieux, etc. Par exemple, si un texte contient l'expression : « ***5 juillet 1962*** » il est plus intéressant de l'indexer globalement par cette date plutôt que les trois termes: « ***5*** », « ***juillet*** » et « ***1962*** ».Si de plus l'indexeur est capable de reconnaître que c'est la date de l'indépendance de l'Algérie, l'indexation sera encore plus précise.

2.3.1.3. Production de fichiers d'indexes :

Après le passage du contenu textuelles par les différents traitements et les niveaux d'analyse on produit les fichiers d'indexes, cet établissement est nécessaires pour l'obtention des indexes mais ça diffère selon les techniques suivis ou apportée sur les documents.

2.3.2. Les techniques de création des tables d'index

Dans le but de répondre le plus rapidement à une requête, des structures de stockage particulières sont nécessaires pour mémoriser les informations sélectionnées lors du processus d'indexation, les plus répandues sont : les fichiers inverses ou les tableaux de suffixes**. [13]**

2.3.2.1. Les fichiers inverses

Dans sa forme la plus simple, un index est tout simplement une structure qui nous donne, pour chaque mot trouvé dans un corpus, la liste des documents où il se trouve. Un index inversé peut prendre plusieurs formes. Certains index peuvent ne donner que la liste des documents où les mots apparaissent, d'autres vont aussi donner la position des mots dans les documents

Exemple :

La	**D1**
Vie	**D1**
Est	**D1**
Belle	**D1, D2**

2.3.2.2. Les tableaux de suffixes

D'abord on fait le tri des suffixes par ordre alphabétique, puis on stocke le tableau des suffixes.

Exemple :

Soit la chaîne « laval », et de ses suffixes :

1 : laval

2 : aval

3 : val

4 : al

5 : l

On peut trier comme suit :

4 : al

2 : aval

5 : l

1 : laval

3 : val

Il suffit alors de stocker le tableau « 4, 2, 5, 1,3 » qui forme le tableau de suffixes.

Il existe d'autres structures couramment utilisées pour l'indexation :

- Les arbres de recherche digitaux : il s'agit d'arbres multi-branches servant au stockage des chaînes de caractères. Ils sont capables retrouver n'importe quelle chaîne de caractères en un temps proportionnel à leur longueur. Chaque arc de l'arbre est étiqueté avec une lettre. Pour chercher un mot dans l'un de ces arbres, il faut partir de la racine et descendre dans l'arbre en suivant les bons arcs.
- Les arbres de recherche binaire.
- Les B-arbres.
- Les tableaux triés.
- Les tables de hachage.

2.4. Les modèles de recherche d'information

Le processus d'indexation aboutit sur la production des fichiers d'index selon un modèle de représentation spécifique. Ce dernier doit être le même avec celui de la représentation de la requête et inclut entre autre une fonction de mesure de similarité. On distingue trois grandes catégories de modèles classiques. La caractérisation formelle de ces modèles est vue comme un quadruplet **[D, Q, F, R (qi, dj)]** tel que **[14] [15]:**

D: est l'ensemble des représentations des documents du corpus;

Q: est l'ensemble des représentations des requêtes de l'utilisateur;

F: est le Framework de modélisation des représentations des documents (ensembles des opérations sur les représentations des documents);

R (qi, dj): est la fonction de classement qui associe au couple (qi, dj) un réel représentant le degré de rapprochement entre qi et dj.

- **Modèle booléen :** les documents et les requêtes sont représentés comme des *ensembles de termes indexent.*
- **Modèle vectoriel :** les documents et les requêtes sont représentés comme des *vecteurs* dans un espace à *n* dimensions.
- **Modèle probabiliste :** la modélisation des documents et des requêtes est basée sur la *théorie des probabilités.*

Nous orientons l'attention du lecteur vers **[16]** pour de plus de détails sur ces modèles.

3. Evaluation d'un SRI

La performance des SRIs se mesure par le classement des documents retournés par scores décroissants, avec de nombreux critères pouvant intervenir *(date du document, qualité/notoriété de la source, liens commerciaux, etc.)* **[17]**

- ✓ ***Evaluation par l'utilisateur***, qui dépend : de la pertinence (relevance) des documents retournés et la quantité de bruit *(cela dépend du point de vue de l'utilisateur, de ses connaissances, etc.)*, du temps de réponse du système, ou encore de l'ergonomie du système *(présentation des résultats, mode d'interaction)*

- ✓ ***Évaluation automatique,*** campagnes d'évaluation de systèmes de recherche d'information, c'est une comparaison booléenne des documents retournés avec des réponses *« idéales » (un document en fait partie ou pas)* qui traite deux notions principales de **précision** et de **rappel**

  Ces deux mesures sont obtenues en partitionnant l'ensemble des documents constitués par le SRI en deux catégories :

  - Documents pertinents.
  - Documents non pertinents

Plus les réponses du système correspondent à celles que l'utilisateur espère, mieux est le système. **[18][19]**

Rappel : capacité du système à fournir en réponse tous les documents pertinents.

Rappel = Nombre de documents pertinents sélectionnées / Nombre de documents pertinents

Précision : capacité du système à ne fournir que des documents pertinents en réponse.

Précision = Nombre de documents pertinents sélectionnées / Nombre de documents sélectionnées

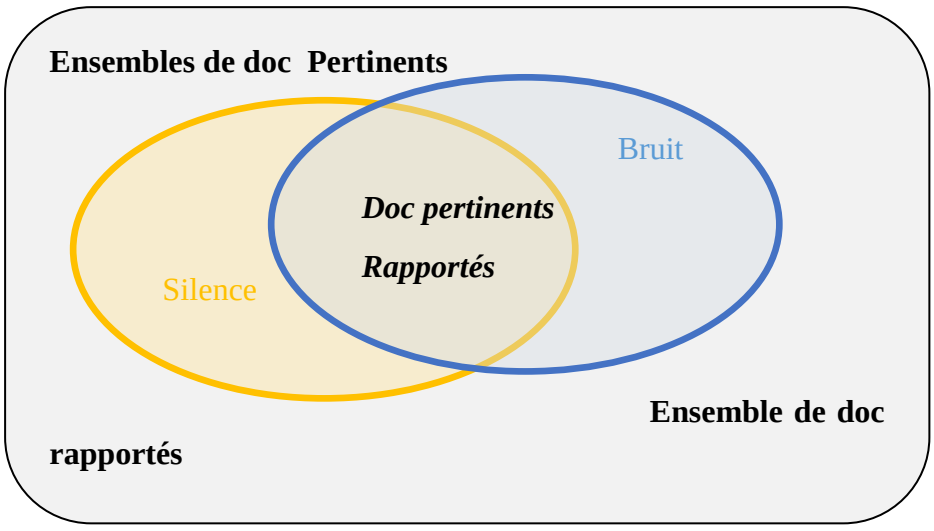


Figure 3. Qualité d'une réponse

4. La sémantique pour les SRI

Les systèmes de recherche d'information classiques traitent la requête de manière à optimiser les temps de recherche et l'identification de documents selon des critères d'appariement entre les mots contenus dans les requêtes utilisateurs (et uniquement ceux-là) et ceux des documents. **[20]**

La recherche sémantique a pour objectif d'améliorer la précision de recherche par la compréhension de l'objectif de recherche et la signification contextuelle des termes tels qu'ils apparaissent dans l'espace de données recherché, que ce soit sur le Web ou dans un système fermé, afin de générer des résultats plus pertinents.

Les méthodes de RIS visent à s'affranchir des problèmes via le passage au niveau conceptuel. Elles reposent souvent sur l'utilisation d'une ressource sémantique externe (thésaurus ou ontologie). En 1968, Salton constate déjà que l'utilisation d'un thésaurus (Harris Synonym) permet d'améliorer les performances, à condition que les termes utilisés pour l'enrichissement soient validés manuellement par l'utilisateur. Il constate également que l'expansion automatique utilisant l'ensemble des termes possibles, dégrade ces performances (Salton, 1968) [22].Ces méthodes ne sont pas comparables parce qu'il est souvent difficile de dissocier les fonctionnalités sémantiques pour pouvoir les évaluer séparément [**21**].

L'utilisation des ontologies pour l'enrichissement des termes en concepts cela peut constituer une solution (parmi d'autres) pour résoudre le problème des variations sémantiques **[23],** en effet, les ontologies offrent des ressources sous la forme de relations sémantiques, ils permettent d'étendre le champ de recherche d'une requête**. [24]**

L'idée est donc d'exploiter le contenu de l'ontologie « wordnet » **[25]** et d'un analyseur morphologique pour pallier le problème des variations lexicales, autrement dit, on interroge wordnet pour récupérer des mots différent lexicalement, mais reliés à ceux de la requête initiale par des relations sémantiques, telles que la synonymie, la généralisation et la spécialisation.

4.1. Concepts fondamentaux

Une base d'une ontologie représente un réseau sémantique définit par des termes et relations entre eux on présente dans ce qui suit les notions clés de cette dernière.

4.1.1. Mot

C'est toute séquence de caractères délimitée par deux séparateurs (blanc ou autre marqueur de séparation, tel que la ponctuation) ou unité minimale de signification appartenant au lexique appelé lexème **[**26].

4.1.2. Terme

Un terme est une expression possédant un sens unique pour un domaine particulier **[**26].

4.1.3. Concept

Un concept est «l'idée générale et abstraite que se fait l'esprit humain d'un objet de pensée concret ou abstrait, et qui lui permet de rattacher à ce même objet les diverses perceptions qu'il en a, et d'en organiser les connaissances » Larousse[1].

4.2. WordNet

WordNet est une base de données lexicale, il offre un grand dépôt d'éléments lexicaux anglais appelés «synsets». WordNet a été conçu pour établir des relations entre les quatre types grammaticaux du mot (POS): nom, verbe, adjectif et adverbe (voir Tableau 1) [27]

Catégorie	Mots	Concepts	Total Paires (mot-sens)
Nom	117 798	82 115	146 312
Verbe	11 529	13 767	25 047
Adjectif	21 479	18 156	30 002
Adverbe	4481	3 621	5580
Total	155 287	117 659	206 941

Tableau 1. Le nombre de mots et de concepts dans WordNet

On peut considérer WordNet comme un graphe ou un réseau sémantique, souvent qu'on qualifie d'ontologie légère (Light Ontology), ou chaque nœud représente un concept du monde réel.

4.2.1. Synset

Une signification particulière d'un mot dans un type de point de vue est appelé un concept/sens elle est la composante atomique sur laquelle repose le système. Chaque sens d'un mot est dans une *synset* différente. WordNet manipule les unités lexicales par un ensemble de synonymes.

4.2.2. Relations dans WordNet

Une relation est définie comme une notion de lien entre les entités, souvent exprimé par un terme ou un symbole littéral ou autre. Deux relations fondamentales interviennent dans WordNet, notamment celle entre les

- ✓ «*word form*» appelés *relations lexicales* par exemple : la synonymie
- ✓ «*word meaning*» appelés relation sémantiques (par exemple : l'hyperonymie).

5. Conclusion

Ce premier chapitre a porté essentiellement sur l'étude des SRI de manière générale, On a présenté l'architecture des systèmes de recherche d'information notamment l'appariement document/requête, les étapes essentielles d'une bonne indexation puis les différents modèles

[1] Larousse *http://www.larousse.fr/dictionnaires/français*

et stratégies utilisés lors de la mise en œuvre d'un SRI et enfin on a parlé d'une approche de l'aspect sémantique présenté par l'ontologie « WordNet ».

Il en ressort que chacun de ces modèles ou stratégies contribue en partie à la résolution des problèmes incohérents à la recherche d'information : perception du besoin en information, représentation du sens véhiculé par les documents, formalisation de la pertinence etc

Chapitre 2 : Les entrepôts textuels

1. Introduction

Les entreprises disposent aujourd'hui d'importants volumes d'informations qui sont généralement hétérogènes : données structurées (bases de données, fichiers…), données semi-structurées et multimédia (XML, HTML…).

Vu cette hétérogénéité des données, il est nécessaire de les contrôler, les homogénéiser, les organiser, les intégrer et enfin les stocker pour donner à leur utilisateur une vue intégrée et orientée métier ; ces informations factuelles et aussi textuelles constituent un facteur de savoir-faire et de connaissances.

2. Les entrepôts de données

Un entrepôt de données (Data Warehouse) est un gigantesque tas d'informations épurées, organisée, historisée provenant de plusieurs sources de données, servant aux analyses et à l'aide à la décision. L'entrepôt de données est l'élément central de l'informatique décisionnelle. Il est le meilleur moyen pour modéliser l'information pour des fins d'analyse.

2.1. Les caractéristiques des entrepôts de données

D'après **W. H. Inmon (1996)** : « Le Data Warehouse est une collection de données *orientées sujet*, *intégrées*, *non volatiles* et *historisées*, organisées pour le support d'un processus d'aide à la décision ».

Le data warehousing : désigne les processus de construction et d'utilisation des entrepôts de données. Ces derniers doivent êtres **[28]**:

- ✓ **Orientées sujet :** les données sont organisées autour de sujets principaux (produits, clients, ventes, etc.) .Pour la modélisation et l'analyse des données ce qui fournit une vue simple et concise autour d'un sujet particulier en excluant les données inutiles pour le processus d'aide à la décision.

- ✓ **Intégrées** : les entrepôts sont des sujets d'entreprise qui requiert une intégration de données sûres, consistantes et complètes, les sources peuvent être multiples et hétérogènes qui doivent être mises en forme et unifiées afin d'avoir un état cohérent.

- ✓ **Non volatiles :** afin de conserver la traçabilité des informations et des décisions prises, les informations stockées au sein de l'entrepôt de données ne peuvent être supprimées.

✓ **Historisées :** point de vue de l'entrepôt de données est plus étendu puisqu'il traite les valeurs des données d'une perspective historique, un référentiel temps doit être associé aux données afin de permettre l'identification dans la durée de valeurs précises.

2.2. Architecture générale d'un entrepôt de données

Les entrepôts de données sont généralement intégrés dans un système d'aide à la prise de décision où l'on distingue deux espaces de stockage : l'entrepôt de données et les magasins de données. Une architecture du processus décisionnel est représentée dans la *Figure 4.* Architecture d'un entrepôt de données.**[28]**.

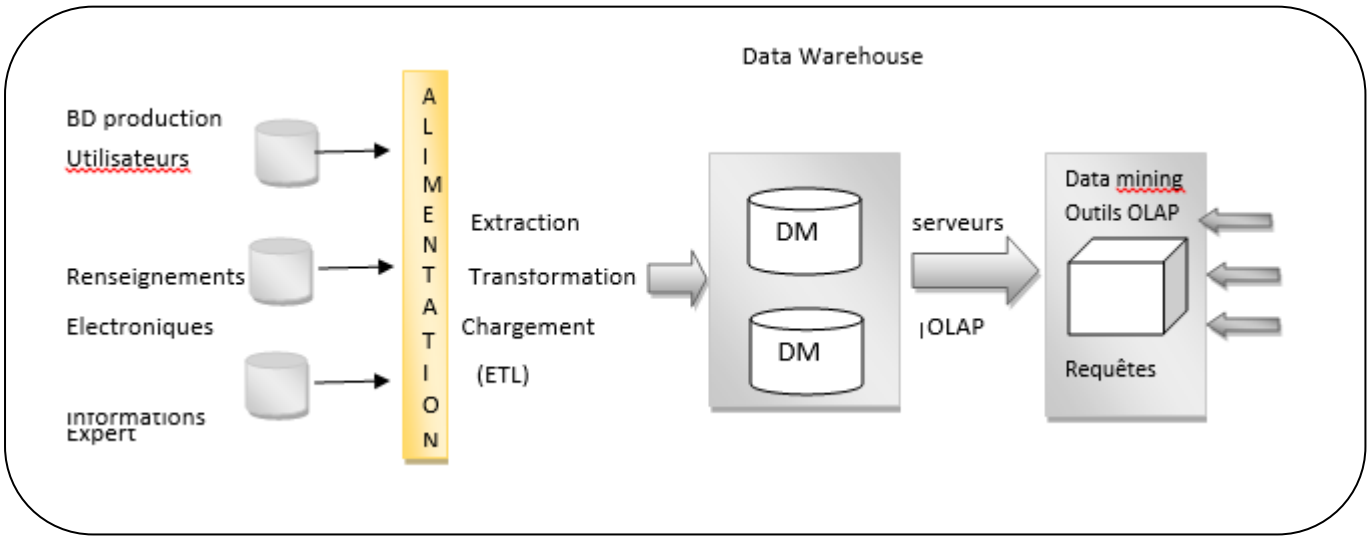


Figure 4. Architecture d'un entrepôt de données.

