الجمهورية الجزائرية الديمقراطية الشعبية

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Dr Yahia Farés de Médéa

Faculté des Sciences

Département de Mathématiques et informatique



MÉMOIRE DE FIN D'ÉTUDES

Pour l'obtention du diplôme de Master en Informatique option
Ingénierie des systèmes et technologies du web

THÈME

L'indexation basée sur la similarité des termes pour améliorer la Recherche d'information en langue Arabe

Proposé et dirigé par :

- **Mr.MAZARI Ahmed Cherif**
- **M^{elle}.ALI KHOUDJA Meriem**

Préparé et présenté par :

- **Mr.BOUGUERRA Elouanes**
- **Mr. ABDELLAH KHATER Esam**

Année universitaire 2016-2017

الجمهورية الجزائرية الديمقراطية الشعبية

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Dr Yahia Farés de Médéa

Faculté des Sciences

Département de Mathématiques et informatique

MÉMOIRE DE FIN D'ÉTUDES

Pour l'obtention du diplôme de Master en Informatique option Ingénierie des systèmes et technologies du web

THÈME

L'indexation basée sur la similarité des termes pour améliorer la Recherche d'information en langue Arabe

Proposé et dirigé par :

- **Mr.MAZARI Ahmed Cherif**
- **M^{elle}.ALI KHOUDJA Meriem**

Préparé et présenté par :

- **Mr.BOUGUERRA Elouanes**
- **Mr. ABDELLAH KHATER Esam**

Année universitaire 2016-2017

À Nos très chers parents qui ont toujours été là pour Nous, qui Nous guidé durant les moments les plus pénibles de ce long chemin et qui Nous ont donné un magnifique modèle de labeur et de persévérance. Nous espérons qu'ils trouveront dans ce travail toute notre reconnaissance et tout notre amour.

À tous Nos chers frères et sœurs.

À tous Nos amis et Nos collègues...

REMERCIMENTS

Nous exprimons notre profonde gratitude à toute personne qui, de près ou de loin, a contribué à la réalisation de ce travail.

Nos remerciements s'adressent plus particulièrement à Mr.MAZARI Ahmed Cherif et M^{elle}.ALI KHOUDJA Meriem qui sont bien accepté de diriger ce mémoire. Ses remarques, ses précieux conseils et ses corrections nous ont été d'une grande utilité. Nous eux disons « Merci ».

Nous aimerions témoigner notre gratitude aux professeurs et enseignants de la Faculté des Sciences qui ont assuré notre formation universitaire pour les conseils prodigués depuis la première année académique, sans lesquels ce travail n'aurait pas abouti à sa fin.

Enfin Nos remerciements iront également vers tous ceux qui ont accepté avec bienveillance de participer au jury de ce mémoire.

« و ما توفيقي إلا بالله عليه توكلت و إليه أنيب »

Résumé

Les systèmes de recherche d'information classiques reposent sur l'indexation par mots-clés pour représenter le contenu des documents et des requêtes, ainsi pour sélectionner les documents à retourner. Effectuer la recherche à ce niveau lexical peut diminuer la précision des résultats de la recherche, et rend très difficile de satisfaire le besoin informationnel de l'utilisateur, du fait de l'absence des liens sémantiques entre la représentation des documents et la représentation de la requête.

Dans ce présent travail, nous proposons une nouvelle approche d'indexation automatique de textes arabes. Elle s'appuie sur les termes et leurs termes similaires, afin de capturer tous les sens voulus par une requête, qui ne donne jamais une description complète et précise du besoin d'information de l'utilisateur. Cette approche se base sur le calcul de l'information mutuelle, en utilisant les relations de cooccurrences, et sur les méthodes d'apprentissage non supervisée (Clustering) pour faire une catégorisation des documents afin d'avoir les différents sens d'un mot dans ses différents domaines d'utilisation.

Mots-Clés :

Recherche d'information arabe, indexation sémantique, apprentissage non supervisée, catégorisation des documents, relations de cooccurrence, information mutuelle, similarité sémantique.

Abstract

Classical information retrieval systems rely on keyword indexing to represent the content of documents and queries, and to select documents to be returned to the user. In such systems, working at lexical level instead of semantic level risk to decrease the precision of results of retrieval process, and make it difficult to satisfy user's informational need, because of the absence of semantic links between documents' and query's representations.

In this work, we propose a novel automatic Arabic texts indexing approach based on terms and their associated terms, in order to capture all the meaning wanted by a query, which cannot give a complete and a precise description of the user's informational need. This approach is based on a mutual information using co-occurrences between terms, and on a clustering analysis to generate a documents' categorization in order to have different meanings of a term in its different fields.

Keywords :

Arabic information retrieval, semantic indexing, clustering, documents' categorization, co-occurrences, mutual information, semantic similarity.

مُلَخَّص

تعتمد أنظمة البحث عن المعلومات على الفهرسة عن طريق الكلمات الرئيسية لتمثيل محتوى الملفات والاستفسارات، ولاختيار الملفات اللازمة للمستخدم. إجراء عملية البحث على أساس المفردات و المطابقة المعجمية بدل استعمال دلالات الألفاظ في مثل هذه الأنظمة يمكن أن يقلل من دقة نتائج البحث، ويصعب إرضاء المستخدم وذلك لغياب الروابط المعنوية بين تمثيل الملفات و تمثيل الاستفسارات.

نقترح في هذا العمل منهجية جديدة لفهرسة النص العربي تلقائيا، والتي تقوم على أساس الكلمات والكلمات المرتبطة بها معنويا لإستخراج كل المعاني المحتواة في استفسار المستخدم الذي لا يعطي في كثير من الأحيان وصفا كاملا ودقيقا للحاجة المعلوماتية للمستخدم. ترتكز هذه المنهجية على الحسابات المعتمدة على تلازم الألفاظ، وعلى تقنيات التجميع لتصنيف الملفات وذلك بهدف إستخراج المعاني المختلفة للكلمة في مجالات استعمالها المختلفة.

كلمات مفتاحية

أنظمة البحث عن المعلومات في اللغة العربية ،الفهرسة الدلالية ، تقنيات التجميع، تصنيف الملفات، تلازم الألفاظ، التقاربات المعنوية.

TABLE DES MATIERES

Introduction générale.....1

PARTIE 1 : ETAT DE L'ART

CHAPITRE I : RECHERCHE D'INFORMATION

I.1. Introduction.....4
I.2. Concepts de base de la RI.....4
I.3. Processus de recherche d'information.....6
I.3.1. Le processus d'indexation.....7
I.3.1.1. Langages d'indexation.....7
I.3.1.2. Approches d'indexation.....8
I.3.1.3. Etapes de l'indexation automatique.....8
I.3.2. Processus d'appariement document / requête.....13
I.3.2.1. Appariement exact.....14
I.3.2.2. Appariement approché.....14
I.3.3. Processus de reformulation de requête.....14
I.3.3.1. Reformulation interactive de la requête.....14
I.3.3.2. Reformulation automatique de la requête.....14
I.4. Modèles de recherche d'information.....15
I.4.1. Le modèle booléen.....15
I.4.2. Modèle vectoriel.....15
I.4.3. Modèles probabiliste.....16
I.5. Recherche d'Information sur le Web.....17
I.5.1. Les outils de Recherche d'Informations.....17
I.5.1.1. Les moteurs de recherche.....17
I.5.1.2. Les annuaires.....17
I.5.1.3. Les métas-moteurs.....19
I.5.2. Algorithmes de classification de résultats.....19
I.5.2.1. Hyperlink-Induced Topic Search (HITS).....19
I.5.2.2. PageRank.....20
I.5.3 Défis de la RI sur le Web.....20
I.6. Evaluation des SRI.....21
I.6.1. Notion de pertinence.....21

I.6.1.1. Pertinence système 21
I.6.1.2. Pertinence utilisateur 22
I.6.2. Collections de test 22
I.6.3. Campagnes d'évaluation 23
I.6.4. Protocole d'évaluation d'un SRI 23
I.6.5. Mesures d'évaluation 23
I.7. Recherche d'information arabe 28
I.7.1. La langue arabe 28
I.7.2. Particularité de la langue arabe 28
I.7.3. Ressources linguistiques 30
I.7.3.1. Lexiques 30
I.7.3.2. Corpus 31
I.7.4. Outils de traitement automatique de la langue arabe 31
I.7.4.1. Analyseurs morphologiques 31
I.7.4.2 Racineurs 33
I.7.4.3 Lemmatiseurs 33
I.8 Conclusion 33

CHAPITRE II : Implémentation et Expérimentation

II.1. Introduction 34
II.2. Notations et définitions préliminaires 34
II.3. La Recherche d'Information basée sur une ressource externe 35
II.3.1. Les ressources externes 35
II.3.1.1. Les ressources structurées 35
II.3.1.1. Les ressources non structurées 36
II.3.2. Utilisation des ressources externes dans la RI 36
II.4. Indexation Sémantique et conceptuelle 36
II.4.1. Indexation sémantique 36
II.4.1.1. Les approches de désambiguïsation des sens des mots (WSD) 37
II.4.2. Indexation conceptuelle 40
II.4.2.1. La méthode de Woods 40
II.4.2.2. La méthode de Aggarwal and al 41
II.4.2.3. La méthode de Baziz 41
II.4.3. Similarité sémantique 41
II.4.3.1. Pondération des attributs 41

II.4.3.2. Types de mesure de similarité sémantique 42
II.4.3.3. Les applications de la similarité 43
II.5. Apprentissage Automatique 44
II.5.1. Apprentissage supervisé 44
II.5.2 Apprentissage non supervisé 44
II.5.2.1. Méthode de Sélection des caractéristiques en classification 45
II.5.2.2 Applications du Clustering 46
II.5.2.3 Les principales étapes du Clustering 47
II.5.2.4 Les méthodes du Clustering 50
II.6 Conclusion 55

PARTIE 2 : Contribution

CHAPITRE III : Ver une nouvelle approche d'indexation automatique de base sur la similarité des termes

III.1.Introduction 73
III.2.Problématique de la langue arabe 73
III.2.1.Propriétés morphologiques et difficultés de l'arabe 73
III.2.2.1. L'alphabet arabe 73
III.2.2.1. Morphologie arabe 74
III.2.2.3.Structure d'un mot 74
III.2.2.4. Les catégories des mots 75
III.2.2. Problèmes de traitement automatique de l'arabe 76
III.3. Approche proposée 76
III.3.1. Présentation de l'architecture proposée 76
III.3.1.1. l'Axe des Documents 77
III.3.1.2. l'Axe des Termes 77
III.3.2. Schéma général de l'approche 78
III.4. Conception détaillée 80
III.4.1. L'indexation 80
III.4.1.1. Prétraitement 80
III.4.1.3. Le regroupement des Termes (Clustering) 84
III.4.1.4 La création de l'Index 86
III.4.2. Exploitation de L'index 87
III.5 Conclusion 89

CHAPITRE IV : Implémentation et Expérimentation

IV.1. Introduction 73
IV.2. Environnement des expérimentations 73
IV.2.1. Matériel 73
IV.2.2. Logiciels 73
IV.2.2.1. Environnement de développement 73
IV.2.2.2. Outils utilisés 74
IV.3. Présentation de l'application 75
IV.3.1. RapidMiner Catégorisation des documents et traitement de texte 75
IV.3.1.1 Prétraitement des documents (Corpus) 76
IV.3.1.2 Extraction de List des termes 78
IV.3.1.3 Catégorisations des documents (documents clustering) 78
IV.3.1.4 Construction de matrix des termes 79
IV.3.2 Java Création et exploitation de l'index 80
IV.3.2.1. Terme Clustering 81
IV.3.2.2. Création de l'index (Cluster Index) 81
IV.3.2.2 Création de l'index (Lucene) 83
IV.3.3 Interface d'évaluation 85
IV.4 Expérimentation 86
IV.4.1 Contexte d'expérimentation 86
IV.4.2 Description de Corpus de Test 86
IV.4.2 Méthodologie de test 86
IV.4.3 Test et Résultats 86
IV.4.4. Interprétation 98
IV.5 Conclusion 99

Conclusion et prespectives 100

LISTE DES FIGURES

Figure I.1 : *SRI en réponse à une requête utilisateur*……5

Figure I.2 *:* Processus de recherche d'information……..7

Figure I.3 : *Courbe de distribution des mots selon la loi de Zipf*……10

Figure I.4 : *Informativité d'un terme en fonction de sa fréquence d'apparition dans un document*……..10

Figure I.5 : *Représentation d'un texte en fichier inverse*……13

Figure I.6 : *Distributions des documents dans une collection suite à une requête*……24

Figure I.7 : *Forme général de la courbe Précision-Rappel*……..25

Figure II.1 *: Processus d'apprentissage supervisé*……..44

Figure II.2 : *Processus d'apprentissage non supervisé*……..45

Figure II.3 *: Les différentes étapes du processus de clustering*……48

Figure II.4 *: Exemple d'une hiérarchie de clusters (dendrogramme)*……51

Figure II.5 *: Exemple de K-moyennes Clustering Process*……54

Figure III.1 : *Schéma général de l'approche proposée* ……62

Figure III.2 : *Processus du Prétraitement des documents* ……63

Figure III.3 : *Les mots vides arabes*……..64

Figure III.4 : *Structure de Fichier Inverse*……65

Figure III.5 *: Processus de Catégorisations des documents*……66

Figure III.6 *: Matrice des Documents*……66

Figure III.7 *: Processus de catégorisation des documents*.....67

Figure III.8 : *Processus de Clustering (regroupement) des termes*……68

Figure III.9 *: Matrice des Termes*.....68

Figure III.10 : *Processus de sélection de la catégorie d'un Requête*……71

Figure IV.1 : *Interface de RapidMiner Studio 7.4*……76

Figure IV.2 : *Processus de Prétraitement des documents*……76

Figure IV.3 : *Exemple de Processus de Prétraitement*……77

Figure IV.4 : *Exemple d'output de processus de Prétraitement*……78

Figure IV.5 : *Processus d'Extraction de la liste des termes*……78

Figure IV.6 : *Processus de Catégorisation des documents*……79

Figure IV.7 : *Exemple d'output de Processus de Catégorisation des documents*……79

Figure IV.8 : *Processus de Construction de matrix des termes*……80

Figure IV.9 : *Exemple de résultat de processus de construction de la matrix des termes*……80

Figure IV.10 : *Exemple de Matrice de Mutuelle Information……81*

Figure IV.11 : *Interface de Détermination de la catégorie d'une requête……82*

Figure IV.12 : *Interface de Recherche dans l'index base sur la clustering……83*

Figure IV.13 *Interface d'indexation par Lucene ……..……82*

Figure IV.14 : *Interface de Recherche par Lucene……85*

Figure IV.15 *: Interface d'Evaluation des Résultat de la Recherche……85*

Figure IV.16 : *La Valeur de Precision Moyenne par Requête……94*

Figure IV.17 : *Les valeurs de Mean Avrage Precision pour tous les index……92*

LISTE DES TABLEAUX

Tableau I.1 : *Comparaison entre les annuaires et les moteurs de recherche……..19*

Tableau I.2 *: Voyelles courtes arabes……..29*

Tableau I.3 : *Voyelles longues arabes……29*

Tableau II.1 : *Table de contingence de Chi2……46*

Tableau III.1 : Les 28 Lettre de L'Arabe…….56

Tableau III.2 : *Exemple de modèles pour le mot « قتل»……57*

Tableau III.3 : *Structure d'un mot Arabe……57*

Tableau III.4 : *Index initial (terme - Catégorie)……69*

Tableau III.5 : *Index Terme Cluster……70*

Tableau IV.1 : *Liste des préfixes et suffixes dans la racinisation légère……77*

Tableau IV.2 : *Listes des class d'indexation de Lucene……83*

Tableau IV.3 : *Listes des class de Recherche de Lucene……84*

Tableau IV.4 : *Description de Corpus BBC Arabic……84*

Tableau IV.5 *: Liste des Requêtes de test avec leurs documents pertinents……87*

Tableau IV.6 *: Liste des Requetés prétraite……88*

Tableau IV.7 *: Résultat de la sélection de la Catégorie des requêtes par les deux mesures……88*

Tableau IV.8 *: Exemple d'extraction des termes similaire par catégorie……89*

Tableau IV.9 *: Les Résultat du processus d'évaluation pour les requêtes de test……93*

Table IV.10 *: Les Résultat d'Evaluations par le mesure Mean Avrage Precision……91*

LISTE DES ABREVIATIONS ET SYMBOLES

AFP	Agence France Presse
AP	Averge precision
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
AWN	Arabic WordNet
BI	Besoin en Information
BIR	Binary Independence Retrieval
CDDL	Common Development and Distribution License
CLEF	Cross Language Evaluation Forum
CLIR	Cross Language Information Retrieval
DARPA	Defense Advanced Research Projects Agency
DF	Document-frequency
DIINAR1	DIctionnaire INformatisé de l'ARabe Version 1
DIINAR-MBC	DIctionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus
DixAF	Dictionnaire bilingue français arabe, arabe français
EWN	Européen WordNet
HITS	Hyperlink-Induced Topic Search
HTML	Hyper Text Markup Language
IA	Intelligence Artificielle
IDF	Inverse Document Frequency
IM	Information Mutuelle
IRSAD	Information Retrieval Syastem for Arabic Document
ISO	International Organization for standardization
JDK	Java Development Kit
KDDA	Knowledge Discovery and Data Analysis
LDC	Linguistic Data Consortium
LN	Langue Naturelle
MAP	Mean Average Precision
NIST	National Institute of Standards and technology
NN	Neural Network
NTCIR	NII Tesbeds and Community for information access research
OS	Operating System
OSAC	Open Source Arabic Corpora
PCC	distance du Clus Court Chemin
PRF	Pseudo Relevance Feedback
PRP	Probability Ranking Principle
PWN	Princeton WordNet
RF	Relevance Feedback
RI	Recherche D'Information
RIS	Recherche D'Information Structurée
RIT	Recherche D'Information Translinguistique
RL	Ressources Linguistiques
RSV	Retrieval Status Value
SGMI	Standard Generalized Markup language
SRI	Système de Recherche D'Information

TA	Traduction Automatique
TAL	Traitement Automatique de Langue
TC	Threshold Class
TF	Terme Frequency
TREC	Text Retrieval Evaluation Conference
URL	Uniform Ressource Language
UTF	Universal character set Transformation format
VM	Virtual Machine
WNA	WordNet Arabe
WSD	Word Sense Disambiguation
XML	eXtensible Markup Language

INTRODUCTION GENERALE

INTRODUCTION GENERALE

I. Contexte et problématique

La recherche d'information (RI) est un domaine qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information. Ce besoin est souvent formulé en langage naturel et exprimé sous forme d'une requête décrite par un ensemble de mot clés. L'essor du web et la popularisation d'internet ont propulsé l'information au premier plan et ont ainsi remis la RI face à de nouveaux défis, à savoir (i) retrouver une information pertinente dans une masse grandissante de documents qui change dans le temps et qui (ii) répond aux besoins spécifiques de l'utilisateur. En effet, la limite majeure des modèles classiques de RI réside en partie dans le fait qu'ils sont basés sur une approche généraliste, qui considère que le besoin en information est complètement représenté par sa requête et délivrant alors des résultats ne tenant compte que des critères de l'adéquation thématique entre les requêtes et les documents.

Dans les SRI classiques, les documents et les requêtes sont représentés(ou indexés) par des mots-clés, manuellement ou automatiquement extraits à partir de leur textes. Dans de tels systèmes, l'appariement (ou mise en correspondance) document-requête est lexical basé sur la présence ou l'absence des mots de la requête dans le document. Un document est alors considéré d'autant plus pertinent pour la requête qu'il a des mots clés en commun avec cette requête. Ou, les mots de la langue sont par nature ambigus. Un même mot utilisé dans le document et la requête peut définir des sens différents (cas de polysémie et d'homonymie), et plusieurs mots lexicalement différents utilisés dans le document et la requête peuvent refléter un même sens (cas synonymie). De ce fait, des documents pourtant non pertinents, contenant des mots de la requête, sont retrouvés, tandis que des documents sémantiquement pertinents, ne contenant aucun mots de la requête, ne sont pas retrouves.

Dans une outre parte qui concerne la recherche d'information arabe, cette langue qui est l'un des six langues officielles des Nations Unies, est langue maternelle de plus de 300 millions de personne. Est devenu un centre de la recherche et du Development commercial. Cela est dû au fait que la langue arabe souffre du manque de ressources électroniques et d'études rigoureuses, et nécessite de tel outils et application pour son traitement dans l'ère électronique.

II. Contribution

L'objectif de ce mémoire est d'apporter une contribution à la recherche d'information arabe, en proposant un nouveau système hybride d'indexation automatique de texte arabe basé sur la similarité des termes Afin de répondre aux problématiques listées précédemment.

L'idée de notre travail consiste à combiner deux Approche :

1. La première est la catégorisation de texte : qui permet d'associer une catégorie (ou classe) à un texte libre, en fonction des informations qu'il contient (les termes), l'objectif de l'utilisation de cette méthode est de réponde à la question : comment trouver l'information dont il a besoin, parmi toutes celles qui est accessible.
2. La seconde est une approche statique qui a pour but de capturer tout le sens voulu par le besoin d'informations qui n'est pas exprimé par la requête, et qui se base sur le principe des cooccurrences d'un mot, qui est basée sur un raisonnement simple :

si deux mots apparaissent souvent ensemble dans la même phrase, alors il est fort probable que ces deux mots aient un sens important, ou tout du moins une relation particulière, lorsqu'ils sont associés.

Dans Notre Approche on fait une démarche à partir de catégorisation de texte, par la suite on applique la Notion des termes similaires dans laquelle on établi le lien entre chaque terme avec leur termes associés (similaires).

Cette approche offre dans un premier temps de sélectionner l'ensemble des documents associés à une requête parmi l'ensemble des documents de différentes catégories par conséquence en augmente la précision par la limitation de bruit qui est effectuée par l'élimination des documents non pertinent (n'appartint pas à la catégorie associés au requête), par la suite d'utiliser les termes similaires pour sélectionner l'ensemble des documents liée aux termes de requête et leur termes similaires va augmenter le rappelé par la limitation de silence qui est effectué la capture de tous le sens de ces termes on utilise ci termes similaires.

Cette approche est focalisée sur la recherche d'information arabe par l'utilisation des méthodes de traitement automatique de langue arabe afin de prendre en considération sa particularité syntaxiquement, morphologiquement et sémantiquement.

III. Organisation du mémoire

Ce mémoire est organisé en quatre chapitres regroupés en deux parties.

La première partie présente le contexte qui a motivé nos travaux. Elle est composée de deux chapitres :

- Le premier chapitre présente les concepts de base de la RI. Nous commençons par donner une définition de la RI et Nous illustrons également le processus de RI en présentant les étapes d'indexation, d'interrogation et de mise en correspondance, ainsi que les techniques de reformulation des requêtes et l'appariement document-requête. nous décrivons les différents modèles servant de cadre théorique pour les Modèles de recherches d'informations de RI, En passant par la présentation des outils de RI sur le web, les algorithmes des moteurs de recherche et l'architecture générale de ce type d'outil. Par la suite on détaille les principales techniques d'évaluation d'un SRI, Enfin, nous terminons par la recherche d'information arabe.
- Le deuxième chapitre traite l'indexation sémantique et les méthodes d'apprentissage non supervisé. il est organisé en deux parties : dans la première, nous abordons les différentes définitions du contexte d'indexation sémantique, nous présentons également les notions de recherche d'information basées sur les ressources externes, En passant par l'indexation conceptuelle, sémantique et les différentes mesures de similarités. Dans la deuxième partie nous définissons l'apprentissage automatique et en détaillant les méthodes, les algorithmes et les principes concernant le clustering.