Cette architecture comporte quatre niveaux **[28]** :

- *Extraction des données* : Elle résout les problèmes liés à l'hétérogénéité et à la distribution des sources pour intégrer les données hétérogènes. De plus, elle extrait les données pertinentes pour la prise de décisions afin de générer l'entrepôt de données.
- *Entrepôt de données* : C'est le premier espace de stockage du système d'information décisionnel. Il est organisé selon un modèle facilitant la gestion des données.
- *Magasin de données* : (Data Mart) C'est le second espace de stockage du système d'information décisionnel. Il correspond à un extrait des données de l'entrepôt destiné à une classe d'utilisateurs, et structure ces données via une modélisation spécifique permettant des analyses multidimensionnelles ou On peut faire des divisions par fonction (un data Mart pour les ventes, pour les commandes, le personnel …) ou par sous-ensemble organisationnel (un data Mart par succursale).
- *Restitution et analyse* : C'est l'exploitation des données contenues dans un magasin afin de restituer au décideur des résultats, généralement agrégés, à travers des outils spécifiques d'analyse.

2.3. Alimentation ETL

L'alimentation est la procédure permet de transférer des données du système opérationnel vers l'entrepôt de données en les adaptant. Personne ne remplit l'entrepôt au sens de la saisie de données car ce dernier est alimenté par les données extraites des bases situées en-dessous. Un moniteur est chargé de garder la cohérence des données et de répercuter vers l'entrepôt toutes les mises à jour des bases opérationnelles. Chaque moniteur envoie ses données vers un intégrateur qui va alimenter l'entrepôt de données. Il vérifie et valide les données, les met en forme afin de garder une cohérence entre les données issues de bases différentes, les analyse et les décrit, afin que les utilisateurs puissent interroger l'entrepôt avec les mots de leur métier.

Donc l'alimentation d'entrepôt de données c'est la problématique de l'ETL (Extracting Transforming and Loading). On peut résumer le processus ETL en :

- **L'extraction des données,** en accédant aux différentes bases et sources de données de l'entreprise,
- **La transformation**, en développant les codifications, résolvant les liens, changeant et uniformisant les différents formats de fichiers d'origine dans un format unique compatible avec le datawarehouse,
- **Le chargement**, pour alimenter les datawarehouses et datamarts, en contrôlant la cohérence des données.

3. La modélisation multidimensionnels

Lorsqu'on fait un schéma de BD pour un système d'information classique, on parle de termes de tables et de relations. Mais les systèmes d'aide à la décision, emploient des bases de données multidimensionnelles (BDM), qui permettent aux décideurs d'avoir une vision des performances d'une entreprise. Pour modéliser les BDM, des structures multidimensionnelles ont été définies permettant la représentation de sujets d'analyse, appelés faits et d'axes d'analyse, appelés dimensions. **[30]**

3.1. Dimention

On entend par dimension les axes avec lesquels on veut faire l'analyse, ils sont composées d'attributs, agencés de manière hiérarchique, qui modélisent les différents niveaux de détails (granularité) ces dernières. Il peut y avoir une dimension client, une dimension produit, etc. **[22]**

3.2. Fait

Les faits, en complément aux dimensions, sont sur quoi va porter l'analyse. Ce sont des tables qui contiennent des informations opérationnelles qui relatent la vie de l'entreprise. On aura des tables faits pour les ventes (chiffre d'affaires, quantités commandées …) par exemple ou sur les stocks (nombre d'exemplaires d'un produit en stock, niveau de remplissage du stock…). **[22]**

4. Les différent modèles du DW

Il existe différents modèles de représentation multidimensionnelles selon les données et le sujet d'analyse on distincts ces derniers :

4.1. Modèle en étoile

Une étoile est une façon de mettre en relation les dimensions et les faits qui contiennent les données à analyser dans un entrepôt. Le principe est que les dimensions sont directement reliées à un fait (schématiquement, ça fait comme une étoile).

On présente ci-dessus deux exemples de modèles en étoiles pour l'exploiter par suite (modèles en constellation)

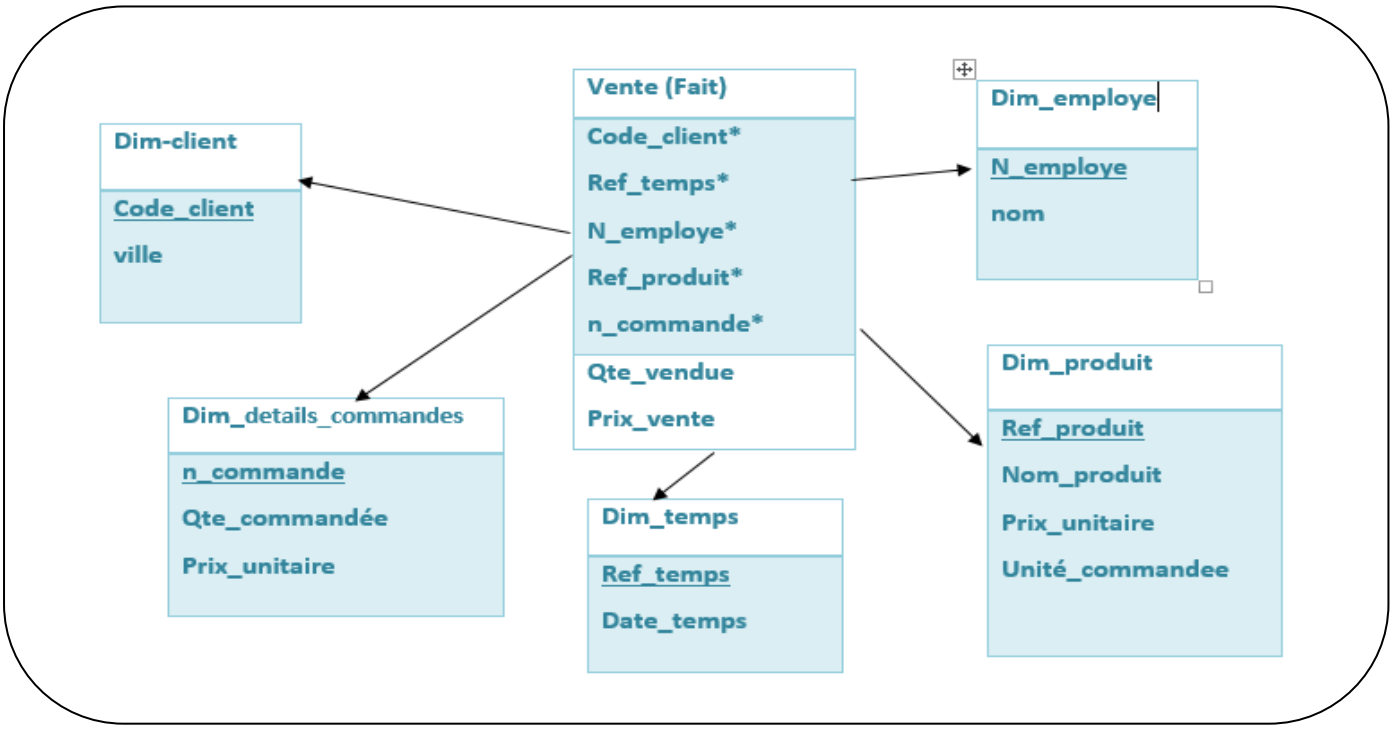


Figure 5. Exemple 1 d'un modèle en étoile(fait :vente)

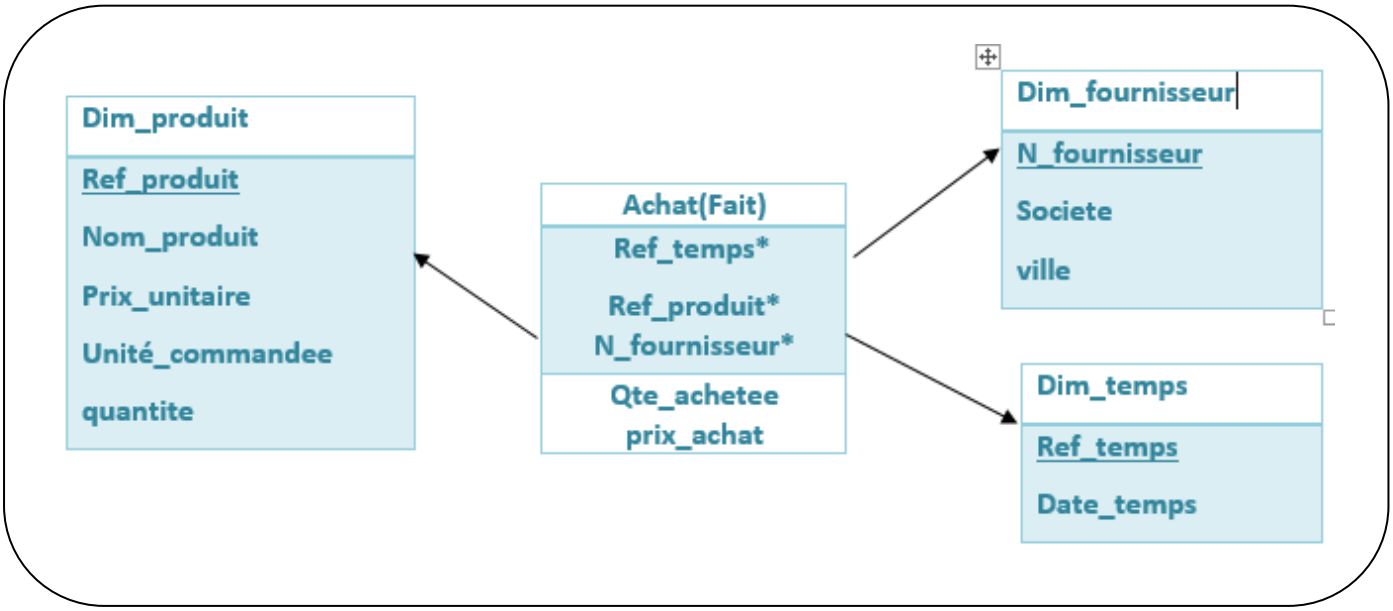


Figure 6. Exemple 2 d'un modèle en étoile (fait achat)

C’est claire que la table de fait contient ce qu'on appelle des " mesures ", des champs (numériques pour la plupart) sur lesquels l’analyse s’exploite, on peut y trouver le montant nettes (prix_achat, prix_achat), les quantités vendues et achetées, ou encore d’autre. Pour analyser une ligne de fait par produit par exemple, il faut qu'il y ait une relation entre cette ligne et la dimension produit.

Les tables de dimension contiennent les éléments qu'utiliseront les décideurs pour voir la table de faits. Les utilisateurs pourront ainsi apprécier les montant des ventes par catégorie, par client, ou meme le Qte_achetee pour un fournisseur pour un produit donnée (pour voir si ce produit est rentable et demande), calculer le coût de revient d'un produit par rapport aux activités des fournisseurs, etc.

4.2. Modèle en flocons

Un autre modèle de mise en relation des dimensions et des faits dans un entrepôt de données. le principe étant qu’il peut exister des hiérarchies de dimensions et qu’elles sont reliées aux faits. **[31] [32]**

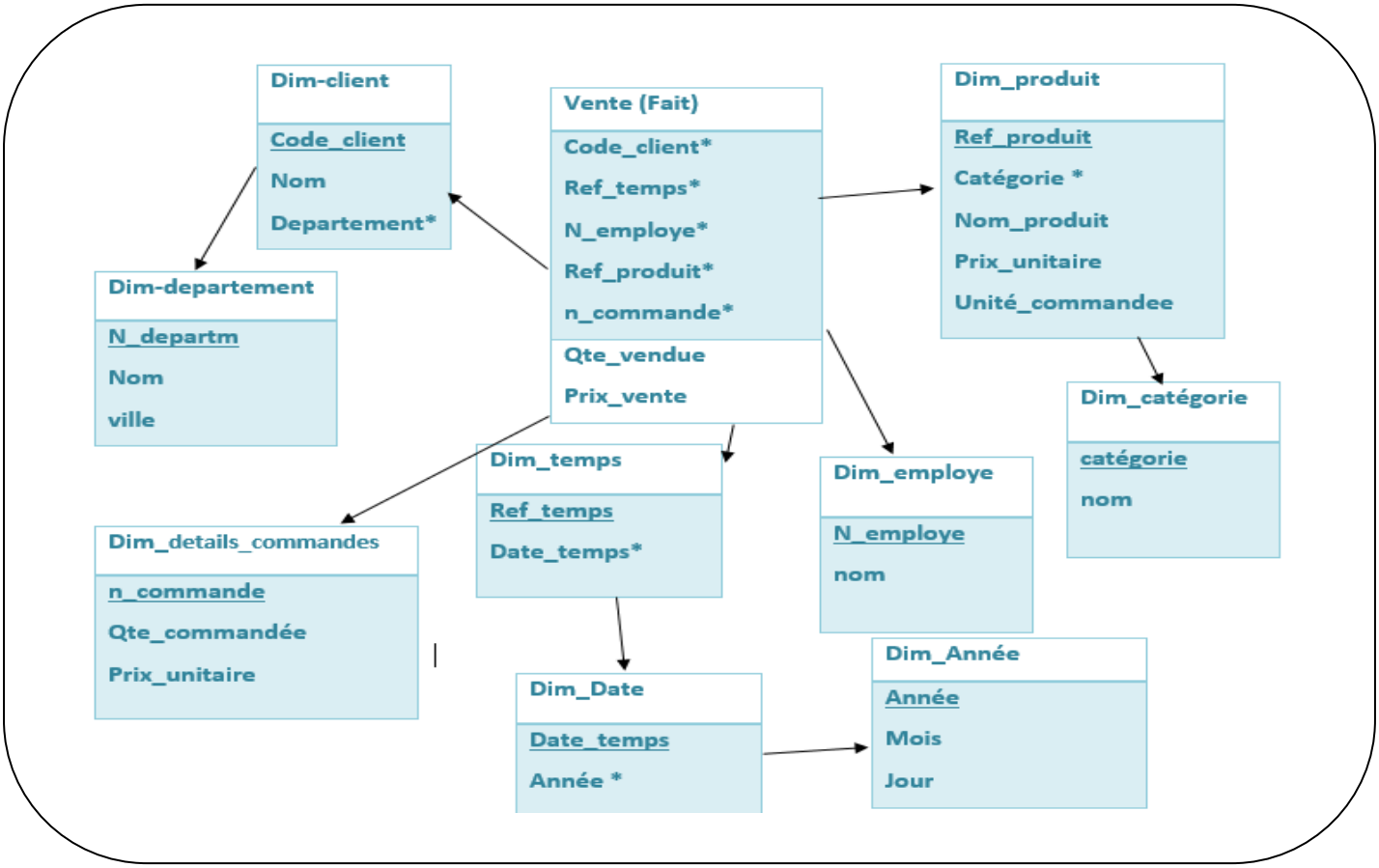


Figure 7. Exemple d’un modèle en flocons

La modélisation en flocon étant une variante de la modélisation en étoile, on prend le même cas avec la même analyse.
Il faut savoir que la modélisation en flocon existe pour des raisons de performances. En effet, des dimensions de plusieurs millions de lignes peuvent poser des problèmes de lenteur lors de l'exploitation des données.
Le principe de la modélisation en flocon est de créer des hiérarchies de dimensions, de telle manière à avoir moins de lignes par dimensions ou la performance prime sur la structure par la définition des degrés de granularité ou l'information seras plus précise plus spécial pour tel axes d'analyse ou encore plus fine.

4.3. Modèle en constellation

Ce modèle permet de représenter plusieurs tables fait en partageant les dimensions. C'est une série d'étoiles ou de flocons reliées entre eux par des dimensions en commun. **[31] [32]** ; Un environnement décisionnel idéal serait une place ou il serait possible de naviguer d'étoile en étoile, de constellation en constellation et de DataMart en DataMart à la recherche de l'information si précieuse.

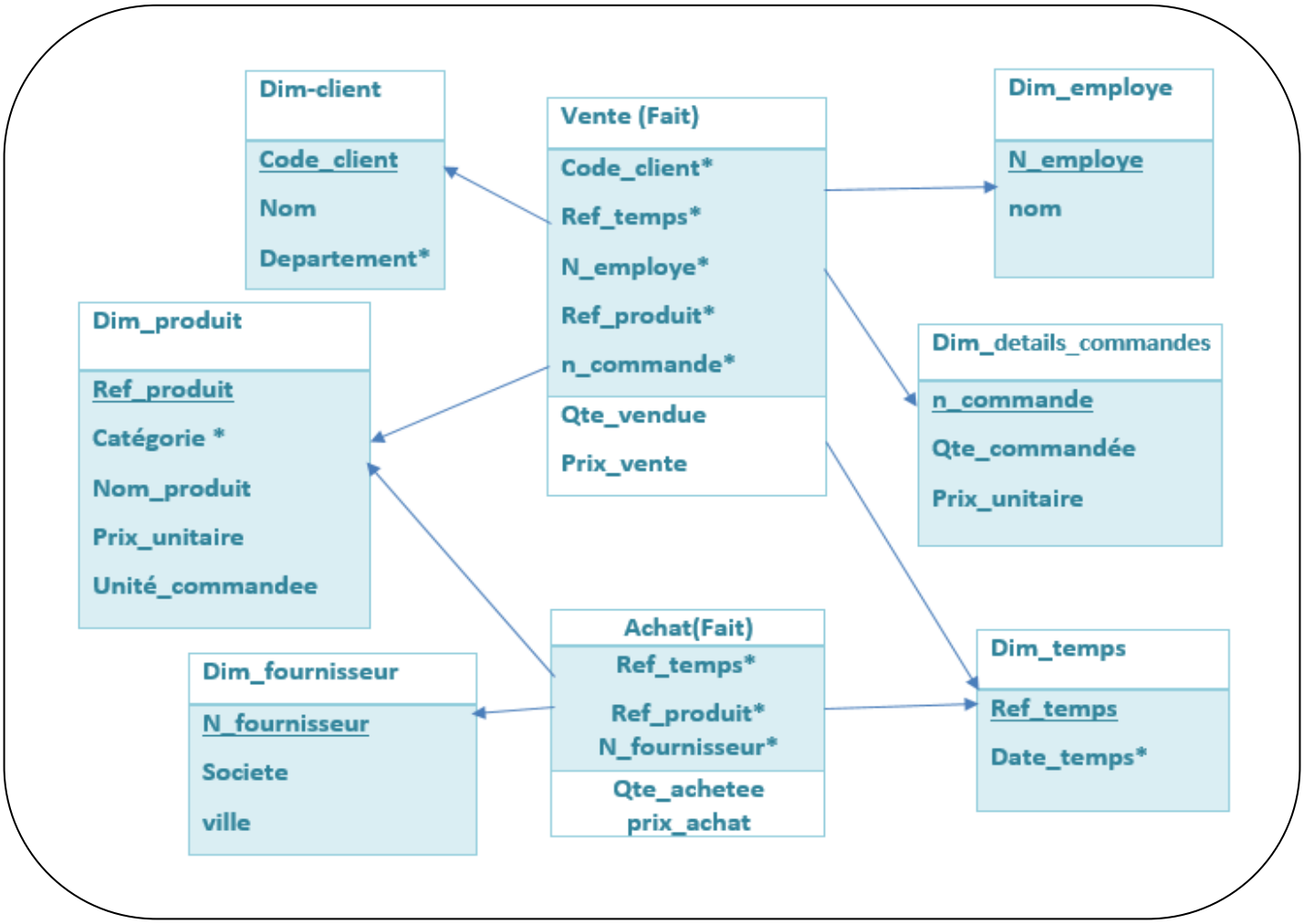


Figure 8. Exemple d'un modèle en constellation

Un des indicateurs clés d'une bonne conception d'entrepôt est la grosseur des constellations. En effet, plus la constellation est grosse, plus cela veut dire que vous avez réutilisé vos dimensions, et qui dit réutilisation de dimension, dit dimensions complètes, centralisées et avec une vue orientée entreprise.

L'analyse multidimensionnelle basée sur les bases de données multidimensionnelles « BDM » factuelles numériques est une tâche bien maîtrisée de nos jours **[33].** Ces BDM sont souvent construites sur des données transactionnelles issues des systèmes d'information (SI) des entreprises. Récemment, le format XML [1] a fourni un vaste environnement permettant l'échange et la diffusion de documents au sein des systèmes d'information des entreprises ou bien sur le Web pour les données textuelles en un format XML qui sont désormais des sources envisageables pour les systèmes OLAP. **[30]**

Un entrepôt de documents XML fournit un environnement de stockage des données peu structurées : c'est le lieu d'entreposage des documents XML orienté-documents issus des sources de données externes et internes.

La construction d'un entrepôt de documents XML est une tâche délicate surtout que, généralement, ces documents possèdent des structures hétérogènes même au sein d'un même domaine donc l'unification de ces structures est nécessaire pour masquer leur hétérogénéité **[34].** De ce fait, on propose par suite une autre approche pour organiser les données sous forme de modèle compréhensible pour une analyse efficace**.**

5. Les entrepôts de données textuels

Actuellement, si l'intuition reste fondamentale dans le processus de prise de décision, elle ne suffit plus, il faut aussi pouvoir prendre ce que les anglophones appellent « la décision informée » (informed décision).Ainsi, face à l'augmentation considérable du nombre de documents gérés par les entreprises, le besoin d'outils pour traiter automatiquement les informations documentaires s'est rapidement fait sentir par les entreprises

Dans ce contexte, un nouveau concept est apparu permettant de gérer cette masse importante d'informations, à savoir : l'entrepôt de documents. C'est « une source d'informations orientées-sujets, filtrées, intégrées, historisées et organisées comme support d'un processus de recherche, d'interrogation ou d'analyse. » **[35].**

D'après cette définition, les données d'un entrepôt doivent être organisées par sujets, jugées pertinentes, intégrées en provenance de multiples sources et historisées, c'est à dire garder leur évolution dans le temps, ainsi que leurs différentes versions ; ce dernier point fera l'objet de ce papier.

Les systèmes d'analyse en ligne OLAP (On-Line Analytical Processing) permettent aux analystes d'améliorer le processus de prise de décision. Ces systèmes facilitent la consultation et l'analyse de données économiques, statistiques ou scientifiques agrégées et historisées via une structuration adaptée au sein de bases de données multidimensionnelles **[36]** Les systèmes

[1] XML, Extensible Markup Language, de http://www.w3.org/XML/.

d’aide à la décision, emploient des bases de données multidimensionnelles (BDM), qui permettent aux décideurs d’avoir une vision des performances d’une entreprise

Ainsi, face à l’augmentation considérable du nombre de documents gérés par les entreprises, le besoin d’outils pour traiter automatiquement les informations documentaires s’est rapidement fait sentir par les entreprises.

5.1. La conception d’un entrepôt de données textuelles

Les entrepôts de données textuelles est un cas particulier des entrepôts de données complexes sauf que les données sont des textes ou des ensembles de textes (documents,corpus).

Pour représenter les textes dans un corpus ou un entrepôt de données, il faut passer par l’étape de prétraitement des textes qui base sur la présentation du texte, plusieurs représentations existe, tel que : la représentation ace N-gramme et avec la notion de sac de mots, dans ce qui suit on va étudier la représentation la plus utilisable, c’est la représentation en sac de mots WordNet.

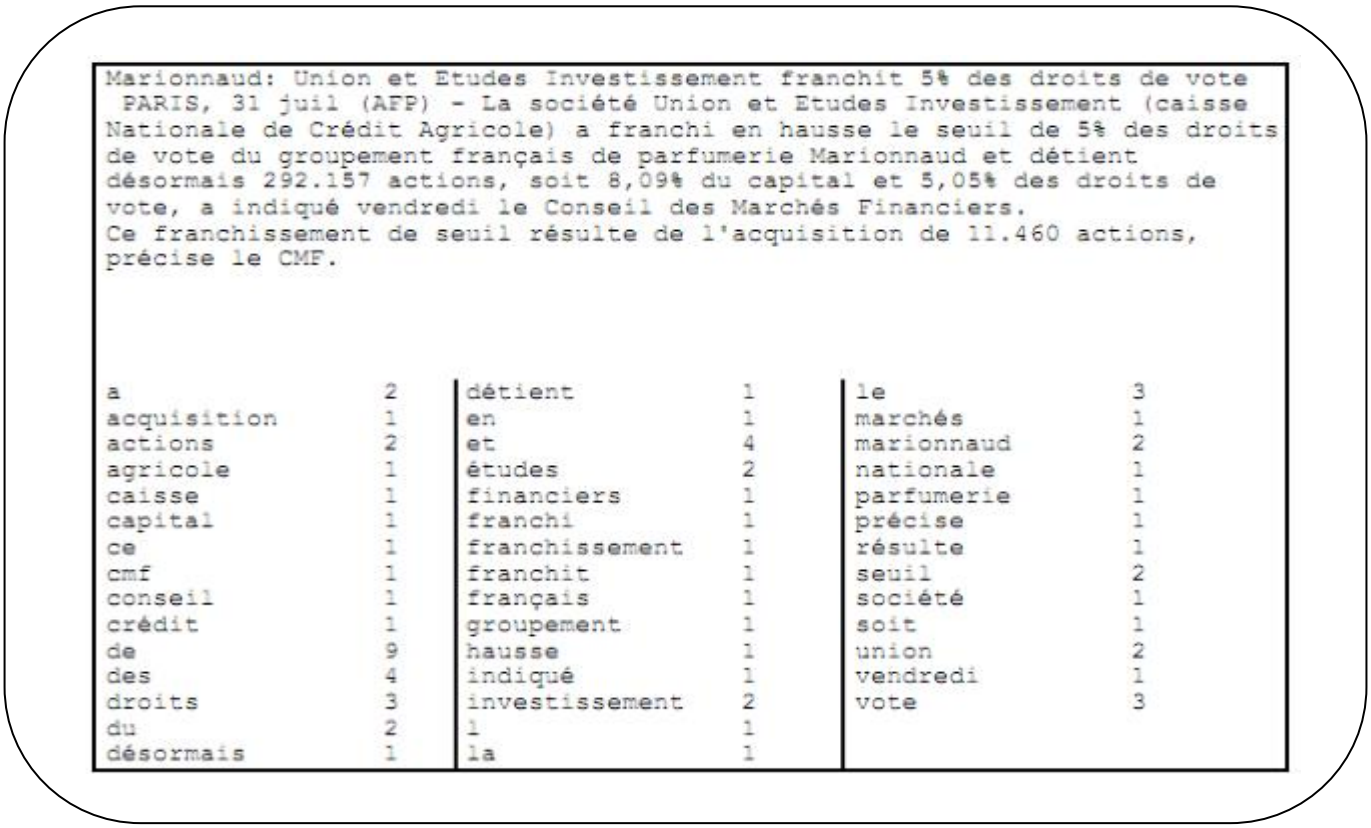
5.2. La représentation en sac de mots

La représentation des textes la plus simple a été introduite dans le cadre du modèle vectoriel, et porte le nom de "sac de mots". Les textes sont transformés simplement en vecteurs dont chaque composante représente un terme. Dans un premier temps, les termes sont les mots qui constituent un texte. Dans les langues comme le français ou l'anglais, les mots sont séparés par des espaces ou des signes de ponctuations ; ces derniers, tout comme les chiffres, sont supprimés de la représentation. On peut choisir de conserver les majuscules pour aider, par exemple, à la reconnaissance de noms propres, mais il faut alors résoudre le problème des débuts de phrase.

Les composantes du vecteur sont une fonction de l'occurrence des mots dans le texte.

À titre d'exemple, nous présentons, sur la Figure 7, une dépêche de l'Agence France Presse qui fournit des informations sur des prises de participations entre des entreprises. La transformation de ce texte en vecteur est présentée sous le texte. À partir de ces informations, un filtre doit détecter que cette dépêche est pertinente pour le thème des participations.

Cette représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots : c'est pourquoi cette représentation est appelée "sac de mots".



Marionnaud: Union et Etudes Investissement franchit 5% des droits de vote
PARIS, 31 juil (AFP) - La société Union et Etudes Investissement (caisse Nationale de Crédit Agricole) a franchi en hausse le seuil de 5% des droits de vote du groupement français de parfumerie Marionnaud et détient désormais 292.157 actions, soit 8,09% du capital et 5,05% des droits de vote, a indiqué vendredi le Conseil des Marchés Financiers.
Ce franchissement de seuil résulte de l'acquisition de 11.460 actions, précise le CMF.

a	2	détient	1	le	3
acquisition	1	en	1	marchés	1
actions	2	et	4	marionnaud	2
agricole	1	études	2	nationale	1
caisse	1	financiers	1	parfumerie	1
capital	1	franchi	1	précise	1
ce	1	franchissement	1	résulte	1
cmf	1	franchit	1	seuil	2
conseil	1	français	1	société	1
crédit	1	groupement	1	soit	1
de	9	hausse	1	union	2
des	4	indiqué	1	vendredi	1
droits	3	investissement	2	vote	3
du	2	l	1		
désormais	1	la	1		

Figure 9. Exemple d'un texte et de son vecteur associé.

6. Conclusion

Les systèmes OLAPs sert à la manipulation des données pour l'aide à la décision par différents sujets ou axes d'analyse, que ce soit les données numériques, additifs ou semi-additifs. Mais l''accès au contenu des documents textuels pour pouvoir les exploités et les analysées directement cause une des problématiques déposé par les grands entreprises prenant l'exemple d'étude les profils des clients, parmi ce qui montrent les limites des domaines classiques d'analyse.

Les domaines du traitement automatique des langues joue le rôle clé dans la tâches parmi ses outils d'adaptation on propose les systèmes de recherche d'information qui sont amené à se développer, récemment la prise en charge de la sémantique dans le texte commence à être maîtrisée via l'introduction des ontologies. On propose dans le chapitre suivant de réaliser un prototype de SRI pour la manipulation des entrepôts textuels

Chapitre 3 : Recherche d'information et entrepôts textuels

1. Introduction

Il s'agit maintenant de mettre en œuvre les étapes explicitées dans les chapitres précédents pour concevoir et implémenter les méthodes de recherche d'informations pour traiter des textes et ouvrir les portes d'analyser par l'indexation d'un corpus pour pouvoir manipuler et chargé les entrepôts et les exploités par des requêtes SQLs.

Dans un second lieu, nous allons concevoir notre système de recherche textuels pour des requêtes d'analyses SQLs de notre entrepôt

2. Problématique

L'interrogation de l'entrepôt pose un problème. En effet, il s'agit de bien cibler le type de requêtes à implémenter, car sur des millions, voire des milliards d'enregistrements, une requête peut prendre plusieurs heures avant d'afficher un résultat.

Le fait que les utilisateurs de ce type d'outil soient des décideurs, qui doivent pouvoir poser interactivement des requêtes, tout en ne pouvant pas mettre à jour les données, et que, par le biais de l'historisation, la quantité de données à manipuler soit très grande, sont les facteurs à l'origine du concept d' entrepôt de données.

La présentation est vue comme l'élément essentiel de l'entrepôt par les personnes qui vont l'interroger. En effet, les utilisateurs ne sont pas des informaticiens et ne connaissent pas le langage SQL; ils doivent néanmoins poser des questions à l'entrepôt et il n'est plus question d'avoir des pages de résultats comme lors de l'interrogation des bases de données. Un état schématique, avec des mots précis, le tout sous forme graphique, est attendu par les usagers. Il existe plusieurs façons de faire : soit l'entrepôt possède son propre outil de présentation (reporting), soit on pourra faire appel à des logiciels annexes qui permettront ce travail.

En outre les BDMs sont construites sur des données transactionnelles issues des systèmes d'information (SI) des entreprises. Ou seules 20% des données d'un SI sont des données transactionnelles et peuvent être traitées. Les 80% restants, restent hors de portée de la technologie OLAP faute d'outils et de méthodes adaptées à la gestion de données textuelles.