La deuxième partie qui concerne notre contribution et sa présentation détaillée. Elle comporte aussi deux chapitres.

- Dans le premier chapitre de cette partie, nous présentons notre nouvelle approche d'indexation automatique de texte arabe. Nous commençons par présenter les problèmes généraux lors du traitement et l'utilisation de langue arabe dans le domaine de recherche

d’information. Par la suite on présente l’architecture générale de notre approche et nous consacrons la suite du chapitre à détailler ses différentes fonctionnalités

- Dans le dernier chapitre, nous présentons l’implémentation de notre approche après avoir définit l’environnement de développement et les différents outils utilisés, Ce chapitre comporte aussi les résultats des tests effectués sur notre index afin d’évaluer et mesurer ses performances.

PARTIE I :

ETAT DE L'ART

CHAPITRE I :

RECHERCHE D'INFORMATION

Plan de Travail

I.1. Introduction 4
I.2. Concepts de base de la RI 4
I.3. Processus de recherche d'information 6
I.3.1. Le processus d'indexation 7
I.3.2. Processus d'appariement document / requête 13
I.3.3. Processus de reformulation de requête 14
I.4. Modèles de recherche d'information 15
I.4.1. Le modèle booléen 15
I.4.2. Modèle vectoriel 15
I.4.3. Modèles probabiliste 16
I.5. Recherche d'Information sur le Web 17
I.5.1. Les outils de Recherche d'Informations 17
I.5.2. Algorithmes de classification de résultats 19
I.5.3 Défis de la RI sur le Web 20
I.6. Evaluation des SRI 21
I.6.1. Notion de pertinence 21
I.6.2. Collections de test 22
I.6.3. Campagnes d'évaluation 23
I.6.4. Protocole d'évaluation d'un SRI 23
I.6.5. Mesures d'évaluation 23
I.7. Recherche d'information arabe 28
I.7.1. La langue arabe 28
I.7.2. Particularité de la langue arabe 28
I.7.3. Ressources linguistiques 30
I.7.4. Outils de traitement automatique de la langue arabe 31
I.7.5.Propriétés morphologiques et difficultés de l'arabe 33
I.7.6. Problèmes de traitement automatique de l'arabe 36
I.8 Conclusion 36

I.1. Introduction

L'être humain est doté de divers types de besoins qui seraient à l'origine de nombreux comportements permettant leur satisfaction**,** parmi lesquels, le besoin informationnel qui portrait l'individu à s'engager dans une activité de recherche d'informations.

La recherche d'information (RI) traite la représentation, le stockage, l'organisation et l'accès à l'information. Le but d'un système de recherche d'information (SRI) est de retrouver, parmi une collection de documents préalablement stockés, les documents qui répondent au besoin de l'utilisateur exprimé sous forme d'une requête. Pour cela, un SRI met en œuvre un ensemble de processus de sélection des documents pertinents pour la requête. (Fatiha, 2008)

Notre but à travers ce chapitre, est de présenter les fondements de base de la RI et des SRI. Après cette introduction, nous définissons les concepts de base de la RI. Nous décrivons ensuite les différentes étapes d'un processus de recherche d'information (indexation, appariement document-requête et reformulation de requête) et nous détaillons le processus d'indexation. Nous présentons par la suite les différents modèles de la RI. La section suivante est dédiée à la recherche d'information sur le Web. Finalement, nous présentons les principales techniques d'évaluation d'un SRI, et la recherche d'information arabe.

I.2. Concepts de base de la RI

La recherche d'information est traditionnellement considérée comme l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents, ceux qui sont susceptibles de répondre aux besoins de l'utilisateur. La gestion de ces informations implique le stockage, la recherche et l'exploration des documents pertinents. (BOURAMOUL, 2011)

De cette constatation, plusieurs concepts clés peuvent être définis, nous avons donc trouvé utile de les clarifier, A cet effet une synthèse des travaux de (BOURAMOUL, 2011), (Baziz, 2005) et (Kompaoré, 2008) nous a permis de dégager les concepts suivants :

- ***Recherche d'information (RI)***

La recherche d'information est l'opération de trouver un matériels (habituellement un document) de nature non structuré (habituellement un texte) qui satisfait un besoin d'information à partir d'une grande collection (généralement stockée sur les ordinateurs). (Christopher D. Manning, 2009)

La Recherche d'Information (*Information Retrieval en anglais*) est une discipline de l'information qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information. (G.Salton, Automatic Information Organization and Retrival, Livre, 1968)

- ***Système de Recherche d'Information (SRI)***

Le système de recherche d'information est un système informatique constitué d'un ensemble de programmes, dont l'objectif principal est de sélectionner, dans une collection de documents préalablement enregistrée, les informations (documents) pertinentes répondants à un besoin en information formellement exprimé par un utilisateur sous forme d'une requête. (AZZOUG, 2013)

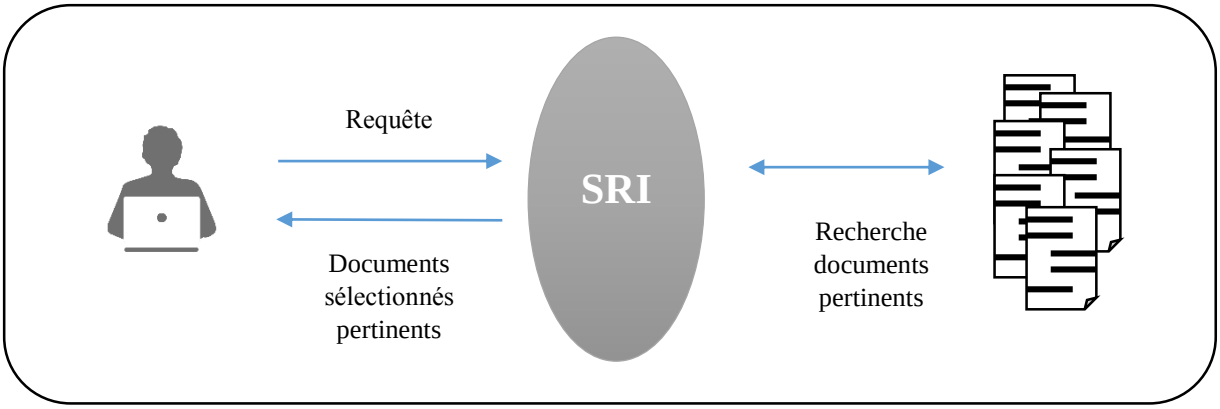


Figure I.1 : *SRI en réponse à une requête utilisateur*

- ***Document***

Le document constitue l'information élémentaire d'une collection de documents. L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document.

- ***Collection de documents***

La collection de documents (ou *fond documentaire*) constitue l'ensemble des informations exploitables et accessibles. Elle est constituée d'un ensemble de documents. Dans le cas général et pour un souci d'optimalité, la base constitue des représentations simplifiées mais suffisantes pour ces documents. Ces représentations sont étudiées de telles sortes que la gestion (ajout suppression d'un document) ou l'interrogation (recherche) de la base se font dans les meilleures conditions de coût.

- ***Besoin d'information***

Un besoin d'information (BI) est l'expression de critères que l'utilisateur souhaite voir apparaître dans les éléments du résultat. Les critères n'ont pas nécessairement la même priorité. Un BI est en général défini par un ensemble de concepts recherchés. Suivant les contextes, d'autres critères s'ajoutent à la définition d'un BI comme par exemple des indications de préférence associées aux concepts (concepts souhaités ou concepts non souhaités).

- ***Requête***

La requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le SRI et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature. Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique.

- ***Modèle de représentation***

Un modèle de représentation est un processus permettant d'extraire d'un document ou d'une requête, une représentation paramétrée qui couvre au mieux son contenu sémantique. Ce processus de conversion est appelé indexation. Le résultat de l'indexation constitue le descripteur du document ou de la requête, qui est une liste de termes ou groupes de termes (concepts) significatifs pour l'unité textuelle correspondante, auxquels sont associés

généralement des poids, pour différencier leurs degrés de représentativité du contenu sémantique de l'unité en question.

- ***Modèle de recherche***

Il représente le modèle du noyau d'un SRI. Il comprend la fonction de décision fondamentale qui permet d'associer à une requête, l'ensemble des documents pertinents à restituer. Il est utilisé pour la recherche d'informations proprement dite et est étroitement lié au modèle de représentation des documents et de la requête. Ces modèles de recherche représentent ce qui différé le plus entre les SRI. Ils sont inspirés de concepts mathématiques afin de pouvoir évaluer certaine relation, notamment la relation d'appariement entre requête et les documents. (Rijsbergen, 2004)

Le modèle de recherche pour un SRI indique comment les documents et requêtes sont représentés et comment ces représentations sont comparées pour évaluer la pertinence

I.3. Processus de recherche d'information

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

Un système de recherche d'information manipule un corpus de documents qu'il transpose à l'aide d'une fonction d'indexation en un corpus indexé. Ce corpus lui permet de résoudre des requêtes traduites à partir des besoins des utilisateurs. Un tel système repose sur la définition d'un modèle de recherche d'information qui effectue ces deux transpositions et qui fait correspondre les documents aux requêtes. La transposition d'un document en un document indexé repose sur un modèle de document. De même, la transformation du besoin utilisateur en requête repose sur un modèle de requête. Enfin, la correspondance entre une requête et des documents s'établit par une relation de pertinence (L.Maisonnasse, 2008).

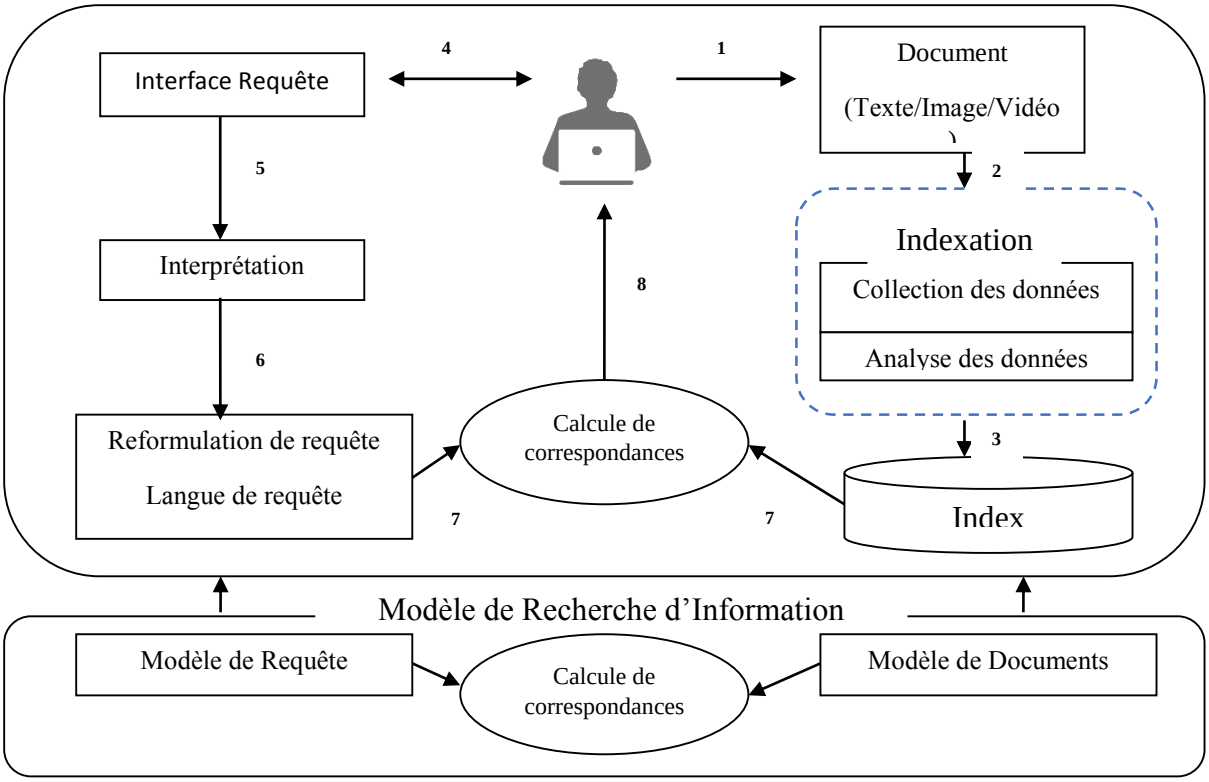


Figure I.2 : *Processus de recherche d'information (Charhad, Novembre 2005)*

I.3.1. Le processus d'indexation

L'indexation est une phase très importante pour un SRI car de sa qualité dépend la qualité des réponses du système et donc les performances de ce dernier. Une bonne indexation doit permettre de retrouver tous les documents pertinents au besoin de l'utilisateur et pas (ou peu) de documents non pertinents pour celui-ci.

L'indexation consiste à extraire des documents les mots les plus discriminants encore appelés index. Cette première tâche est généralement effectuée en marge du processus de recherche car, la construction des index peut être assez longue en fonction du nombre de documents de la collection ainsi que de la taille des documents. Les index ont un caractère réducteur car tous les termes d'un document ne sont pas importants à prendre en compte pour la recherche (Kompaoré, 2008).

Techniquement, l'indexation peut-être manuelle, automatique ou semi-automatique, basé sur un langage libre ou contrôlé.

I.3.1.1. Langages d'indexation

I.3.1.1.1. Langage libre

Le langage libre est un langage évolutif, proche de notre langue naturelle (LN). Son vocabulaire est l'ensemble des éléments qui composent le langage , il est choisi à posteriori et n'est pas limité par un contrôle. Le vocabulaire est donc composé de tous les descripteurs choisis librement pour indexer les documents (GASAMI, 2009). Le langage libre n'est pas régi par une syntaxe car aucune contrainte n'est spécifiée à priori. Par conséquent, le vocabulaire évolue rapidement et peut contenir des synonymes, polysémiques etc. Ce qui entraîne des

incohérences et diminue les performances du système de recherche d'information (Catherine, 2010).

I.3.1.12. Langage contrôlé

Le langage contrôlé ou langage documentaire est un langage normalisé, il sert à éviter les problèmes d'ambiguïté (dû à l'homonymie et à la polysémie de certains termes) ainsi que les problèmes de redondance (synonymie, etc.) du langage libre. Une liste de termes d'indexation est définie; cette liste appelée liste d'autorité, ainsi, un terme d'indexation ne possède qu'un seul sens et inversement un sens n'est associé qu'à un seul terme d'index action donc les descripteurs retenus seront les seuls mots clés acceptés lors de la requête (Catherine, 2010).

L'indexation en langage contrôlé réduit le nombre de représentations possibles d'un document. Cela n'empêche pas l'indexation d'être subjective si elle est réalisée par un sujet humain, même si les sens et les termes sont bien délimités.

I.3.1.2. Approches d'indexation

I.3.1.21. Indexation manuelle

C'est un opérateur humain, généralement expert du domaine, qui se charge de caractériser, selon ses connaissances propres, le contenu sémantique d'un document. Cette approche présente deux inconvénients :

- elle est subjective, puisque le choix des termes d'indexation dépend de l'indexeur et de ses connaissances du domaine
- elle est pratiquement inapplicable aux corpus de textes volumineux

I.3.1.2.2. Indexation automatique

C'est un processus complètement automatisé qui se charge d'extraire les termes caractéristiques du document. L'intérêt d'une telle approche réside dans sa capacité à traiter les textes nettement plus rapidement que l'approche précédente, et de ce fait, elle est particulièrement adaptée aux corpus volumineux. L'indexation automatique est l'approche la plus étudiée en RI, nous la détaillons en section suivante

I.3.1.2.3. Indexation semi-automatique (indexation supervisée)

C'est une combinaison des deux approches d'indexation précédentes. Dans ce cas, les indexeurs utilisent un vocabulaire contrôlé sous forme de thésaurus ou de base terminologique. Le choix final des termes d'indexation à partir du vocabulaire fourni, est laissé ainsi à l'indexeur humain (généralement spécialiste du domaine).

I.3.1.3. Etapes de l'indexation automatique

L'indexation automatique repose sur un ensemble des traitements automatisés sur les documents, qui sont les suivants : *L'analyse lexicale* (l'extraction automatique des mots), *l'émanation des mots vides*, *la lemmatisation*, et la *pondération des termes* et la création de l'index.

I.3.1.1. L'analyse lexicale

L'analyse lexicale est l'étape qui permet de convertir le texte d'un document en un ensemble de termes. Un terme est une unité ou entité lexicale, appelée aussi Token en anglais. L'analyseur lexical permet de détecter les espaces de séparation des mots, les chiffres, les ponctuations, etc… et les supprimer (El GHALI, 2016).

I.3.1.2. L'élimination des mots vides

Les mots vides (pronoms, personnels, prépositions, conjonctions de coordination, …etc.) sont des mots non significatifs dans un document, car ils ne traitent pas le sujet du document (Hlaoua L. , 2007).

L'élimination des mots vide consiste à la suppression de ces mots qui sont trop fréquents dans les documents de la collection et, par conséquent, n'affectent pas le sens du contenu. Ces mots sont ceux qui se répètent pratiquement dans tous les documents de la collection, ils ne sont pas considérés comme des termes à indexer. (R. A. Baeza-Yates & B. A. Ribeiro-Neto, 1999)

Il existe deux techniques pour éliminer les mots vides :

- L'utilisation d'une liste de mots vides (appelée anti-dictionnaire ou stoplist en anglais).
- L'élimination des mots qui dépassent un certain nombre d'occurrences dans la collection (permet de réduit la taille de l'index, et améliore le temps de réponse du système).

I.3.1.3. La normalisation (lemmatisation ou radicalisation)

Cette phase est liée à la lemmatisation (ou radicalisation), il s'agit d'un traitement morphologique des mots permettant de regrouper les variantes d'un mot. En effet, dans un texte, il peut y avoir différentes formes d'un mot désignant le même sens. Le but de ce processus est de les représenter par un seul mot qui porte un concept commun (ex. biologie, biologiste, biologique, par : biologie) (BADACHE, 2016).

Elle se réfère à la transformation d'un mot (terme) à sa base ou sa racine. L' utilisateur spécifie souvent un mot dans une requête, mais un document pertinent ne sera pas retourné si une variante de ce mot et pas le mot lui-même existe dans le document. Les préfixes et les suffixes sont des exemples des variations syntaxiques qui empêchent un mappage parfait entre les termes de la requête et ceux du document correspondant. (R. A. Baeza-Yates & B. A. Ribeiro-Neto, 1999)

Il reste cependant à mentionner que ces traitements peuvent induire certains inconvénients tels que la production de normalisation agressive, par exemple les mots university/universe, organization/organ, policy/police sont normalisée par l'algorithme de Porter (Ourdia, 2007).

I.3.1.4. La pondération des termes

Cette étape vient après l'identification des termes des documents et leur normalisation. Les termes qui représentent un document n'ont pas la même importance. Donc, La pondération permet d'associer à chaque termes un poids qui représente son importance dans la collection de documents. Cette importance est souvent calculée sur la base d'aspects statistiques (ou parfois linguistiques). L'objectif est de trouver les termes qui représentent le mieux le contenu d'un document.

Les approches statistiques tirent leur origine de la loi de *Zipf*, et de la conjecture de *Luhn*

- **Loi de Zipf**

La loi de Zipf est une loi empirique énoncée en 1949 par G.K Zipf. (G.K Zipf , 1949) Selon Zipf, les mots dans les documents ne sont pas organisés de manière aléatoire mais suivant une loi inversement proportionnelle à leurs rangs, qui représentent les classements des mots dans la

liste décroissante des fréquences d'apparition des mots dans le corpus. Ainsi, la fréquence du second mot le plus fréquent dans le corpus est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. Formellement, cette loi s'exprime par l'égalité approximative suivante :

$$Fréquence \times Rang \cong Constante$$

Selon la loi de Zipf, la distribution des mots suit la courbe illustrée dans la figure suivante :

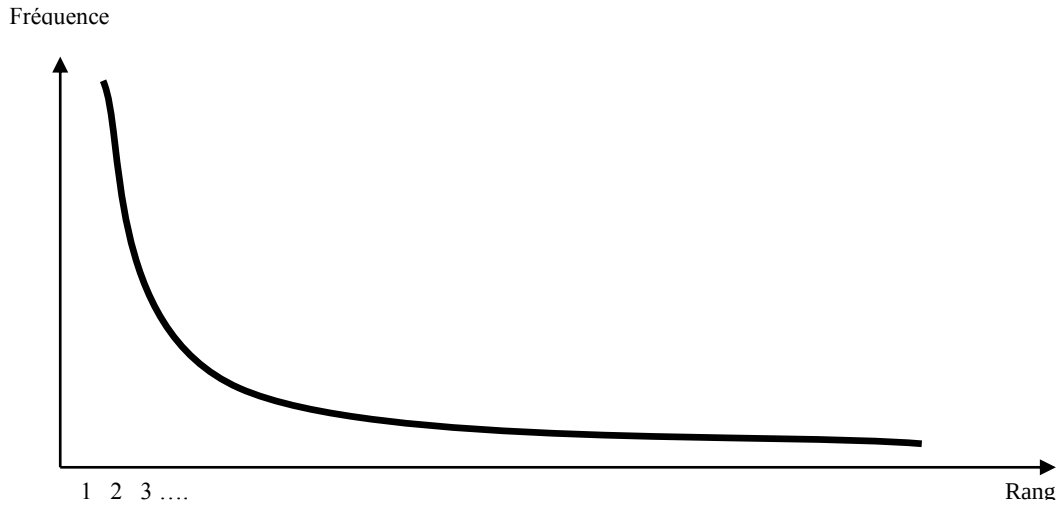


Figure I.3 : *Courbe de distribution des mots selon la loi de Zipf.*

Zipf qui ne put expliquer cette curieuse loi, invoqua le principe du moindre effort : il considéré qu'il est plus facile pour un auteur d'un document de répéter certains termes que d'en utiliser de nouveaux. C'est la conjecture de Luhn (Luhan, 1958) qui permit de définir comment utilisé la loi de Zipf pour sélectionner les termes représentatifs d'un document (Ourdia, 2007).

- **Conjecture de Luhn**

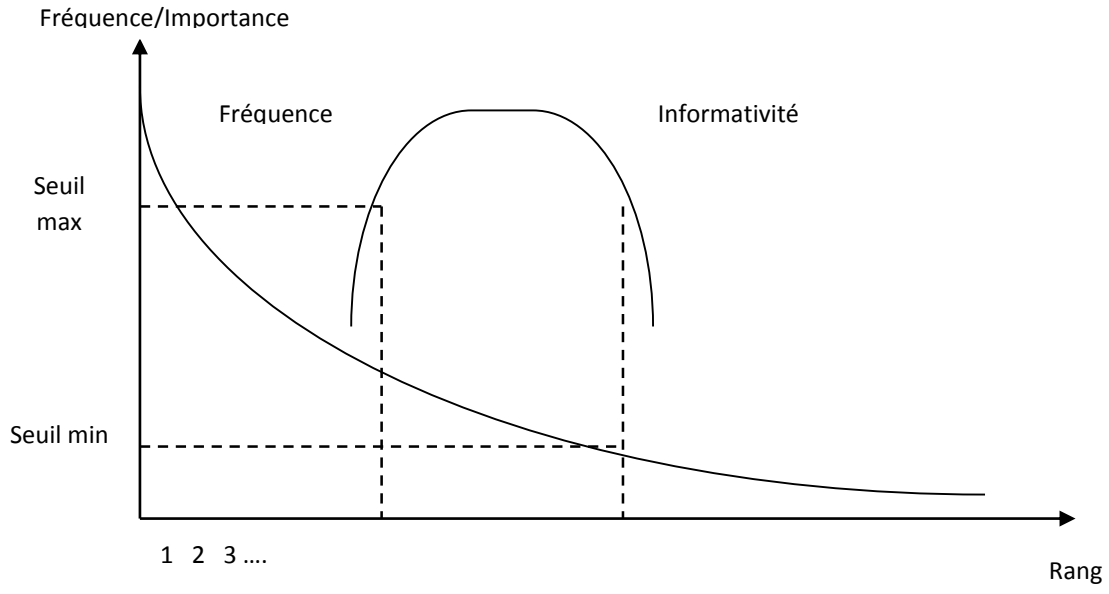


Figure I.4 : *Informativité d'un terme en fonction de sa fréquence d'apparition dans un document.*

La loi de Zipf est la base de conjecture de Luhn. Cette dernière émet une hypothèse sur l'information contenue dans les termes (informativité) d'un document schématisée par les deux courbes de fréquence et d'informativité dans la Figure I.4 Elle considère que les mots qui reviennent souvent de rangs faibles (très fréquents), sont non significatifs car ils n'ont pas de pouvoir discriminant. Elle considéré aussi ceux de rangs élevés (très rares), comme peu significatif. En effet ces mots sont rares et donc peu utilisés. Les termes significatifs sont ceux de rangs intermédiaires. Cette conjecture permet de diminuer la taille des index des documents. Deux seuils de fréquence (*seuil max* et *seuil min* dans la Figure I.4), sont fixés pour éliminer les termes dont le contenu informatif est jugé faible. Seuls les termes entre ces deux seuils sont alors significatifs pour représenter les documents.