3. Processus d'élaboration de l'entrepôt

L'entrepôt de données stocke des informations collectées à partir d'une source globale; ce processus vise à construire le schéma de l'entrepôt, à partir de celui de la source globale. Olivier Teste a proposé un processus de construction de l'entrepôt décomposé en deux phases:

- Une phase d'extraction qui consiste à définir la structure des classes entrepôt, en dérivant les propriétés des classes source pertinentes pour les utilisateurs, et éventuellement,

en augmentant les classes entrepôt d'attributs calculés et de propriétés spécifiques selon notre corpus.

- Une phase de charge qui consiste à organiser et implémenté la hiérarchie d'héritage des classes entrepôt, en fonction des besoins spécifiques de l'entrepôt liés aux applications décisionnelles.

4. Intégration des ressources logicielles

Notre application interroge la ressource Ontologie anglaise : Princeton WordNet data.On a déjà présentés *WordNet* dans les chapitres précédents. Le navigateur *WN* sert à faire les alignements nécessaires il utilise l'analyseur morphologique pour examiner d'autres formes possibles du mot non reconnu dans la requête exprimé par la synonymie, l'hypermimie, l'hyponymie et l'homonyme.

5. Architecture d'application

On présente dans ce (Figure 10) l'architecture globale de notre application :

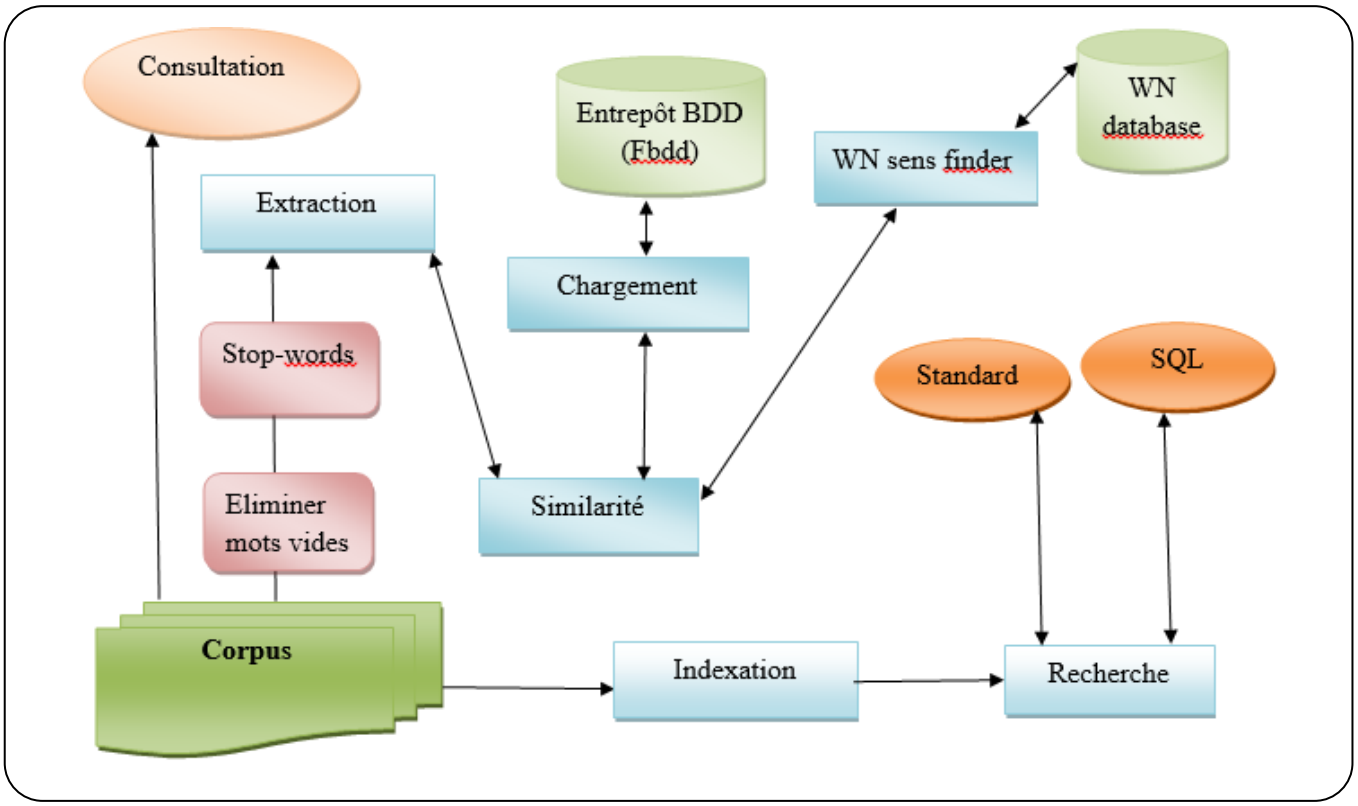


Figure 10. Architecture d'application

Le corpus d'entré seras examiner par l'élimination des mots-vides et les stops-words pour la préparation du terrain d'extraction ou on calcule la similarité des termes absorber entre eux puis on les accorde au synsets du WordNet tout cas concerne le prétraitement de notre entrepôt textuels ou le chargement des données est le complémentaire

Après tout ça une recherche seras effectué soit sur le corpus textuels comme étends une recherche standard ou bien on auras une vision plus étendu concernant la manipulation et la traduction des requêtes en langage SQL pour le lancement des résultats sur l'entrepôt chargé ce qui reste comme travail supplémentaire et l'un des perspectives

5.1. Extraction d'information

L'importation de données textuelles pour l'implémentation de l'entrepôt consiste à notre objectif principale définit par les trois fonctions citées (éliminations et simulation) qui sont des facteurs ou taches solides dans les systèmes de recherche d'information pour ces deux phases on va les détaillés plus dans le chapitre suivant.

Une des avantages de notre systèmes réside dans l'interprétation d'aspect sémantique qui est une tache supplémentaire sert à l'amélioration des résultats par l'interprétation des concepts au lieu des termes.

Pas mal de systèmes de recherches interroge l'ontologie WordNet dans la reformulation des requêtes, Or on vas l'introduire comme étant un sac de mots pour l'augmentation d'occurrences des termes au niveau des documents suivant un seuil prédéfinit entré par l'utilisateurs selon son besoin pour plus de flexibilité.

On démontre dans la figure suivante l'architecture du processus plus précise d'un schéma de sélection des termes selon le WordNet

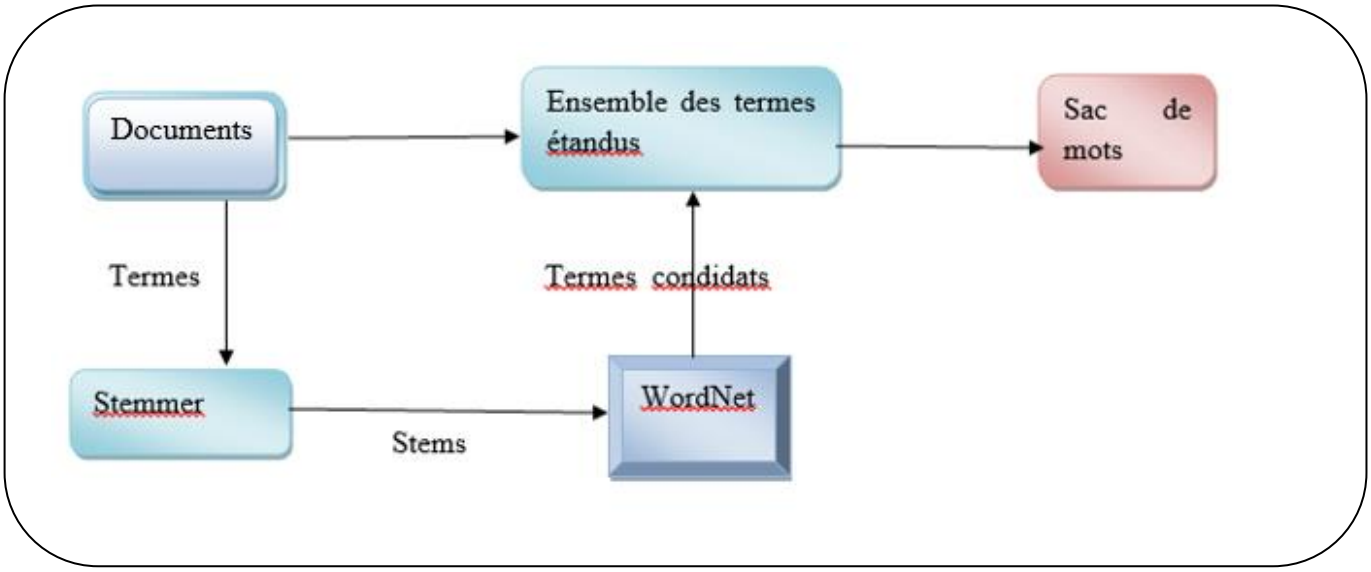


Figure11. Schéma du processus de sélection des termes

5.2. Architecture d'entrepôt textuels

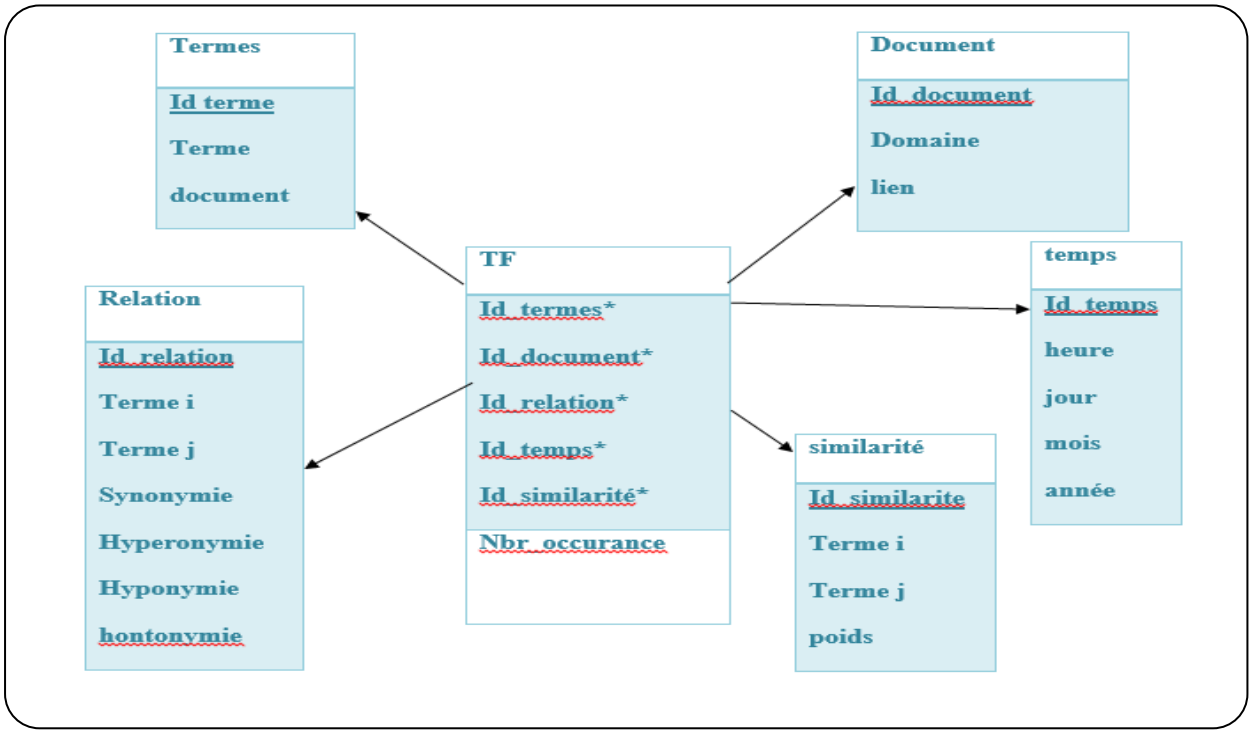


Figure 12. Le schéma en étoile en accord avec le corpus.

5.3. Module d'indexation

L'indexation des données textuels comme a était cité dans le second chapitre en details dans notre cas on va travailler par le modèle vectoriel pour le détection des termes et la création des tokens ou fichier d'index

5.4. Module de Recherche

Le codage "tf.idf" seras introduit dans le cadre du modèle vectoriel et donne parfois son nom à la méthode vectorielle. Le codage utilise une fonction de l'occurrence multipliée par une fonction qui fait intervenir l'inverse du nombre de documents différents dans lequel un terme apparaît. Ce sigle provient de l'anglais et signifie : term frequency * inverse document frequency.

Il est admis que l'occurrence d'un terme est une information importante, mais cependant, on considère généralement que la fonction f ne doit pas être l'identité. Si, par exemple, un mot apparaît deux fois dans un texte, son importance n'est pas nécessairement deux fois plus grande que s'il n'apparaissait qu'une seule fois. Pour cette raison, si l'occurrence n'est pas nulle, la fonction souvent utilisée est de la forme :

$$1 + Log\left(TF(m,t)\right)$$

En fait, les mots-clefs ont tendance à apparaître plusieurs fois dans un document pertinent : ainsi, deux mots peuvent avoir la même d'occurrence sur l'entrepôt, mais l'un des deux peut apparaître dans beaucoup plus de documents que l'autre. Pour compenser cet effet, le modèle vectoriel utilise souvent le codage appelé tf.idf qui consiste à choisir une fonction telle que

$$f = \begin{cases} (1 + Log(TF(m,t))).Log\frac{N}{DF(m)} & si\ TF(m,t) > 0 \\ 0 & si\ TF(m,t) = 0 \end{cases}$$

DF(m) est le nombre de documents différents où le terme « m » apparaît au moins une fois et N est le nombre de documents contenus dans l'entrepôt. Ainsi, si un terme apparaît dans tous les documents, son poids a une valeur nulle puisqu'il n'apporte aucune information**.[6]**

6. Représentation des diagrammes UML

IL est nécessaire pour chaque développement d'un tel logiciels de modéliser et mettre en correspondances les différents diagrammes UMLs qui s'agit. L'UML en anglais « *Unified Modeling Language* » est Le langage de modélisation unifié graphique à base de pictogrammes conçu pour fournir une méthode normalisée pour visualiser la conception d'un système. On met en œuvre le diagramme de cas d'utulisation et celui des classes.

6.1. Diagramme de cas d'utilisation

Un diagramme de cas d'utilisation capture le comportement d'un système, d'un sous-système, d'une classe ou d'un composant tel qu'un utilisateur extérieur le voit. On présente dans la Figure 13 ci-dessus le diagramme de cas d'utilisation global de notre application.