A partir de ces constatations, différentes techniques de pondération ont vu le jour. Parmi les nombreuses formules de pondération définies dans le domaine, la mesure

TF-IDF est de loin la plus connue et la plus utilisée. Ainsi que sa version normalisée appelée LTC. (Goker & Davies, 2009), (Manning, Raghvan, & Schutze, 2009), (Rijsbergen, 2004)

- **Pondération en TF-IDF**

La formule de pondération Tf-Idf est construite par la combinaison de deux facteurs :

- Un facteur de pondération locale (***TF : Terme Frequency***), qui quantifie la représentativité locale d'un terme dans le document. La pondération tf_{ij} du terme t_i dans le document D_j est donnée selon l'une des formulations suivants : (Baziz, 2005)

$$tf_{ij} = 1 \; ou \; 0 \tag{I.1}$$

Dans le cas d'une **fonction binaire** qui rend 1 si le terme apparait dans le document une ou plusieurs fois, et rend 0 si non.

$$tf_{ij} = freq(t_i, D_j) \tag{I.2}$$

Dans le cas d'une **fonction brute** de calcul de la fréquence qui compte le nombre d'occurrence du terme t_i dans le document D_j.

$$tf_{ij} = \alpha + \log freq(t_i, D_j) \tag{I.3}$$

Dans le cas d'une **fonction logarithmique**, ou α est une constante. Cette fonction a pour but d'atténuer les effets de larges différences entre les fréquences d'occurrence des termes dans le document.

$$tf_{ij} = \frac{freq(t_i, D_j)}{max_{t_i \in D_j} (freq(t_i, D_j))} \tag{I.4}$$

Dans le cas d'une **fonction normalisée**, où $max_{t_i \in D_j} (freq(t_i, D_j))$ est la fréquence maximale des termes dans D_j ; cette fonction permet de réduire les différences entre les valeurs associées aux termes du document.

$$tf_{ij} = \frac{freq(t_i, D_j) \times (K_1 + 1)}{K_1 \times ((1-b) + b \times \frac{dl}{avdl})} \tag{I.5}$$

Dans son modèle BM25 utilise une normalisation, de la fréquence d'un terme ti dans un document Dj en fonction de la longueur de Dj noté par dl et la moyenne de longueur des documents dans la collection traitée avdl, où K1et b sont des constantes.

- Un second facteur, de pondération globale (***IDF : Inverse Document Frequency***), qui mesure la représentativité globale du terme vis-à-vis de la collection des documents, En effet, un terme fréquent dans la collection, a moins d'importance qu'un terme moins fréquent. Cette mesure est exprimée selon l'une des déclinaisons suivantes :

$$IDF_i = \log(\frac{N}{n_i}) \quad \text{(I.6)}$$

$$IDF_i = \log(\frac{N}{n_i + 1}) \quad \text{(I.7)}$$

Avec N représente le nombre de documents de la collection et n_i est le nombre de documents dans lesquels le terme t_i apparaît.

La formule TF-IDF pour le calcul du poids d'un terme t_i dans un document d_j=[r_{1j}, r_{2j},… r_{ij},…, r_{mj}] est donnée par : (Hamada , El Hady, & Abd El-Wahed, 2010)

$$r_{ij} = tf_{ij} \times idf_i \quad \text{(I.8)}$$

La mesure tf ∗ idf donne une bonne approximation de l'importance du terme dans le document, particulièrement dans les corpus de documents de taille homogène. Cependant, elle ne tient pas compte d'un aspect important du document : sa longueur. En général, les documents les plus longs ont tendance à utiliser les mêmes termes de façon répétée, ou à utiliser plus de termes pour d'écrire un sujet. Par conséquent, les fréquences des termes dans les documents seront plus élevées, et les similarités à la requête seront également plus grandes. Pour pallier cet inconvénient, Robertson (E.Robertson & S.Walker, 1994) et Singhal et al. (A.Singhal, G.Salton, M.Mitra, & C.Buckley, 1996) Proposent d'intégrer la taille des documents à la formule de pondération, on parle de facteur de normalisation.

- **LTC**

L'approche de pondération LTC emploie également les deux facteurs de pondération locale et globale, mais avec une formule différente :

$$r_{ij} = \frac{\log(tf_{ij} + 1) \times idf_i}{\sqrt{\sum_{k=1}^{m}[\log(tf_{ik} + 1) \times idf_i]}} \quad \text{(I.9)}$$

En utilisant le logarithme du facteur de pondération locale, qui représente la fréquence d'apparition d'un terme dans le document, la formule présentée par l'équation (I.9) réduit les effets de grandes différences dans les fréquences.

I.3.1.5. La création de l'index

L'index, défini comme étant la structure de stockage utilisée pour mémoriser les informations sélectionnées pour la représentation du texte. Cette structure permet de sélectionner, pour

n'importe quel terme, tous les documents où il apparaît. Plusieurs solutions de stockage ont été proposées parmi lesquelles : les fichiers inverses (inerte files), les fichiers de signatures (signature files) et les tableaux de suffixes (suffixe Arras).

La solution la plus utilisée actuellement est celle des fichiers inverses. Ces fichiers sont composés de deux éléments principaux: Le dictionnaire et le fichier potin. Le vocabulaire consiste en la liste de tous les mots distincts extraits du texte. Pour chaque mot est assigné l'ensemble des positions dans lesquelles ce dernier apparaît (P. Bellot, 2000).

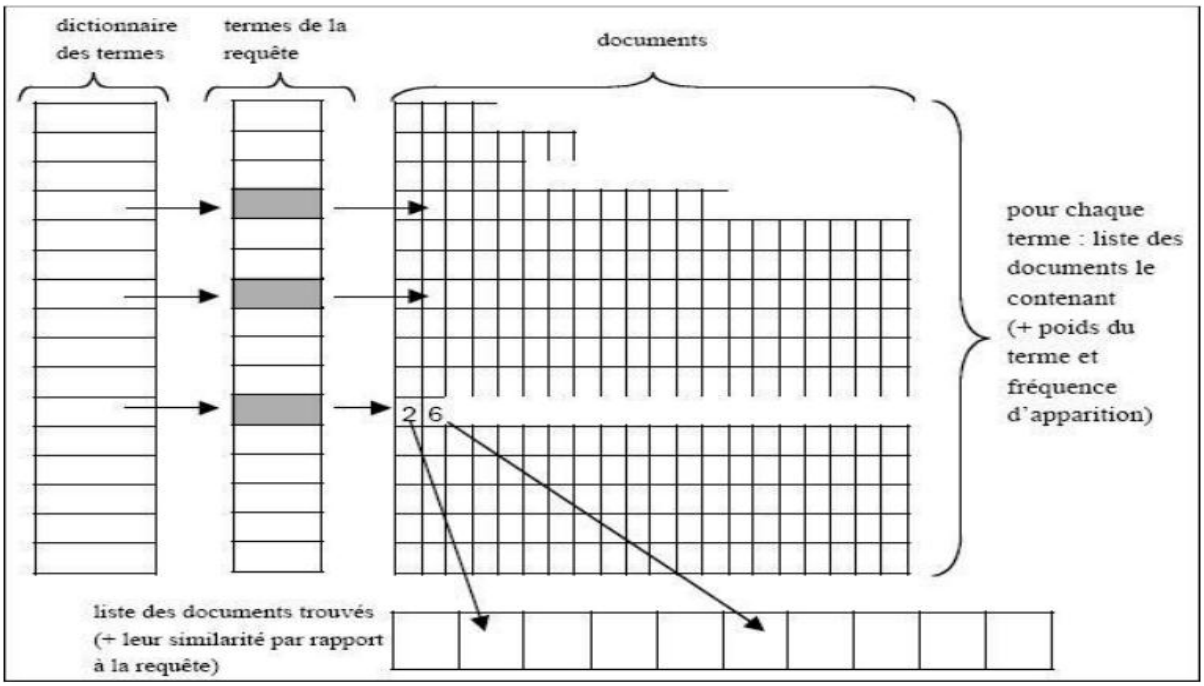


Figure I.5 : *Représentation d'un texte en fichier inverse (P. Bellot, 2000)*

I.3.2. Processus d'appariement document / requête

Le processus de mise en correspondance ou d'appariement document-requête revient à mesurer un score, supposé représenter la pertinence d'un document vis-à-vis d'une requête, en calculant un score de correspondance entre la représentation de chaque document et celle de la requête. Ce score traduit un degré de pertinence système, que l'on essaye de rapprocher le plus possible du jugement de pertinence de l'utilisateur vis-à-vis du document. Cette valeur est calculée en utilisant une fonction ou une probabilité de similarité appelée : **RSV (q, d)** (Retrieval Statu Value), où q est une requête et d'un document. Cette mesure tient compte des poids des termes calculés dans l'étape précédente du processus de recherche d'information. (Ourdia, 2007)

La fonction de correspondance est très étroitement liée aux opérations d'indexation et de pondération des termes de la requête et des documents du corpus. Notons que d'une façon générale, l'appariement document-requête et le modèle d'indexation permettent de caractériser et d'identifier un modèle de recherche d'information. Il existe un certain nombre de modèles théoriques, dans la littérature les plus connus étant le « Modèle Booléen », le « Modèle Vectoriel », et le « Modèle Probabiliste ». (El GHALI, 2016)

Il existe deux types d'appariement :

I.3.2.1. Appariement exact

Le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés.

I.3.2.2. Appariement approché

Le résultat est une liste de documents censés être pertinents pour la requête. Les documents retournés sont triés selon un ordre de mesure qui reflète le degré de pertinence des documents par rapport à la requête.

I.3.3. Processus de reformulation de requête

En RI, l'utilisateur formule son besoin en information par le biais d'une requête dans l'espoir de trouver des réponses pertinentes à ce qu'il recherche. La qualité des réponses dépend d'une part des termes utilisés par l'utilisateur pour formuler son besoin, et d'autre part des termes utilisés dans l'indexation des documents. Du fait de l'ambiguïté/imprécision de la langue naturelle, l'utilisateur peut formuler sa requête dans un vocabulaire différent de celui utilisé par les auteurs/indexeurs des documents, ce qui a pour conséquence d'influer négativement sur la qualité des résultats de la recherche. Pour résoudre ces problèmes, Van Riejsbergen (Rijsbergen, 2004) introduit le mécanisme de reformulation de requête dans les SRI.

La reformulation de la requête est alors considérée comme un processus ayant pour objectif de générer une nouvelle requête plus ciblée permettant d'obtenir des résultats de recherche plus pertinents que ceux obtenus par la requête initialement formulée par l'utilisateur. Le processus de reformulation de requête se base sur les deux étapes suivantes : la première consiste à étendre la requête initiale par des termes jugés pertinents par l'utilisateur ou par le système de recherche, et la seconde consiste à réajuster les poids des termes de la nouvelle requête (AZZOUG, 2013).