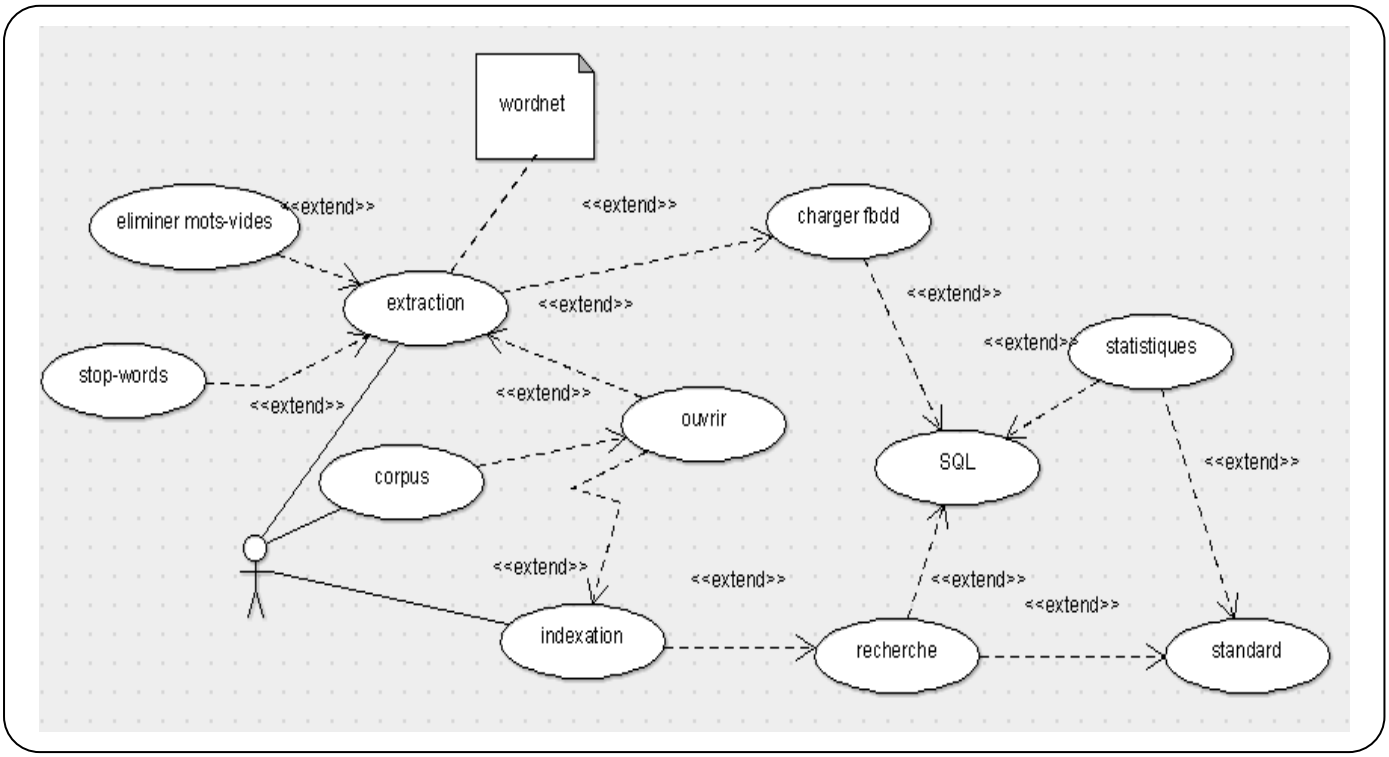


Figure13. Diagramme de cas d'utilisation d'application

7. Les travaux réalisés dans le domaine

AKNOUCHE, Rachid.2014. «*Entrepôt de textes* » : Pour aborder la problématique que les données ne peuvent être intégrées et entreposées par l'application de simples techniques employées dans les systèmes décisionnels actuels, cette démarche a étais proposé pour la construction d'entrepôts de textes, ils ont focalisé sur deux phases qui sont l'intégration des données textuelles et leur modélisation multidimensionnelle. Ils ont eu recours aux techniques de recherche d'information (RI) et du traitement automatique du langage naturel (TALN). Pour cela, ils ont conçu un processus d'ETL (Extract-Transform-Load) adapté aux données textuelles. Il s'agit d'un framework d'intégration, nommé ETL-Text, qui permet de déployer différentes tâches d'extraction, de filtrage et de transformation des données textuelles originelles sous une forme leur permettant d'être entreposées. Certaines de ces tâches sont réalisées dans une approche, baptisée RICSH (Recherche d'information contextuelle par segmentation thématique de documents), de prétraitement et de recherche de données textuelles. D'autre part, l'organisation des données textuelles à des fins d'analyse est effectuée selon TWM (Text Warehouse Modelling), un nouveau modèle multidimensionnel adapté à ce type de donnée.
Tseng et. Al, **2006.** a proposé un nouveau type de langage de requête appelé MD2X « *Multi-Dimensional expression Document* » spécialement conçu pour le document cube. Lin et al. **[37]** a proposé un nouveau cube de données appelées « *Cube texte* ». Il a une nouvelle dimension pour la navigation sémantique de données textuelles, la hiérarchie terme, qui implique deux nouvelles opérations OLAP telles que pull-up et push-down. Il dispose également de deux mesures RI spéciaux, la fréquence des termes et index inversé afin que d'autres techniques et applications RI puissent être construites de manière efficace.

Lee et McCabe et al. **[38,39 ,40**] ont proposé un moteur de RI multidimensionnelle « *MIRE* » qui est basé sur un modèle de données multidimensionnelle à utiliser la technologie OLAP. L'idée est de construire un modèle de données multi-dimensionnelle pour le texte, et permet aux utilisateurs de rechercher avec la facilité de techniques OLAP, *MIRE* est une approche pour construire un système d'RI sur le dessus de OLAP installation où des tables de faits contiennent la mesure, mot-semble-en-document et les tableaux des dimensions contiennent des données structurées hiérarchiques. *MIRE* est un nouveau système de RI intégrant un index inversé pour le texte et une méthode d'accès multidimensionnelle des données dimensionnelles.

Perez et al. 2007, « *le R-Cubes* » sont des cubes OLAP contextualisées avec les documents texte, est une architecture pour l'intégration d'un entrepôt de données structurées avec un entrepôt de documents textuels puisque le entrepôt de documents peut contenir des documents sur de nombreux sujets différents, un utilisateur spécifie d'abord un cadre d'analyse en fournissant une séquence de mots-clés un R-cube « *Pertinence Cube* » est matérialisée par la récupération des documents et les faits liés au contexte sélectionné. Chaque fait dans le R-cube est lié à l'ensemble des documents qui décrivent son contexte, et à une valeur numérique représentant sa pertinence par rapport au contexte spécifié **[41].**

Roy et al.2005, le score de système « *Context-Oriented symbiotique Information Retrieval* » pour intégrer de façon transparente des informations critiques de l'entreprise répartis sur deux

sources de données structurées et non structurées. Dans les solutions d'intégration de l'information existante, l'application a besoin de formuler la logique SQL pour récupérer les données structurées nécessaires, d'une part, et d'identifier un ensemble de mots-clés pour récupérer les données non structurées connexes sur l'autre. Certaines techniques sont utilisées dans *SCORE* qui utilise des mots-clés pour récupérer les contenus non structurés pertinente en utilisant un moteur de recherche, Car *SCORE* peut travailler avec tous les SGBD SQL justificatives et de toute source de données non structurées avec une interface de recherche par mots-clés. **[41]**

8. Conclusion

Ce chapitre a été consacré à la description conceptuelle de notre application. Différent méthodes de logiciels et outils linguistiques ont été notés. Notre approche a pu mettre en œuvre plusieurs concepts et stratégie, qui paraissaient abstraits, dans une seule interface simplifiée par suite, on a démontré notre architecture logiciel globale en spécifiant le module d'indexation vectoriel et la recherche par similarité suggéré a appliqué sur l'intégration des données textuels a notre base d'entrepôts d'un part et la fonction tf_idf pour le système de recherche d'information d'autre part .

Chapitre 4 : Conception et intégration

1. Introduction

Il s'agit maintenant de mettre en œuvre les étapes explicitées dans les chapitres précédents pour concevoir et implémenter les méthodes de recherche d'informations pour traiter des textes et ouvrir les portes d'analyser par l'indexation d'un corpus pour pouvoir manipuler et chargé les entrepôts et les exploités par des requêtes SQLs.

2. Environnement de l'application

L'implémentation et les tests de notre application ont été réalisés dans l'environnement matériel et logiciel suivant :

- Processeur : Intel ® Core ™i3-3110M CPU @ 2.40 GHz
- Mémoire installée (RAM) : 2.00 GO
- MS-Windows 8.1 : professionnel. Type de système : système d'exploitation 64 bits.
- Java sous l'environnement Netbeans 7.3.1.
- wampServer 2.2 : wampserver2.2e-php5.4.3-httpd2.2.22-mysql5.5.24-x64

2.1. Langage d'application

Java est un langage de programmation et une plate-forme informatique créée par Sun Microsystems en 1995, racheté plus tard par Oracle. Il s'agit de la technologie sous-jacente qui permet l'exécution des applications modernes sur différentes plateformes.

La portabilité, des programmes Java sur différents systèmes d'exploitation, représente son atout principal. Java est utilisée sur plus de 850 millions d'ordinateurs de bureau et un milliard de périphériques dans le monde, dont des périphériques mobiles et des systèmes de diffusion télévisuelle. Il est doté d'une riche bibliothèque de classes, comprenant la gestion des interfaces graphiques (fenêtres, menus, graphismes, boites de dialogue, contrôles)

Il se caractérise aussi par l'existence d'une API (Interface de programmation d'applications) JAVA qui fournit l'accéder à l'ontologie

2.2. IDE NetBeans

NetBeans, créé à l'initiative de Sun Microsystems (Noyau de Forte4J/SunOne), présente toutes les caractéristiques indispensables à un IDE de qualité, que ce soit pour développer en Java, Ruby, C/C++ ou même PHP.

De licence Open Source, NetBeans permet de développer et déployer rapidement et gratuitement des applications graphiques Swing, des Applets, des JSP/Servlets, de l'architecture J2EE, dans un environnement fortement personnalisable.

2.3. Intégration des ressources logiciels

Notre application interroge la ressource Ontologie anglaise : Princeton WordNet data.On a déjà présentés *WordNet* dans les chapitres précédents. Le navigateur *WN* sert à faire les

alignements nécessaires il utilise l'analyseur morphologique pour examiner d'autres formes possibles du mot non reconnu dans la requête exprimé.

3. Interfaces et Intégration

Une fois l'application seras lancé une fenêtre d'accueil devrais apparais qui permet l'accès a la fenêtre principale du projet, l'exécution de cette dernière doit permettre l'autorisation et la permission de connexion a la structure de base de donnée définit sur le WanpServer.

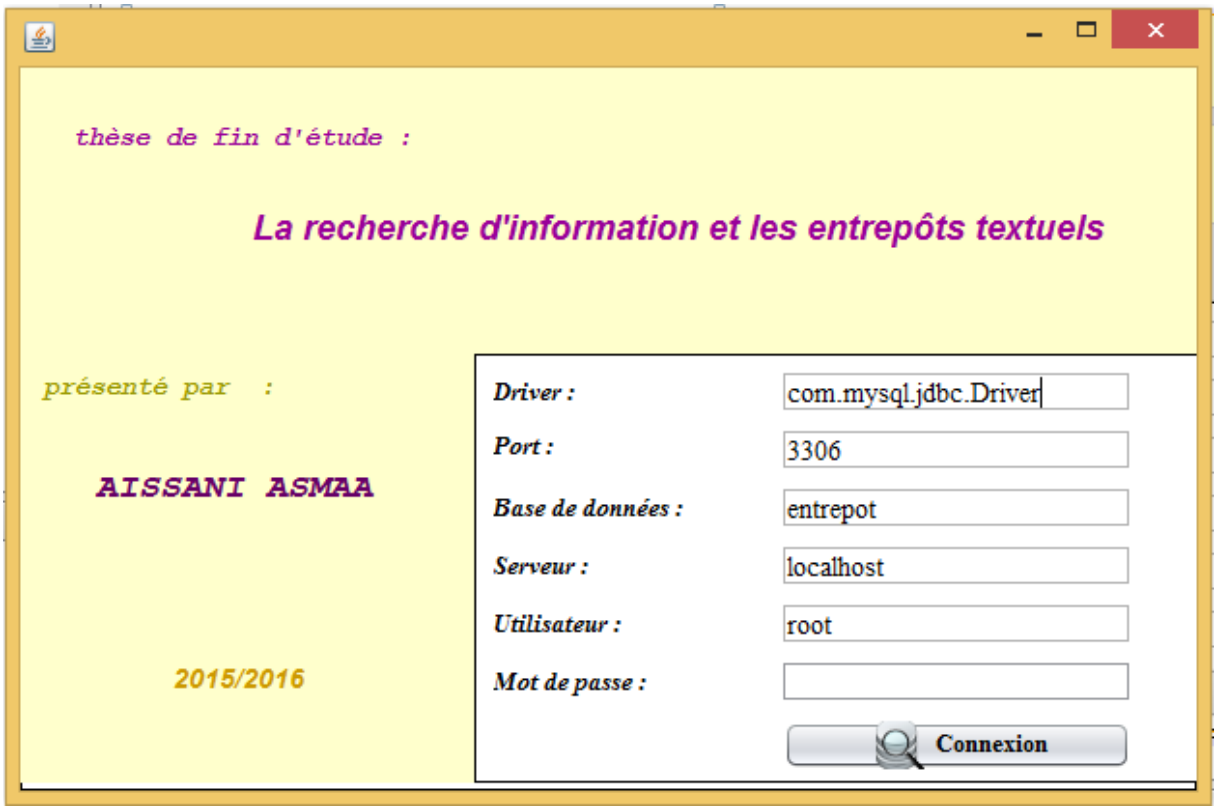


Figure15. Fenêtre d'accueil du projet

4. Traitement du corpus

Lorsque l'utilisateur accède à l'application il seras donc connecté et relié automatiquement à son architecture de base d'entrepôts :

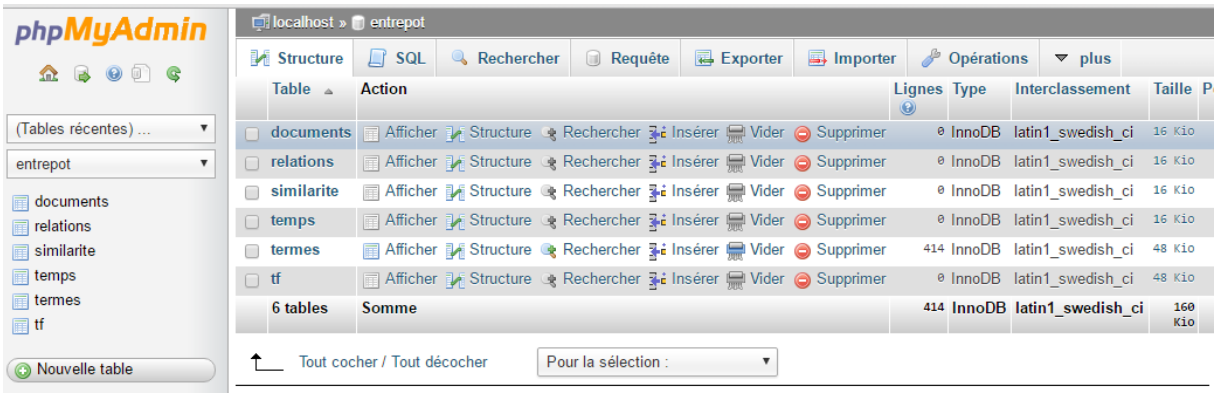


Figure16. Structure de base d'entrepôt

Tout d'abord ne faut que l'utilisateur sélectionne son corpus ou son source de documentation en cliquant une fois sur le bouton « chargé »

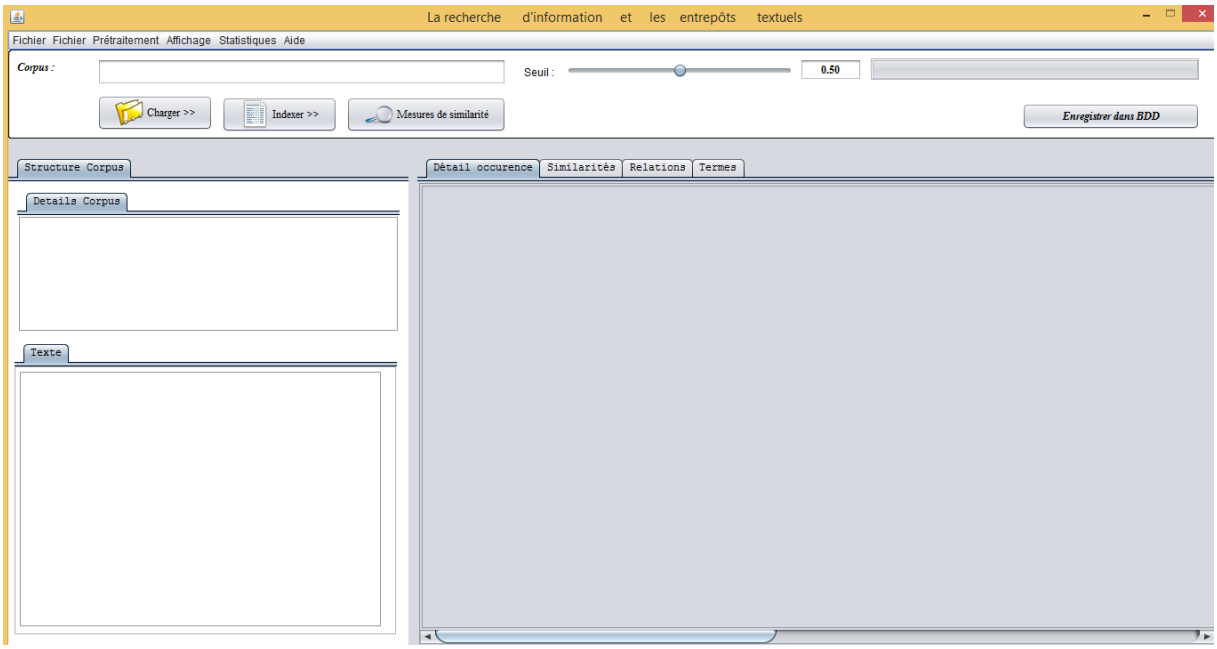


Figure17. Fenêtre principale d'application

4.1. Chargement

En cliquant sur le bouton « charger » le système doit lancés une fenêtre pour que l'utulisateur définit son corpus d'etude qui correspond à sa structure de base d'entrepôt sur le quelle on veut extraire les données.

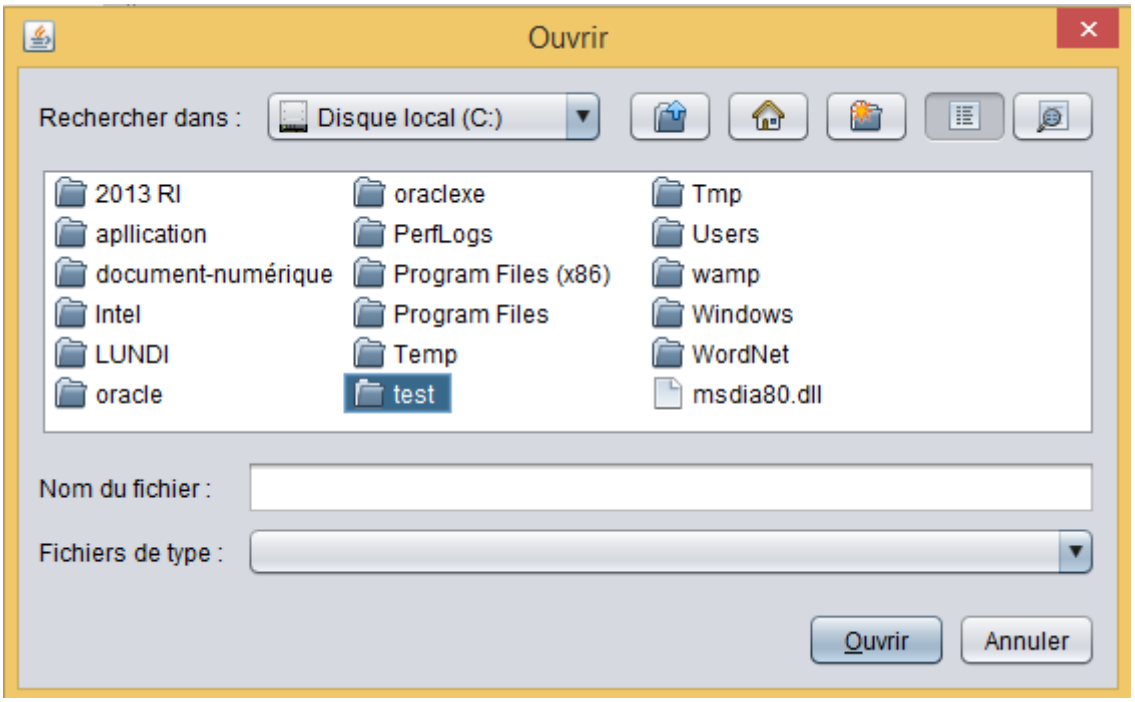


Figure18. Choix du corpus de test

Notre corpus est sous forme d'un dossier contient plusieurs autres dossiers classé selon des domaines d'analyse différents ou chaqu'un englobe plusieurs fichiers :

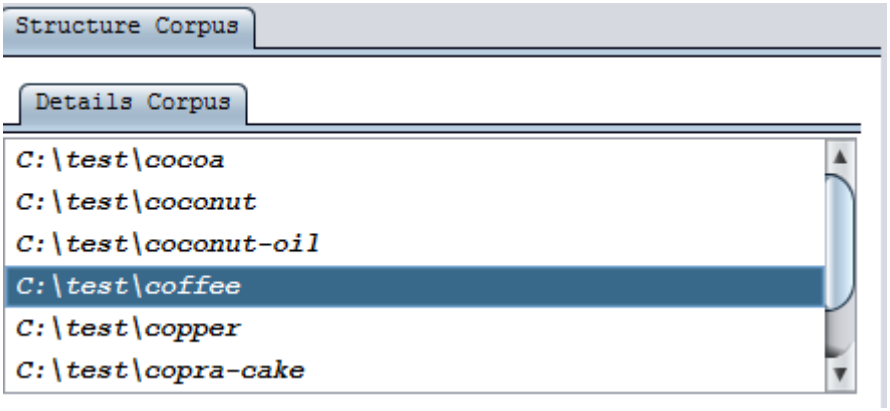


Figure19. Liste des dossiers apparus du corpus

Il est possible de visualisé le contenu du corpus sélectionnée ce qui mène à une flexibilité par suite voyant les termes et les donnais extraites (la figure 20).

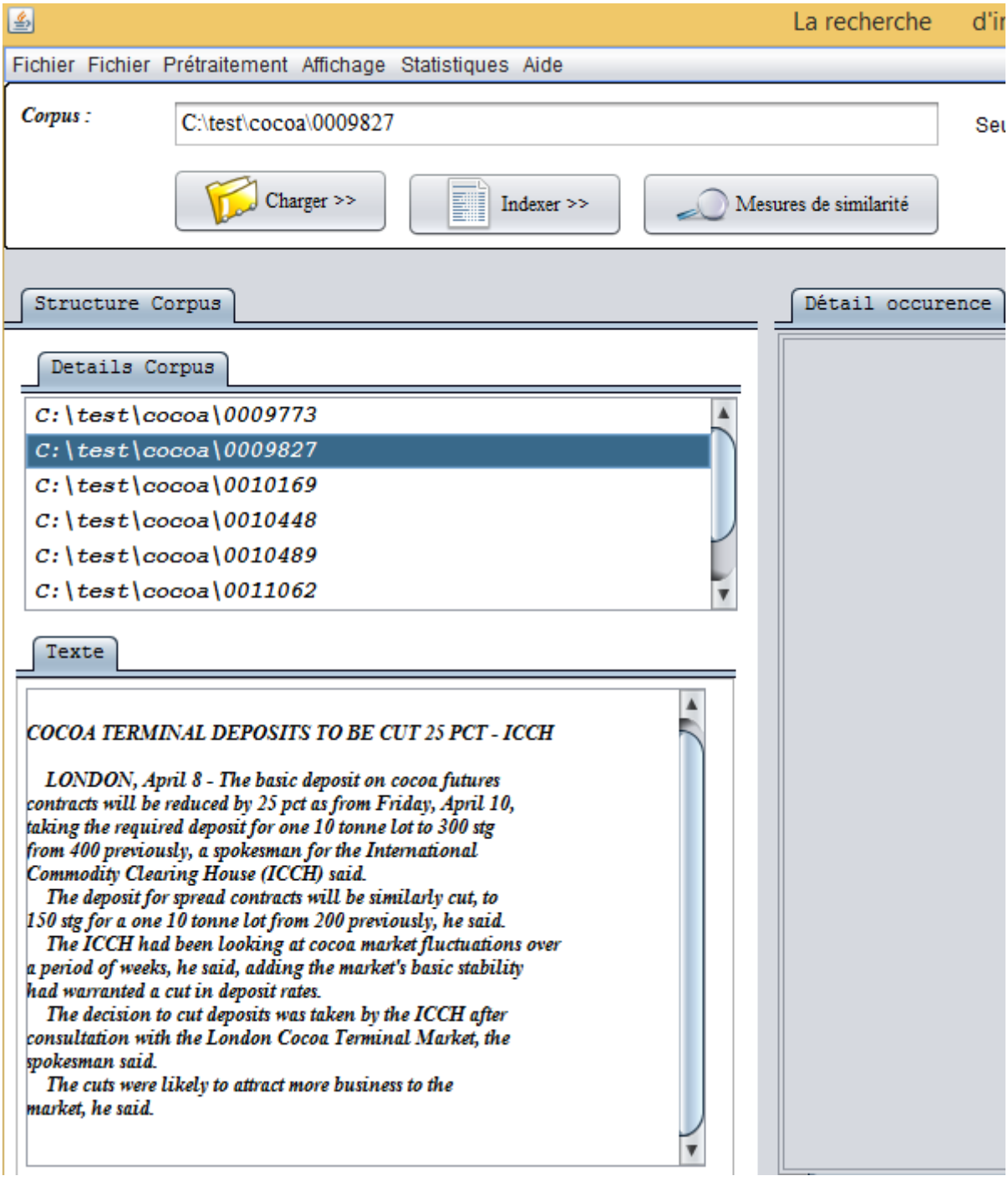
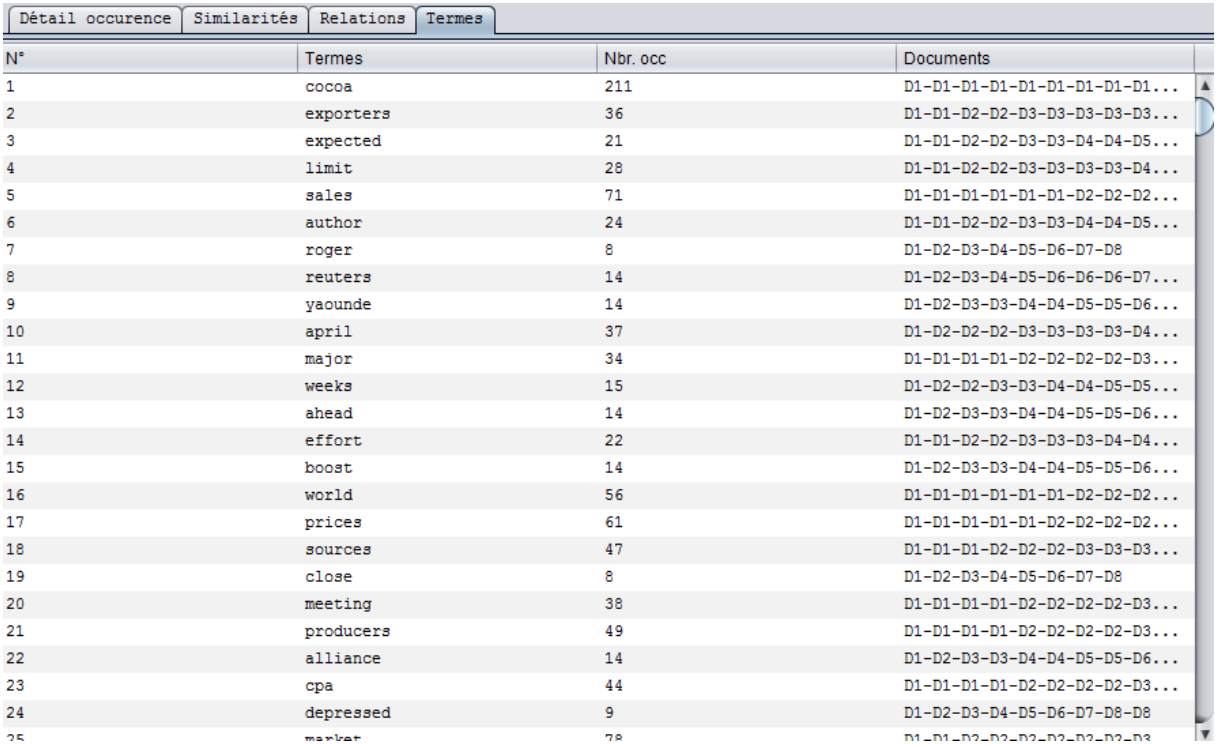


Figure 20. Ouverture et lecture du contenu d'un corpus

4.2. Extraction

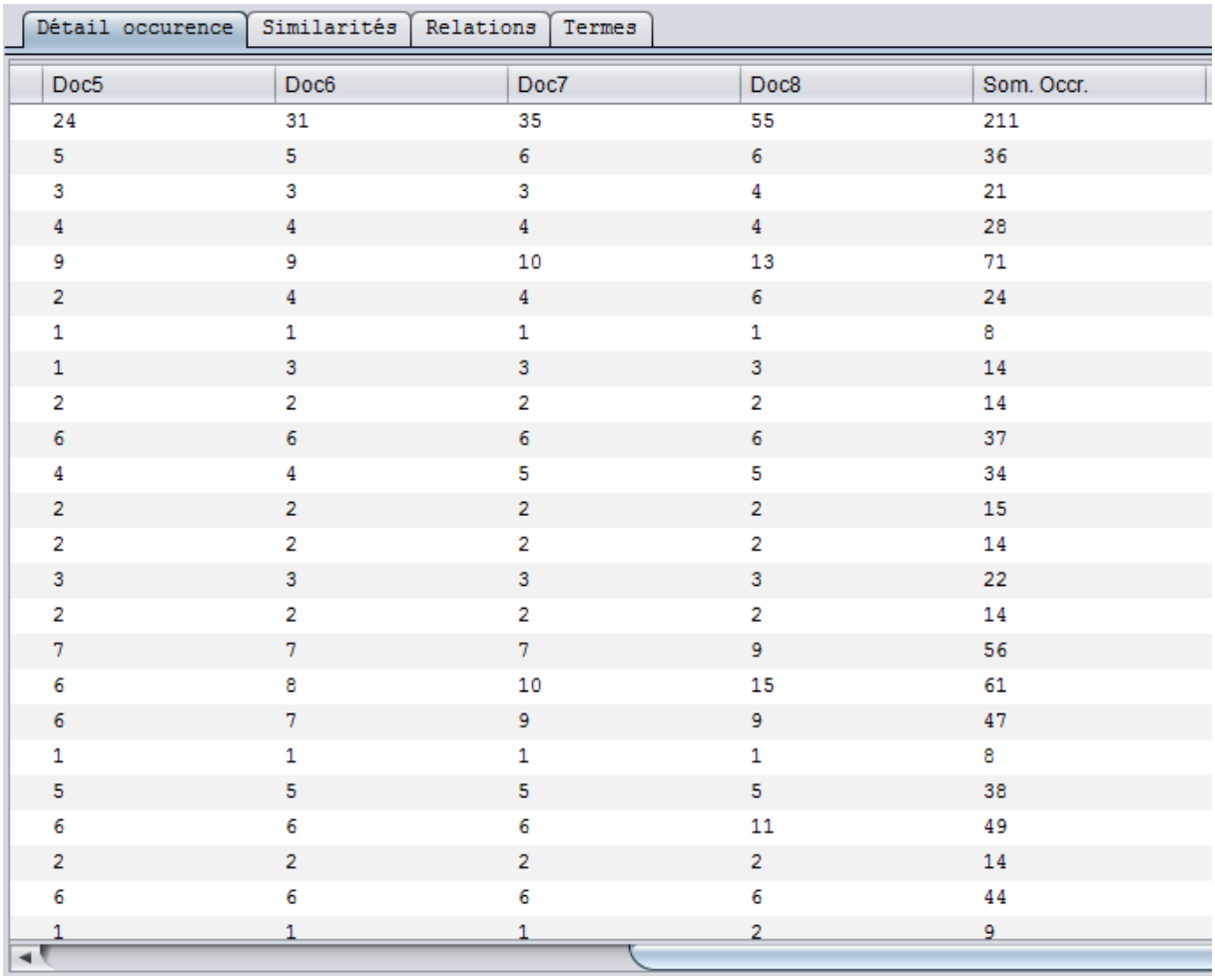
L'élimination des stopes-words et des mots vides est une étape nécessaire pour effectuer une indexation des données textuels, une fois le corpus est chargé l'utilisateur doit cliquer sur le bouton « Indexer » pour pouvoirs appliquer l'extraction des termes, deux tables seront remplit celle des termes et l'autre des détails d'occurrence