Deux techniques de reformulation sont à distinguer : la reformulation interactive et la reformulation automatique.

I.3.3.1. Reformulation interactive de la requête

Appelée aussi *réinjection de pertinence* ou *relevance feedback* en anglais. Elle consiste à présenter à l'utilisateur une liste de documents jugés pertinents en réponse à la requête initiale. C'est à l'utilisateur de sélectionner à partir des documents pertinents ceux dont lesquels le système va extraire les termes à rajouter à la requête initiale dans le but d'effectuer une nouvelle recherche (Boujamaa & Tawes, 2009). Ces termes sont choisis en se basant sur un poids calculé par un algorithme de sélection de termes. La repondération des termes avant qu'ils soient ajoutés à la nouvelle requête, et l'interaction avec l'utilisateur peuvent servir pour améliorer globalement la performance du SRI.

I.3.3.2. Reformulation automatique de la requête

Les SRI peuvent comporter des composantes supplémentaires pour étendre automatiquement la requête. L'extension de la requête est faite sans intervention de l'utilisateur grâce à l'utilisation d'un thésaurus contenant des informations de type linguistique (équivalence,

association, hiérarchie) et statistique (pondération des termes) (Boujamaa & Tawes, 2009). La construction de ce thesaurus est faite pendant le processus d'indexation de manière interactive ou automatique.

I.4. Modèles de recherche d'information

Si l'indexation permet de déterminer les termes représentatifs des documents et requêtes, alors c'est le modèle de RI qui assure leur interprétation dans un formalisme de représentation propre, et qui offre le mécanisme de leur appariement en vue de calculer les degrés de pertinence des documents pour les requêtes.

Un *modèle* de RI a pour rôle de fournir une formalisation du processus de RI et un cadre théorique pour la modélisation de la mesure de pertinence.

Un modèle de RI est défini par un quadruplet (D, Q, F, R (q, d)) :

– D est l'ensemble de documents

– Q est l'ensemble de requêtes

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

– R (q, d) est la fonction de pertinence du document d à la requête q

Nous présentons par la suite les principaux modèles de RI : le modèle booléen, le modèle vectoriel et le modèle probabiliste. (BOURAMOUL, 2011)

I.4.1. Le modèle booléen

Le modèle booléen, ou logique, est le modèle de base qui repose sur la théorie des ensembles et modélise les documents et les requêtes selon des ensembles de termes pondérés de façon binaire, correspondant à la présence ou l'absence du terme dans le document. Les termes au sein des documents et des requêtes sont reliés par des opérateurs logiques : OU (∨), ET (∧) et NON (¬). (Salton, 1971)

Le score de similarité entre un document d_i et une requête q_h est inclu dans l'ensemble {0, 1} et repose sur un appariement exact entre l'expression logique L (d_i) du document di et l'expression logique L (q_h) de la requête q_h, comme défini ci-après :

$$RSV\ (d_i, q_h) = \begin{cases} 1 & si\ L(d_i) = L(q_h) \\ 0 & sinon \end{cases} \qquad \text{(I.10)}$$

I.4.2. Modèle vectoriel

Dans ce modèle , la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Le modèle vectoriel Salton (G.Salton, 1971) représente les documents et les requêtes par des vecteurs d'un espace à dimensions, les dimensions étant constituées par les termes du vocabulaire d'indexation. L'index d'un document d_j est le vecteur$\overrightarrow{d_j} = (W_{1,j}, W_{2,j}, W_{3,j}, \ldots, W_{n,j})$, où $W_{k,j} \in [0\ 1]$ dénote le poids du terme t_k dans le document d_j. Une requête est également représentée par un vecteur $\vec{q} = (W_{1,q}, W_{2,q}, W_{3,q}, \ldots, W_{3,q})$, où $W_{k,q}$ est le poids du terme dans la requête q.

La fonction de correspondance mesure la similarité entre le vecteur requête et les vecteurs documents. Il existe à ce jour plusieurs mesures dont les plus connues sont les suivantes

- *Le produit scalaire :*

$$RSV\ (\vec{q},\overrightarrow{d_j}) = \cos(\vec{q},\overrightarrow{d_j}) \tag{I.11}$$

- *La mesure de cosinus* ou *RSV* $(\vec{q}, \overrightarrow{d_j})$.

$$RSV\ \left(\vec{q},\overrightarrow{d_j}\right) = \frac{\vec{q} \times \overrightarrow{d_j}}{\|\vec{q}\| \times \|\overrightarrow{d_j}\|} \tag{I.12}$$

$ou\ \|x^{\rightarrow}\|\ répresent\ le\ nombre\ euclidienne\ du\ vectore\ x^{\rightarrow}$

- *La mesure de Jaccard ou RSV* $(\vec{q}, \overrightarrow{d_j})$.

$$RSV\ \left(\vec{q},\overrightarrow{d_j}\right) = \frac{\left|\vec{q} \cap \overrightarrow{d_j}\right|}{\left|\vec{q} \cup \overrightarrow{d_j}\right|} ou\ \left|\vec{q} \cap \overrightarrow{d_j}\right| \tag{I.13}$$

Elle correspond au nombre de termes présents à la fois dans la requête q et le document d_j tandis que $|\vec{q} \cup \vec{d_j}|$ représente le nombre de termes contenus dans la requête q ou le document d_j.

- *La mesure de Dice ou RSV* $(\vec{q}, \overrightarrow{d_j})$.

$$RSV\ \left(\vec{q},\overrightarrow{d_j}\right) = \frac{2 \times \left|\vec{q} \cap \overrightarrow{d_j}\right|}{|\vec{q}| + |\overrightarrow{d_j}|} \tag{I.14}$$

Où $\left|\overrightarrow{d_j}\right|$, respectivement $|\vec{q}|$, annote le nombre de termes dans le document $d_j$respectivement la requête q.

Plus les vecteurs sont similaires, plus l'angle formé est petit, et plus le cosinus de cet angle est grand. A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante. (Salton, 1971)

I.4.3. Modèles probabiliste

L'approche classique d'un modèle probabiliste consiste à évaluer par une probabilité le degré de pertinence que peut avoir un document par rapport à une requête (ROBERTSON, 1977). Le calcul d'une telle probabilité, que l'on peut noter P (Pert| D_i), utilise les principes fondamentaux de la théorie des probabilités :

$$P(Pert|D_i) = \frac{P(Pert)}{P(D_i)} \times P(D_i|Pert) \tag{I.15}$$

Où P(Pert) désigne la probabilité qu'un document quelconque soit pertinent pour la requête, $P(D_i)$ la probabilité d'obtenir le document D_i. Cette probabilité peut se décomposer comme suit :

$$P(D_i) = P(D_i|Pert) \times P(Pert) + P(D_i|\overline{Pert}) \times P(\overline{Pert}) \quad \text{(I.16)}$$

Pour calculer P (D_i| Pert) (ou $P(D_i|\overline{Pert})$), on considère que le document D_i peut être représenté par les termes t_j qui l'indexent (à la manière des approches vectorielles) et on suppose que ces termes sont mutuellement indépendants. On a alors

$$P(D_i|Pert) = \prod_{j=1}^{j=m} P(t_j|Pert) \quad \text{(I.17)}$$

$$P(t_j|Pert) = \frac{r_j}{R}\left(et\ P(t_j|\overline{Pert}) = \frac{n_j - r_j}{N - R}\right) \quad \text{(I.18)}$$

Où R est le nombre de documents pertinents pour la requête, r_j le nombre de documents pertinents contenant le terme t_j, N le nombre total de documents et n_j le nombre de documents contenant t_j.

I.5. Recherche d'Information sur le Web

Le web est la collection de documents la plus large dans le monde. A titre indicatif, le projet d'information de Berkeley "How Match Information", en 2003, estimait que la quantité d'information du web est 20 fois plus la taille de la collection entière de la bibliothèque du congress. (Berkeley, 2003)

I.5.1. Les outils de Recherche d'Informations

Il existe de nombreux outils de recherche d'information sur le Web, ces outils se spécialisent en fonction des services utilisés et du type d'information qu'ils fournissent. Aujourd'hui toute interface d'interrogation et de recherche est le plus souvent qualifiée de moteur de recherche et ce, quelle que soit la source interrogée et le système informatique utilisé. (Largouet Anita, 2005)

En se basant sur le mode de fonctionnement, On peut distinguer donc trois catégories d'outils pour la recherche d'information sur le web : les moteurs de recherche, les annuaires et les méta-mouturs. (ALI KHOUDJA, 2015)

I.5.1.1. Les moteurs de recherche

Un moteur de recherche est une application qui permet de retrouver des ressources (pages web, images, vidéo, lieu, fichiers, etc.) associées à des mots-clés soumis par un utilisateur.

Ces outils de recherche sur Internet sont constitués de trois parties :

- *Le robot (encore appelés spider ou crawlers) :* le robot parcourent le site à intervalles réguliers et de façon automatique pour découvrir de nouvelles adresses (URL). Ils suivent les liens hypertextes rencontrés sur chaque page atteinte ;
- *L'index :* le moteur indexe chaque page identifiée dans une base de données accessible par l'internaute ;

- *Logiciel d'interrogation :* représenté par l'interface à travers laquelle les internautes peuvent interroger l'index en utilisant des mots-clés ;

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

- *Le correcteur orthographique :* il permet de corriger les erreurs introduites dans les mots de la requête, et s'assurer que la pertinence d'un mot sera bien prise en compte sous sa forme canonique ;
- *Le lemmatiseur :* il permet de réduire les mots recherchés à leur lemme et ainsi d'étendre leur portée de recherche ;
- *L'anti dictionnaire :* utilisé pour supprimer à la fois dans l'index et dans les requêtes tous les mots «vides" (tels que «de", "le", "la") qui sont non discriminants et perturbent le score de recherche en introduisant du bruit.

Enfin la popularité des moteurs de recherche n'est pas absolue, généralement, nous ne considérons que le moteur de recherche Google. Toutefois, malgré le fait qu'il a envahi un grand pourcentage du marché mondiale, son hégémonie n'est pas aussi importante dans le monde entier, car certains pays résistent encore à cet envahisseur : Russie, Chine …etc.

I.5.1.2. Les annuaires

L'annuaire (*directory en anglais*) est une liste de liens subdivisés en catégories suivant une structure arborescente, accompagnée d'une brève description (BOURAMOUL, 2011). Bien que ce procédé fût pionnier en la matière, il tend à disparaître. En effet, le fait qu'on devrait sélectionner les catégories dans lesquelles on recherche supposé que l'on sache où chercher exactement. Néanmoins, on doit reconnaître aux annuaires un grand avantage, celui de mettre en quelque sorte dans le contexte, ainsi les recherches dans la base de données sont diminuées, en plus d'obtenir des résultats plus pertinents

Les annuaires sont donc des outils basés sur le recensement humain de l'information. Ils signalent des sites et des ressources de l'Internet comme un catalogue de bibliothèque signale des livres ou bien encore comme les pages jaunes signalent des entreprises. On distingue dans ce contexte deux catégories d'annuaires : les annuaires commerciaux, qui se financent grâce à la publicité, et les annuaires non commerciaux, qui sont élaborés par des individus de façon bénévole ou bien par des institutions (Ils sont soit généraux soit spécialisés) et leur préoccupation consiste toujours à identifier les ressources et les sites en tenant comptes de leurs qualités.

Enfin, la distinction est importante entre les annuaires et les moteurs de recherche. Le tableau I.1 présente une comparaison entre ces deux outils pour la RI (BOURAMOUL, 2011).

Annuaires	Moteurs
Indexation de sites par des documentalistes	Indexation de mots par des robots
Recherche sur des sites et sur des catégories	Recherche en texte intégral sur des pages web
Avantages : choix des informations, classement raisonné par catégories et sous-catégories	Avantages : plus d'exhaustivité et mise à jour plus rapide
Inconvénients : moins d'exhaustivité mise à jour moins rapide	Inconvénients : capture de pages web sans classement raisonné
À retenir : L'exploration des catégories s'avère souvent plus fructueuse que celle des sites	À retenir : La recherche par mots-clés donne de meilleurs résultats sur les moteurs

Tableau I.1. *Comparaison entre les annuaires et les moteurs de recherche.*

I.5.1.3. Les métas-moteurs

Ils sont de création plus récente. Ils constituent en fait la première génération des agents dits "intelligents". Ils permettent d'interroger en une seule fois différents outils de recherche, qu'ils soient de type annuaire ou de type moteur, afin de fournir une réponse plus exhaustive.

Les méta-moteurs existent en deux catégories (Largouet Anita, 2005): ceux en ligne et ceux « Desktop », consistant en un logiciel client à installer sur son ordinateur (le plus connu est COPERNIC). Le principe de fonctionnement des méta-moteurs également représente un critère pour les classifier. Certains indexent l'information contenue dans différents annuaires et moteurs, d'autres les interrogent simultanément de façon dynamique. Certains de ces méta-moteurs retraitent plus au moins les réponses (tri, dé-doublonnage). Ils permettent ainsi de rechercher de façon plus large sur le Web, ce qui peut également générer du "bruit" (des réponses non-pertinentes).

I.5.2. Algorithmes de classification de résultats

Différentes études ont suggéré de tenir compte de la popularité des documents afin d'améliorer les performances de la recherche d'information. Le PageRank (Brin, 1998) de Google et le HITS (M.Kleinberg, 1999) de Kleinberg sont deux algorithmes fondamentaux qui utilisent les liens hypertextes pour classer les résultats d'une requête

I.5.2.1. Hyperlink-Induced Topic Search (HITS)

Kleinberg fut un des premiers à s'intéresser aux propriétés de connectivité du graphe représentatif d'Internet et de son apport dans la détection de la pertinence d'une page à une requête (M.Kleinberg, 1999). Quelques constatations simples sont à l'origine de ses travaux dans ce domaine.

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

bonnes autorités et les bonnes autorités comme des pages pointées par beaucoup de bons hubs (F.Picarougne, 2004)

En pratique, le résultat de HITS est composé de deux listes ordonnées : une liste de bonnes pages autorités et une autre de bonnes pages pivots qui seront renvoyées à l'utilisateur. L'utilisateur a l'embarras du choix entre les deux listes et il peut être intéressé par une liste au détriment de l'autre selon la recherche demandée (I.Chibane, 2008).

I.5.2.2. PageRank

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

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

La mesure de PageRank (PR) proposée par (Brin, 1998) est une distribution de probabilité sur les pages. Elle mesure en effet la probabilité PR, pour un utilisateur navigant au hasard, d'atteindre une page donnée. Elle reposer un concept très simple : un lien émis par une page A vers une page B est assimilé à un vote de A pour B. Plus une page reçoit de votes, plus cette page est considérée comme importante. Le PageRank se calcule de la façon suivante :

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

$$PR(A) = (1-d) + d \frac{PR(T)}{S(T)} \quad \text{(I.19)}$$

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

I.5.3 Défis de la RI sur le Web

La RI sur le web utilise des techniques de la RI classique, mais avec de nouveaux défis, Parmi les Plus Importent (Bellaour, 2009):

- Le Web est Dynamique

L'aspect statique caractérise la RI classique. La dynamicité du web peut être résumée dans les points suivants :

- *La création* : de nouvelles pages sont créées d'une façon continue.

- *Les mises à jour :* les mises à jour font référence au changement du contenu de pages.
- *La suppression :* une page web devient inexistante si elle est supprimée, ou si tous les liens menant à cette page sont supprimés.

- Le Web est auto organisé

Contrairement aux collections de documents classiques, qui sont collectées et classifiées par des spécialistes, sur le web, n'importe quelle personne peut poster une page web et faire des liens à volonté.

- Les hyperliens

Un hyperlien ou lien hypertexte ou simplement un lien est une référence ou un élément de navigation qui permet d'associer plusieurs nœuds entre eux dans un hypertexte.

- La duplication du contenu

La duplication du contenu est la reproduction d'un contenu d'une page sur une autre, les pages concernées sont celles présentant des URL différentes et un contenu similaire ou trop approchant

I.6. Evaluation des SRI

Il existe plusieurs modèles de recherche d'information qui utilisent différentes techniques et méthodes. Il est difficile de connaitre quelle est la meilleure technique pour une telle collection de documents. Pour ce faire, un axe de recherche indépendant, est créé et qui s'intéresse à l'évaluation des systèmes de recherche d'information et comparer entre les résultats de chaque modèle (D. Manning & Prabhakar , 2009) (Hlaoua L. , 2007).

L'évaluation des SRI est donc un thème de recherche important en sciences de l'information. La validation d'un modèle, d'une technique ou d'un algorithme en RI est une fonction fondée sur le paradigme de l'évaluation comparative des résultats qu'ils produisent dans des situations de recherche simulées à l'aide de collections de test (Ressad-Bouidghaghen, 2011), Ces dernières fournissent des ressources de base pour l'évaluation des SRI . Afin de comparer et d'évaluer les systèmes de RI, des mesures d'évaluation basées essentiellement sur le rappel et la précision sont utilisées.

I.6.1. Notion de pertinence

La pertinence représente la question fondamentale posée lors de l'accès à l'information. Elle dépend notamment du centre d'intérêt ou du domaine d'application, mais également du moment, du lieu et du support que l'utilisateur a choisi pour accéder à l'information, et enfin du système qui délivre cette information (W.N.Zemirli, 2008).

I.6.1.1. Pertinence système

Les SRI doivent s'appuyer sur un modèle de pertinence qui leur permet de calculer pour chaque document un score de pertinence. La pertinence apparaît donc ici comme une valeur numérique calculée par les SRI. Cette pertinence système a cependant des limites car elle est estimée à partir d'un score de ressemblance entre la requête et les documents, et détermine une pertinence supposée des documents pour l'utilisateur (BOURAMOUL, 2011).

I.6.1.2. Pertinence utilisateur

La pertinence utilisateur est une notion subjective permet à ce dernier d'exprimer sa satisfaction par rapport aux documents que le système lui restitue. En effet, deux utilisateurs différents ayant soumis la même requête au SRI ne jugent pas de la même manière les réponses du système. Dans le cas où le jugement de pertinence est donné par un degré de pertinence des documents, le désaccord entre plusieurs utilisateurs est dû au fait que les besoins sont différents et que le même besoin peut être exprimé différemment en fonction de l'utilisateur. De plus, l'interprétation que l'utilisateur fait des documents qu'il reçoit dépend en partie de ses connaissances personnelles et de son expérience, ainsi que du contexte dans lequel s'effectue sa recherche (Kompaoré, 2008).

Outre ce classement basé sur l'entité qui produit le jugement, nous trouvons dans les travaux une distinction de quatre types de pertinence :

- Pertinence algorithmique : la pertinence est traduite par une mesure algorithmique dépendant des caractéristiques des requêtes d'une part et des documents d'autre part. C'est le seul type de pertinence qui est indépendant du contexte ;
- Pertinence thématique : elle traduit le degré d'adéquation de l'information retrouvée au thème et non au contenu de la requête, C'est le type de pertinence adressé par les assesseurs (experts) de la campagne d'évaluation TREC (Text Retrieval Conférence) ;
- Pertinence cognitive : il s'agit de la pertinence liée au thème de la requête, pondérée par la perception ou les connaissances de l'utilisateur sur ce même thème ;
- Pertinence situationnelle : c'est la pertinence liée à la tâche de recherche. Ce type de pertinence traduit essentiellement l'utilité de l'information relativement au but de recherche de l'utilisateur.

I.6.2. Collections de test

L'utilisation des collections de test remonte aux années 1970, quand un protocole complet d'évaluation des systèmes de recherche d'information a été élaboré, dans le cadre du projet Smart de Gerald Salton (Gerard, 1968) (G.Salton & M.McGill, 1983), Un groupe d'experts d'un domaine, a travaillé à la constitution d'un corpus documentaire homogène et cohérent couvrant le Domain. Une fois la collection documentaire constituée, ces experts définissent un ensemble de questions types (besoin en information) dont les caractéristiques divers reflètent l'ensemble des demandes pouvant être adressées au système (spécialisées, généralistes et abordant différents sujets thématiques) (Laine-Cruzel, 2001).

Pour chacune des questions-test, les experts effectuent alors la tâche la plus complexe qui consiste à construire collaborative ment la « réponse idéale », susceptible d'être retournée par le système. La collection de test est finalement constituée lorsque l'on dispose, pour une collection de document donnée, d'un ensemble de question-test avec leur « réponse idéale » associées.

La collection de tests permettant d'évaluer la qualité de la réponse fournie par des SRI a évalué. Il suffit de lui soumettre les question-test, et de comparer la réponse qu'il fournit aux réponses idéales de la collection.

De nombreux projets bases sur des corpus d'évaluation se multiplient depuis les années 70. Nous citons à titre d'exemple la collection CACM, CISI, TIME, CLEF et le projet d'évaluation le plus ambitieux est sans aucune doute : TREC.

TREC (*Text Retrieval conference*) est une compagne d'évaluation organisée par le NIST (Nationale Institute of standards and Technology) et la DARPA (Defines Advanced Research Projects Agency). TREC est un programme international initie au début des années 90 dans le but d'améliorer la recherche documentaire basée sur de grandes collections d'essai. Il offre des moyens homogènes d'évaluation des systèmes de recherche d'information. En effet, il a permis de définir plusieurs taches en recherche d'information et de construire de larges collections de test.

I.6.3. Campagnes d'évaluation

L'objectifs des compagnes d'évaluation est de fournir aux chercheurs en RI un cadre unifié (collections de test, protocole d'évaluation et mesures d'évaluation) qui permettre d'évaluer les performances de leurs systèmes de recherche, et de juger leurs qualités (ALI KHOUDJA, 2015).

Parmi l'importante compagne d'évaluation nous citons : le programme américain TREC5 (*Text Retrieval Evaluation Conference)* qui en est à sa 14 émet campagne, le programme CLEF (*Cross Language Evaluation Forum*) qui en est à sa sixième, le programme NTCIR7 conçu pour les langues asiatique, le programme Amaryllis …etc.

Ces principaux programmes sont tous inspirés Dun même protocole qui est celui de TREC. Ils fournissent l'infrastructure nécessaire pour l'évaluation des méthodologies de recherche et de filtrage d'information.

I.6.4. Protocole d'évaluation d'un SRI

Un protocole d'évaluation a pour objectif de décrire la stratégie a appliqué pour évaluer les performances d'un système de recherche. Le protocole d'évaluation le plus connu et le plus utilisé en RI est le protocole TREC disposent d'une collection de test sur laquelle ils testent leur système. Les 10000 premiers documents pertinents restitués par un système participant, pour chaque requête de la collection de test, sont examinés. Puis, différentes métriques sont utilisées pour calculer, à partir de ces résultats de recherche, différentes mesures de performance du système (AZZOUG, 2013).

I.6.5. Mesures d'évaluation

Etant donnée une requête Q, les documents de la collection sont partitionnés selon le rapport à la requête en deux catégories : documents pertinents et documents non pertinents. Les documents sélectionnés par le SRI en réponse à la requête Q peuvent être pertinents ou non (AZZOUG, 2013).

Dans ce qui suit, on notera :

- C : l'ensemble de tous les documents de la collection ;
- PS : l'ensemble de documents pertinents sélectionnés ;
- P : l'ensemble de documents pertinents ;
- S : l'ensemble de documents sélectionnés.

La figure I.5 résume cette distribution dans la collection. Plusieurs métriques d'évaluation des SRI sont proposées selon cette distribution, nous résumons les principales dans ce qui suit.