N°	Termes	Nbr. occ	Documents
1	cocoa	211	D1-D1-D1-D1-D1-D1-D1-D1-D1...
2	exporters	36	D1-D1-D2-D2-D3-D3-D3-D3-D3...
3	expected	21	D1-D1-D2-D2-D3-D3-D4-D4-D5...
4	limit	28	D1-D1-D2-D2-D3-D3-D3-D3-D4...
5	sales	71	D1-D1-D1-D1-D1-D1-D2-D2-D2...
6	author	24	D1-D1-D2-D2-D3-D3-D4-D4-D5...
7	roger	8	D1-D2-D3-D4-D5-D6-D7-D8
8	reuters	14	D1-D2-D3-D4-D5-D6-D6-D6-D7...
9	yaounde	14	D1-D2-D3-D3-D4-D4-D5-D5-D6...
10	april	37	D1-D2-D2-D2-D3-D3-D3-D3-D4...
11	major	34	D1-D1-D1-D1-D2-D2-D2-D2-D3...
12	weeks	15	D1-D2-D2-D3-D3-D4-D4-D5-D5...
13	ahead	14	D1-D2-D3-D3-D4-D4-D5-D5-D6...
14	effort	22	D1-D1-D2-D2-D3-D3-D3-D4-D4...
15	boost	14	D1-D2-D3-D3-D4-D4-D5-D5-D6...
16	world	56	D1-D1-D1-D1-D1-D1-D2-D2-D2...
17	prices	61	D1-D1-D1-D1-D1-D2-D2-D2-D2...
18	sources	47	D1-D1-D1-D2-D2-D2-D3-D3-D3...
19	close	8	D1-D2-D3-D4-D5-D6-D7-D8
20	meeting	38	D1-D1-D1-D1-D2-D2-D2-D2-D3...
21	producers	49	D1-D1-D1-D1-D2-D2-D2-D2-D3...
22	alliance	14	D1-D2-D3-D3-D4-D4-D5-D5-D6...
23	cpa	44	D1-D1-D1-D1-D2-D2-D2-D2-D3...
24	depressed	9	D1-D2-D3-D4-D5-D6-D7-D8-D8
25	market	78	D1-D1-D2-D2-D2-D2-D2-D2-D3

Figure 21. Extraction et apparences des termes

Les occurrences des termes indique l'existence ou l'apparence des termes dans tel fichiers à chaque fois ou les termes se répète augmente la variable d'occurrence selon des documents spécifiés bien sûr et ça aide au niveau d'analyse des requêtes SQL, par exemple le nombre d'occurrence de terme 'author' par documents

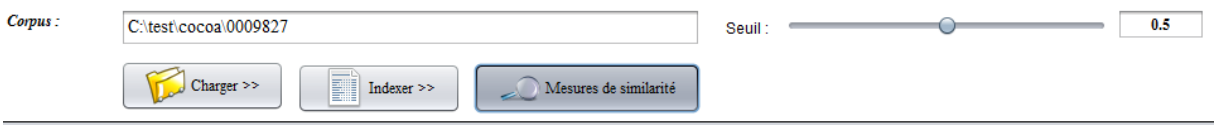


Doc5	Doc6	Doc7	Doc8	Som. Occr.
24	31	35	55	211
5	5	6	6	36
3	3	3	4	21
4	4	4	4	28
9	9	10	13	71
2	4	4	6	24
1	1	1	1	8
1	3	3	3	14
2	2	2	2	14
6	6	6	6	37
4	4	5	5	34
2	2	2	2	15
2	2	2	2	14
3	3	3	3	22
2	2	2	2	14
7	7	7	9	56
6	8	10	15	61
6	7	9	9	47
1	1	1	1	8
5	5	5	5	38
6	6	6	11	49
2	2	2	2	14
6	6	6	6	44
1	1	1	2	9

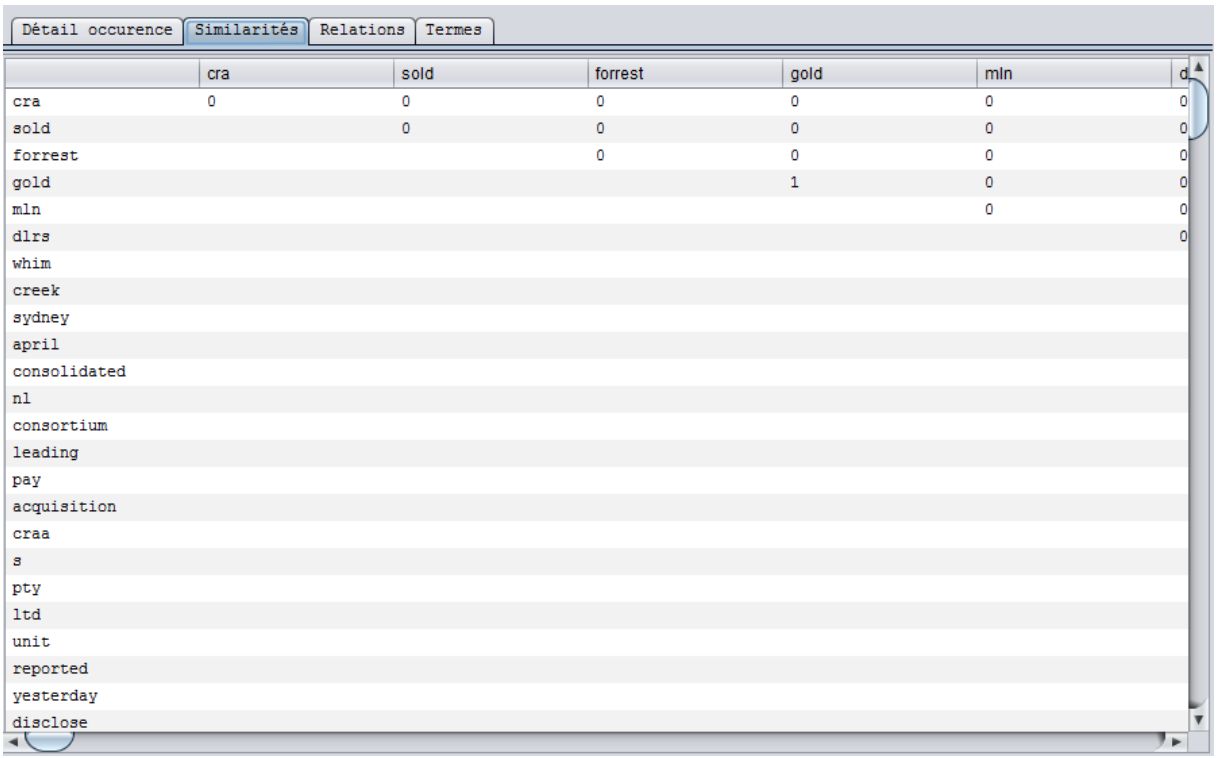
Figure22. Détails d'occurrences des termes sur les documents

4.3. Etude de similarité (WordNet)

On a interpréter la similarité des termes du corpus selon un seuil prédéfinit entré par l'utilisateur avant le lancement du mesures cet fonctionnement et fus grâce à la classe Stemmer( ) définit dans le code source



Quand l'utilisateur clique sur le bouton « Mesures de similarité » la fonction stemmer ( ) feras un calcule de similarité entre tous les termes extraire (la figure21) jusqu'à le parcours de toute les termes du corpus cela se résume par une matrice finale symétrique dans la figure 23.



	cra	sold	forrest	gold	mln	d
cra	0	0	0	0	0	0
sold		0	0	0	0	0
forrest			0	0	0	0
gold				1	0	0
mln					0	0
dlrs						0
whim						
creek						
sydney						
april						
consolidated						
nl						
consortium						
leading						
pay						
acquisition						
craa						
s						
pty						
ltd						
unit						
reported						
yesterday						
disclose						

Figure23. Calcule de similarité avec un seuil de 0.5

Après le calcule de similarité on affichera dans la table relations tous les termes qui ont une mesure de similarité plus que le seuil entre par l'utilisateur sauf les synonymes

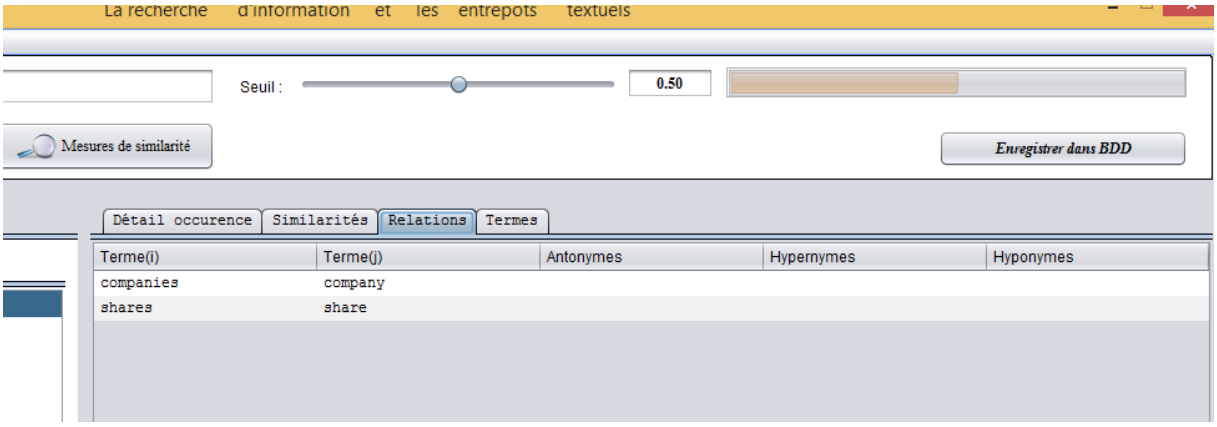


Figure24. Les relations déduites pour un seuil de 0.5

On va lancé le même traitement pour un seuil de 0.2 pour voir la différence entre les deux résultats de relations en sorties et assurée le fonctionnement (figure 25).

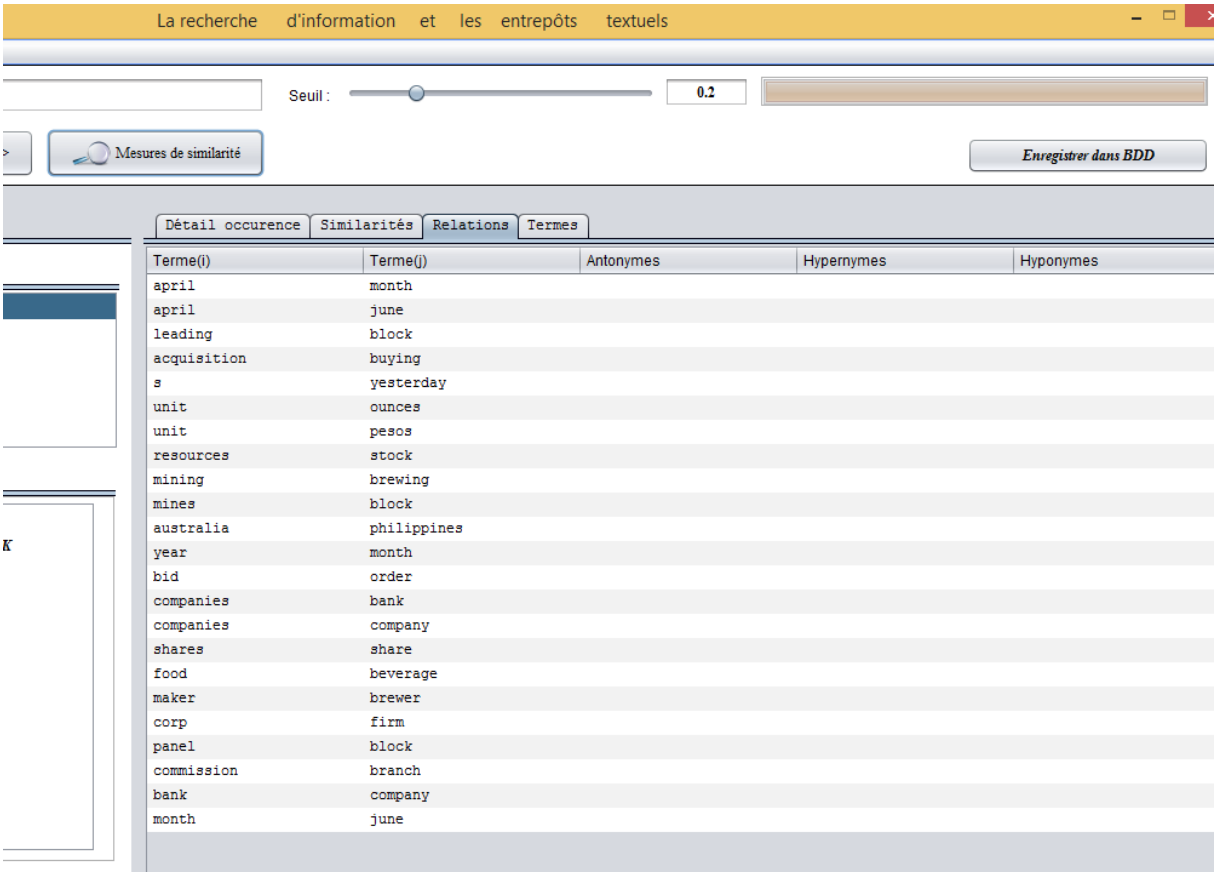


Figure25. Les relations déduites pour un seuil de 0.2

4.4. Intégration dans l'entrepôt et traitements des tables

Quand l'extraction et les mesures de similarité sont exécuté le système permet le chargement des données dans la base d'entrepôt en cliquant sur le bouton tout en haut à gauche « enregistré dans la BDD » ou les enregistrement des tables seront affecté, on pourras après affiché cet enregistrement voyant dans la bare de menu le module « affichage » puis on doit cliquer sur « BDD » d'où un tableau doit apparait (figure 26).

Toute ce traitement d'enregistrement,d'ajout et d'affichage est traiter par la fonction « fbdd » dans le code java.

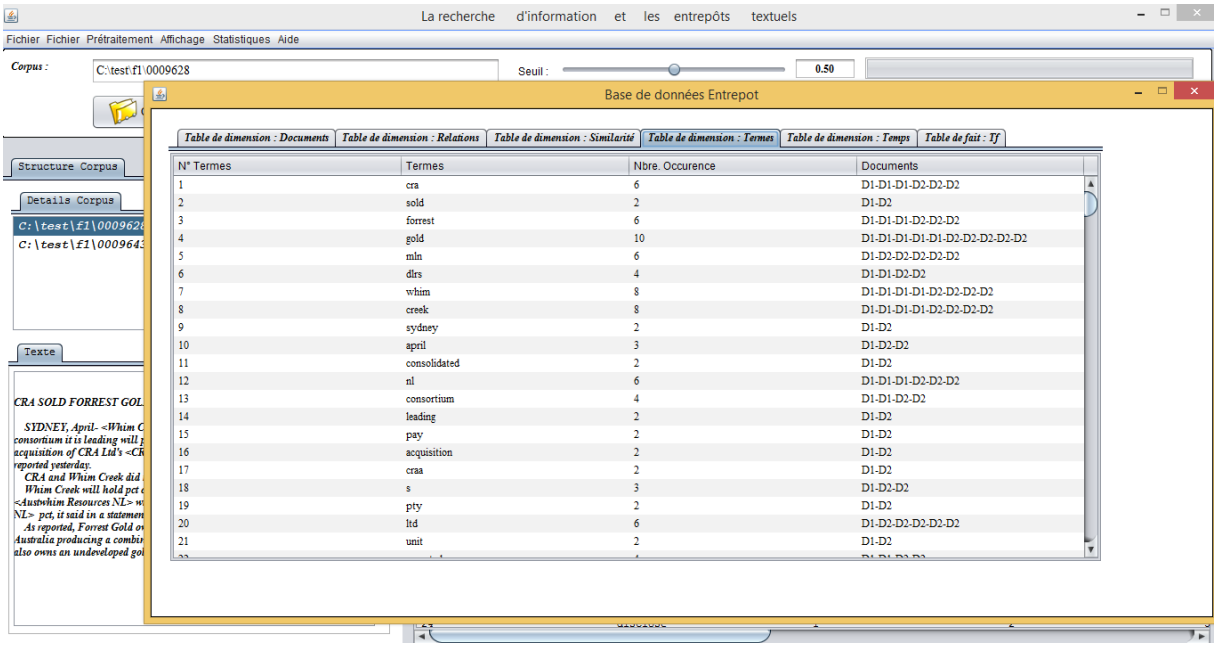


Figure26. Les enregistrements des données dans la base d'entrepôt

Pour le module de la sémantique le WordNet nous fournis les différentes relations (figure 27), comme un sac de mots pour les différents termes déduite du corpus cela seras afficher dans la tables relations en appuyant dans la barre de menu sur le « prétraitement » et on feras le choix de la relation souhaité.

Détail occurence | Similarités | Relations | Termes

Terme(i)	Terme(j)	Antonymes	Hypernymes	Hyponymes
april	month		the month following March and preceding May	
april	june		the month following March and preceding May	
leading	block		thin strip of metal used to separate lines of...	
acquisi...	buying		the act of contracting or assuming or acquiri...	[PointerTargetNode: [...
s	yesterday		1/60 of a minute; the basic unit of time adop...	
unit	ounces		any division of quantity accepted as a standa...	
unit	pesos		any division of quantity accepted as a standa...	
resources	stock		available source of wealth; a new or reserve ...	
mining	brewing		the act of extracting ores or coal etc from t...	
mines	block		excavation in the earth from which ores and m...	
australia	philippines		a nation occupying the whole of the Australia...	
year	month		a period of time containing 365 (or 366	[PointerTargetNode: [...
bid	order		an authoritative direction or instruction to ...	[PointerTargetNode: [...
companies	bank		an institution created to conduct business; "...	
companies	company	[PointerTargetNo...	an institution created to conduct business; "...	[PointerTargetNode: [...
shares	share	[PointerTargetNo...	any of the equal portions into which the capi...	[PointerTargetNode: [...
food	beverage		any substance that can be metabolized by an o...	[PointerTargetNode: [...
maker	brewer		a person who makes things	
corp	firm		a business firm whose articles of incorporati...	[PointerTargetNode: [...
panel	block		sheet that forms a distinct (usually flat	
commission	branch		a special group delegated to consider some ma...	
bank	company		a financial institution that accepts deposits...	
month	june		one of the twelve divisions of the calendar y...	

Figure 27. Les relations du WordNet pour le corpus

```
----------------------------
Le mot :neptunia
----------------------------
Le mot :subsidiary
----------------------------
          Synonyme :subordinate
          Synonyme :subsidiary
          Synonyme :underling
          Synonyme :foot_soldier
          Synonyme :subsidiary_company
          Synonyme :subsidiary
Le mot :talks
----------------------------
          Synonyme :negotiation
          Synonyme :dialogue
          Synonyme :talks
Le mot :break
----------------------------
          Synonyme :interruption
          Synonyme :break
          Synonyme :break
          Synonyme :good_luck
          Synonyme :happy_chance
```

Figure 28. La synonymie des termes détecter par WN

Ces relations permet une amélioration de performance de système car si un utilisateur entre une requête par un mot inexistant dans la base ou le corpus donc les résultats retournées seront en accords de ses relations c'est-à-dire si le mot entré est inexistant le résultat seras peut etre par son synonyme

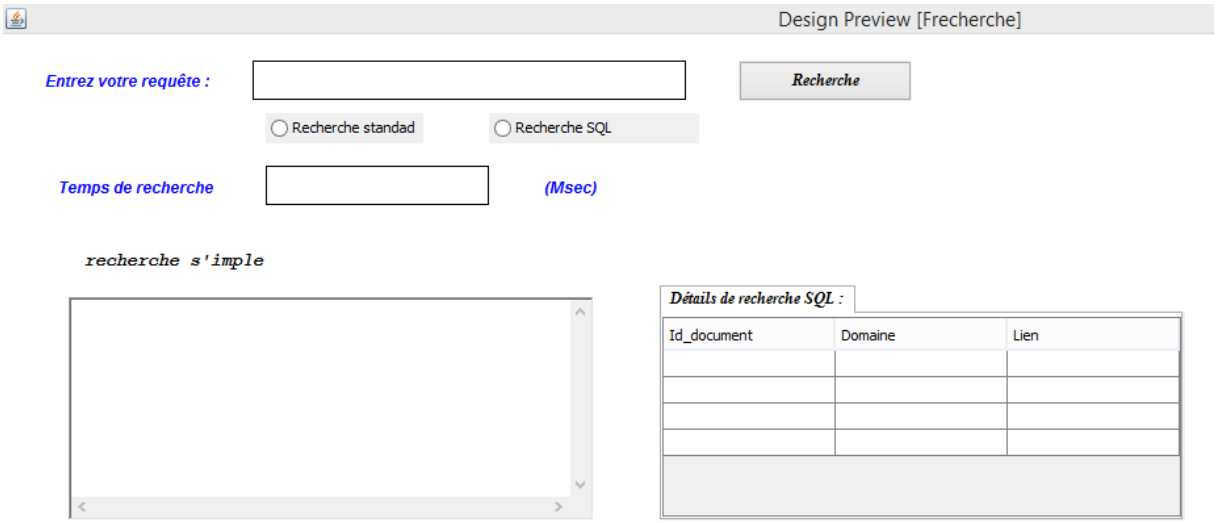


Figure 29. Interface de recherche

5. Conclusion

On a présenté à travers ce dernier chapitre la conception et l'intégration de notre système de recherche ainsi que son réalisation. L'implémentation a pu mettre en œuvre plusieurs stratégie et approches, comme amélioration, nous avons combiné l'ontologie pour obtenir des bons résultats.

Notre système permet de fournir une structure d'extraction et chargement pour la construction d'une base d'entrepôt à travers un corpus entré par l'utilisateur par la combinaison des méthodes du système de recherche qui retourne les données pertinentes souhaité selon le besoin des utilisateurs. Les tests appliqués sur notre corpus anglais sont encourageants et nous motivent à approfondir nos recherches dans ce domaine.

Conclusion générale

Dans ce travail on a étudié les aspects théoriques et pratiques relatives à la recherche d'information et qui sont rattachés à un contexte plus général d'analyse de données textuelles pour les systèmes d'aide à la décision.

Premièrement on a exposé la recherche d'information et son noyau «l'indexation» montrant ces principes arrivant à un SRI afin d'offrir à l'utilisateur un accès facile à l'information qui l'intéresse. Cette information étant située dans une masse de documents textuels (corpus), l'étude suivante s'est portée sur la nature des entrepôts textuels en décrivant plus particulièrement l'intérêt d'un SRI pour le traitements de cet dernières.

Un intérêt particulier était orienté vers la prise en charge des aspects sémantiques pour s'approcher le maximum du besoin en information de l'utilisateur. Notre travail explore la faisabilité d'intégrer des ressource linguistique ontologique pour améliorer la performance de recherche à travers la reformulation de la requête. Des outils très solides dans ce domaine ont pu être déployés et intégrés dans notre projet : on cite l'occurrence le systeme de recherche d'information spécifié par l'indexation *vectoriel*, l'architecture de notre *entrepôt* et *le navigateur d'ontologie WordNet.*

Les résultats nous ont amené à conclure que la RI peut réellement permettre l'extraction et une préparation de terrain d'analyse pour des gigantes entrepôts textuels pour l'avoir exploite par des requêtes SQL qui devrait être plus fiable que l'analyse classique. On a laissé le choix de la méthode ou la structure de recherche à l'utilisateur pour plus de flexibilité selon son besoin entre une recherche simple ou celle traiter au langage SQL.

Finalement l'interprétation de notre application mérite toute l'attention, et on espère que cela va être développé davantage avec de nouveaux outils et applications. Nous estimons que ce travail a pu ouvrir de nouveaux horizons pour pouvoir la recherche d'information conceptuelle dans les entrepôts textuels.

Bibliographie

1. KHROUF K., SOULÉ-DUPUY C., A Textual Warehouse Approach: a Web Data Repository, Chapter VII, Intelligent Agents for Data Mining and Information Retrieval, Idea Group Publishing, p. 101-124, 2004
2. Aissani Asmaa,Ould Hadri Imene, « réalisation d'une interface de recherche dans un texte arabe à base d'ontologie »,thème de licence université Abd Elhamid ben Badis ,université de Mostaganem 2014.
3. MAHMOUDI (Seyed Mohammad), Le rôle de la technologie de l'information et de la communication dans le reengineering des systèmes, in "Journal of Management Culture" , Qom Campus of University of Tehran", en persan, N° 11, L'hiver 2002. pp. 171-194.
4. DAL (Georgette), HATHOUT (Nabil), NAMER (Fiammetta), « *Morphologie Constructionnelle et Traitement Automatique des Langues »* : le projet MorTAL, in Temple, M., éditeur, Lexique, Volume 16, Presses Universitaire de Lille, 2004.
5. A.Rozenknop, introduction :« Recherche et Extraction d'Information », source : Romaric Besançon : CEA-LIST/LIC2M, Master Micr Paris, 13.
6. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "*Introduction to Information Retrieval"* , Cambridge University Press,2008.
7. .HABIB ZAHMANI Mohamed, ZAOUI Lynda : mémoire master « Quel modèle de Recherche d'Information pour les Entrepôts de Données Etude et implémentation » Université des Sciences et de la Technologie d'Oran – MB Faculté des Sciences 2011/2012.
8. Siham Boulaknadel,'Traitement Automatique Des Langues et Recherche D'Information En Langue Arabe Dans Un Domaine de Spécialité ' ; Apport Des Connaissances Morphologiques et Syntaxiques Pour L'Indexation Siham Boulaknadel To Cite This Version , 2010.
9. Karen SAUVAGNAT, Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés, Thèse de doctorat, Université Paul Sabatier de Toulouse, 2005.
10. Abderrahim, Med-El-Amine and Med-Alaeddine A. Using Arabic Wordnet for Query Expansion in Information Retrieval System. In *IEEE The Third International Conference on Web and Information Technologies*, Marrakech, Morocco, June 2010.
11. Xavier Tannier, introduction " *Indexation et Recherche D'Information*" Qu'Est-Ce Que La Recherche Les Acteurs de La Recherche D'Information , cours Master 2 IRI Université Paris-Sud,xavier.tannier@limsi.fr
12. Aurélien Max, « Indexation pour la Recherche d'Information » Cours d'Indexation et Recherche d'Information, Master 2 Professionnel Informatique et MIAGE Université Paris-Sud 11, Année 2010-11.
13. Populaire Universit and others, 'Construction Semi-Automatique D ' Ontologies à Partir de Textes Arabes', 2012.

14. Paul McNamee, Cours 605.744 « Information Retrieval », Johns Hopkins University, 2011. http://www.apl.jhu.edu/~paulmac/ir.html
15. Jian-Yun Nie, Cours IFT6255 « Recherche d'information », Université de Montréal, 2008. http://www.iro.umontreal.ca/~nie/IFT6255/
16. Ricardo Baeza-Yates and Berthier Ribeiro-Neto : Information Textuelle, 'RECHERCHE D'INFORMATION.', "Modern Information Retrieval",Addison-Wesley Educational Publishers Inc, 1999.
17. Christopher D, Prabhakar Raghavan and Hinrich Schütze, Textuelle Manning, *"An Introduction to Information Retrieval".*Cambridge University Press, 2008.
18. Brahmi A. Contribution à la Recherche Intelligente sur le Web: « Indexation Sémantique des Textes Non-Structurés ». Thèse de Doctorat, USTO-Oran, 2013.
19. Alain Boucher Ifi, 'Recherche D ' Information.' Source : C.D. Manning, P. Raghavan and H. Schütze, « Introduction to Information Retrieval »,Cambridge University Press. 2008.
20. Philippe Lefèvre, *La recherche d'information, du texte intégral au thésaurus*, Hermes,2000.
21. Alain Boucher – IFI, cours « Recherche d'information » *Fouille de données et recherche d'information) Source : Module « Indexation et recherche d'informations », Aurélien Max*, Université Paris 11, 2010-2011.
22. Sais, Fatiha : Transformation d'informations structures en documents XML guidée par une ontologie. Mémoire de DEA Information-Interaction-Intelligence, université Paris-Sud, septembre (2004)
23. Abderrahim, Med-El-Amine and Med-Alaeddine A. Réinjection automatique de la pertinence pour la recherche d'informations dans les textes arabes. In *IEEE 4th International Conference on Arabic Language Processing (CITALA)*, pages 77–81, Rabat, Morocco, May 2012.
24. Boucham Souhila, Une approche basée Ontologies pour l'indexation automatique et la recherche d'information Multilingue, mémoire de magister, Université M'hamed Bougara Boumerdes, 2009.
25. Baziz, Mustapha: 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, Année 2001/2002 (2002)
26. Tournier Ronan, Analyse en ligne (OLAP) des documents. Thèse de doctorat en Informatique, Universtié Toulouse III, Paul Sabatier, Toulouse, France, 2007.
27. KHROUF K., SOULÉ-DUPUY C., A Textual Warehouse Approach: a Web Data Repository, Chapter VII, Intelligent Agents for Data Mining and Information Retrieval, Idea Group Publishing, p. 101-124, 2004
28. Colliat, G. (1996). OLAP, Relational and Multidimensional database systems. SIGMOD
29. Record, vol.25(3), ACM Press, p. 64–69.
30. Chaudhuri, S. et U. Dayal (1997). An Overview of Data Warehousing and OLAP Technology. SIGMOD Rec. 26(1), 65–74.

31. Kimball, R., L. Reeves, M. Ross, et W. Thornthwaite (2000). Concevoir et déployer un data warehouse. Eyrolles.
32. Sullivan D. (2001). *Document Warehousing and Text Mining*. Wiley John & Sons, 2001
33. Journal Officiel, 'Juillet 1992, Publiée Au Journal Officiel Du 2 Juillet 1992', 1992. « *Approche de construction d'entrepôts de documents XML* », Université Toulouse I Capitole, 2014.
34. Fuhr Norbert et Grobjohann Kai, *XIRQL: a query language for information retrieval in XML documents*. 24th International ACM Conference on Research and Developement in Information Retrieval (SIGIR), ACM Press, 2001, pp. 172-180.
35. Kamps Jaap, Marx Maarten, De Rijke Maarten et Sigurbjornsson Borkur, *Best-Match Querying from Document-Centric XML*. Proceedings of the Seventh International Workshop the Web and Databases, 2004, pp. 55-60.
36. Tseng Frank S. C. et Chou Annie Y. H., *The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence.* Decision Support Systems, 2006, pp. 727-744
37. Franck Ravat, Olivier Teste,Ronan Tournier et Gilles Zurfluh : Top_Keyword : fonction d'agrégation OLAP « agrégation de mots-clefs dans un environnement d'analyse en ligne (OLAP) » IRIT SIG/ED, UMR5505, 118 rte. de Narbonne,F31062 Toulouse CEDEX 9, France 1996.
38. Lin, C. X., Ding, B., Han, J., Zhu, F., & Zhao, B. (2008). Text Cube: Computing IR measures for multidimensional text database analysis. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 905-910.
39. Lee, J., Grossman, D., & Orlandic, R. (2002). MIRE: A multidimensional information retrieval engine for structured data and text. *Proceedings of the International Conference on Information Technology: Coding and Computing*, Las Vegas, NV, USA. 224-229.
40. Boucham Souhila, Une approche basée Ontologies pour l'indexation automatique et la recherche d'information Multilingue, mémoire de magister, Université M'hamed Bougara Boumerdes, 2009.
41. Zargayouna H., « Quelle évaluation pour la Recherche d'Information Sémantique », Troisième Atelier Recherche d'Information Semantique RISE@CORIA'2011, 2011.
42. http://www.torrefacteurjava.fr/content/utiliser-apache-lucene-pour-effectuer-des recherches-textuelles
43. Harman D., « Towards interactive query expansion », Proceedings of the 11th annualinternational ACM SIGIR conference on Research and development in information retrieval, SIGIR '88, ACM, New York, NY, USA, p. 321-331, 1988.
44. http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html
45. Lee, J., Grossman, D., Frieder, O., & McCabe, M. C. (2000). Integrating structured data and text: A multi-dimensional approach. *Proceedings of the Internationa*
46. *Conference on Information Technology: Coding and Computing*, Las Vegas, NV, USA. 264-269.
47. McCabe, M. C., Lee, J., Chowdhury, A., Grossman, D., & Frieder, O. (2000). On the design and evaluation of a multi-dimensional approach to information retrieval (poster

session). *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece. 363-365.
48. Byung-kwon Park, 'Toward Total Business Intelligence Incorporating Structured and Unstructured Data', 2011.