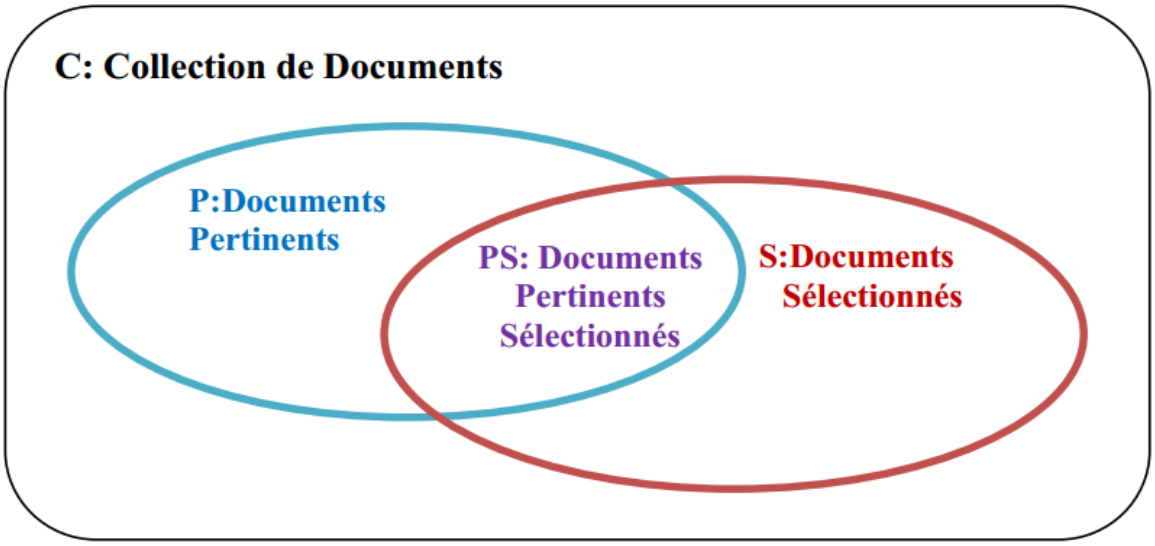


Figure I.6 : *Distributions des documents dans une collection suite à une requête*

I.6.5.1. Les mesures Rappel/Précision

La comparaison des réponses d'un système pour une requête avec les répondes idéales nous permet d'évaluer les deux métriques suivantes (Goker & Davies, 2009), (Riejsbergen, 1979) (D. Manning & Prabhakar , 2009) :

Taux de rappel : le rappel mesure la proportion des documents pertinents sélectionnés par le système relativement à l'ensemble des documents pertinents contenus dans la base documentaire. Il mesure la capacité de système à sélectionner tous les documents pertinents répondant à une requête.

Taux de précision : la précision mesure la proportion de document pertinent sélectionné par le système relativement à l'ensemble des documents sélectionnés par le système. Elle mesure la capacité du système à sélectionner que le document pertinent à une requête, i.e., sa capacité à rejeter tous les documents non pertinents. Les taux de rappel et de précision sont donnes par les formulations suivantes :

$$rappel = \frac{nombre\ de\ documents\ pertinents\ sélectionnés}{nombre\ totale\ de\ documents\ pertinents} = \frac{PS}{P} \qquad (I.20)$$

$$précision = \frac{nombre\ de\ documents\ pertinents\ sélectionnés}{nombre\ totale\ de\ documents\ \ selectionnés} = \frac{PS}{S} \qquad (I.21)$$

Les mesures complémentaires du rappel et de la précision sont respectivement : le silence et le bruit.

On parle de silence lorsque les documents pertinents ne sont pas sélectionnés par le SRI, alors qu'ils existent dans la collection de documents. Le silence est défini par la formule suivante :

$$silence = 1 - rappel = \frac{P - PS}{P} \quad \text{(I.22)}$$

On parle de bruit lorsque des documents non pertinents sont proposés à l'utilisateur par le SRI. Le bruit est calculé par la formule suivante :

$$bruit = 1 - précision = \frac{S - PS}{S} \quad \text{(I.22)}$$

I.6.5.2. La courbe de Rappel-Précision

Pour pouvoir examiner les résultats efficacement, on calcule la paire des mesures (taux de rappel, taux précision) à chaque document restitué pour les premiers documents dans la liste des réponses du système à la requête. Nous dessinions ensuite la courbe Rappel-Précision correspondante

L'idéal serait d'avoir une précision et un rappel égaux à 1, signifiant que tous les documents pertinents ont été retrouvés et qu'aucun document non pertinent n'a été sélectionné. En pratique, ceci n'est jamais atteint. En effet, comme indiqué dans la figure I.7, ces deux mesures évoluent en sens inverse. Intuitivement, si nous augmentons le rappel pour retrouve plus de documents pertinents, la précision du système diminuera engendrant ainsi trop de documents non pertinents sélectionnés. Inversement, une plus grande précision risque de rejeter des documents pertinents diminuant ainsi le rappel.

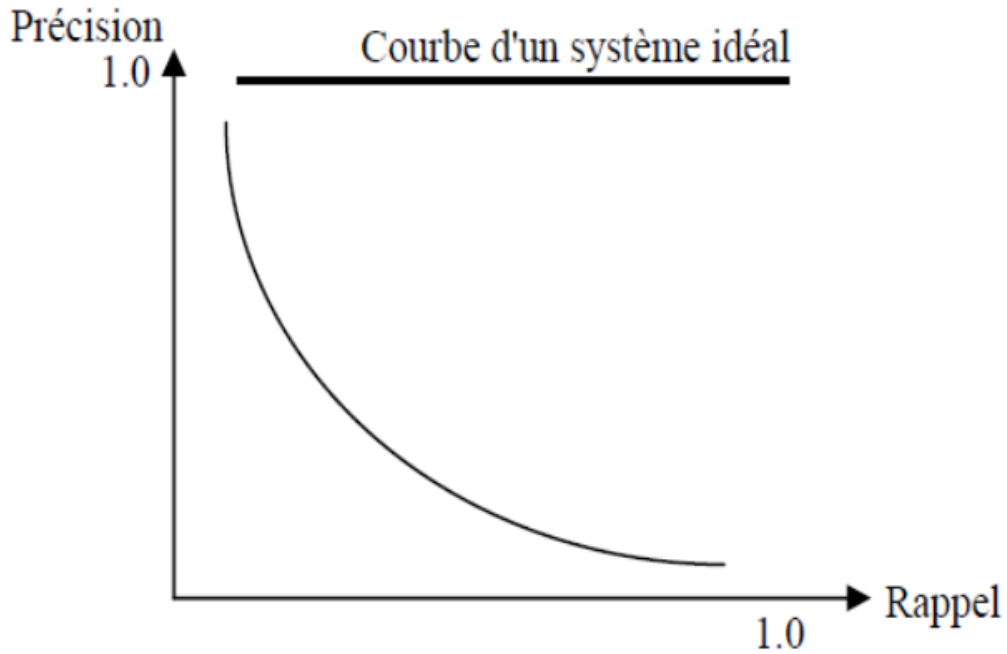


Figure I.7 : *Forme général de la courbe Précision-Rappel*

I.6.5.3. Précision à x documents

L'efficacité d'un system de recherche d'information peut être mesurée par le nombre de documents pertinents sélectionnés dans les premiers rangs (x=5, 10, 15,20, …etc.) qui est note P@x. La précision à x documents mesure la proportion des documents pertinents sélectionnés parmi les x premiers documents restitués par le SRI

$$P@x = \frac{PS}{x} \tag{I.23}$$

I.6.5.4. R-Précision (précision exacte)

La R-Précision est intéressante lorsque la collection comporte un nombre considérable de document pertinents. Elle mesure la proportion des documents pertinents sélectionnés parmi les R premiers documents retournés par le SRI

$$R - Précision = P@R = \frac{PS}{R} = \frac{PS}{P} \tag{I.24}$$

I.6.5.5. Précision moyenne

C'est une moyenne de précision sur un ensemble de points de rappel. Nous distinguons les deux mesures suivantes :

- *AP : Average Precision* : elle représente la précision moyenne de la requête *q*. Elle est définir par la moyenne des précisions (non interpolées) calculées à chaque document pertinent restitué en réponse à chaque requête donnée *q*. La précision moyenne peur être calculée pour une requête donnée *q* par la formule :

$$AP_q = \frac{1}{PS} \sum_{j=1}^{k} P(j) \times A(j) \tag{I.25}$$

Ou :

A(j) = 1 si le j^{éme} document restitué est pertinent, *A(j)*=0 si le j^{éme} document restitué est non pertinent

P(j) est la précision à *j* documents restitués.

PS le nombre de documents restitués.

K le nombre total de documents retournés par le système de recherche.

- *MAP : Mean Avrage Precision* : elle présente la précision moyenne de l'ensemble de requêtes. Elle est définie par la moyenne arithmétique des précisions moyennes AP_q de chaque requête *q*. Elle permet de mesurer les performances du SRI utilisé. Cette mesure est formellement donnée par :

$$MAP = \frac{1}{N} \sum_{q=1}^{N} AP_q \tag{I.26}$$

Ou :

N est les nombre de requetés.

AP_q est la précision moyenne de la requête *q*.

I.6.5.6. Interpolation de la précision

On s'intéresse à interpoler la précision pour chaque point r_j de rappel, Pour les valeurs de rappel r_j, on retint dans la littérature les 11 points standard de 0.0 à 1.0 :

Rj= {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}

La précision interpolée au point de rappel r_j est égale à la valeur maximal des précisions obtenues aux points de rappel r, tel que r>= r_j, Elle est ainsi calculée comme suit :

$$P(r_j) = max_{r \geq r_j} P(r) \tag{I.27}$$

I.6.5.7. F-mesure

La F-mesure prend en considération la précision et le rappel simultanément. Elle est définie comme la combinaison pondérée du taux de rappel et du taux de précision par :

$$F(j) = \frac{2 \times P(j) \times R(j)}{P(j) \times R(j)} \tag{I.28}$$

Où : P(j) et R(j) représentent respectivement les résultats de précision et de rappel.

I.6.5.8. Mesure Harmonique

La mesure harmonique est une fonction de paires de valeur de précision et de rappel, qui tient ses valeurs dans un intervalle fermé [0,1]. Elle retourne une valeur égale à 0 lorsqu'aucun document pertinent n'est sélectionné, et à 1 lorsque tous les documents sélectionnés sont pertinents

$$H(j) = \frac{2}{\frac{1}{R(j)} + \frac{1}{P(j)}} \tag{I.29}$$

La valeur de H est élevée quand les valeurs de rappel et de précision sont élevées, Ainsi, La mesure Hamonique garantie le compromis entre les deux mesures de rappel et de précision.

I.6.5.9. E-mesure

Une mesure permettant de combiner le rappel et la précision. Elle a pour but de permette à l'utilisateur de spécifier laquelle des valeurs de précision ou de rappel, est plus intéressante.

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{R(j)} + \frac{1}{P(j)}} \tag{I.30}$$

Où b est un paramètre de l'utilisateur lui permettant de spécifier l'importance du rappel ou de la précision {Si b est supérieurs à 1, plus d'importance à la précision, et inversement}.

I.6.5.10. Autres mesure de performance

Il existe aussi d'autres mesure de performance des SRI telles que :

- *Le temps de réponse acceptable ;*
- *La présentation des résultats claire avec facilité d'utilisation* ;
- *Le nombre total de documents pertinents retournés ;*
- *Le range du premier document pertinent ;*
- *La longueur de recherche.*

I.7. Recherche d'information arabe

I.7.1. La langue arabe

La langue arabe est la langue officielle, d'enseignement et de communication, de près de 22 pays avec plus de 450 millions de locuteurs (BENZATER, 2015). Le développement de la langue arabe a été associé à la naissance et la diffusion de l'islam. L'arabe s'est imposée, depuis l'époque arabo-musulmane, comme langue religieuse mais plus encore comme langue de l'administration, de la culture et de la pensée, des dictionnaires, des traités des sciences et des techniques.

L'arabe peut être considéré comme un terme générique rassemblant plusieurs variétés (Boulaknadel, 2008) :

- l'arabe classique : la langue du Coran, parlée au VIIe siècle
- l'arabe standard moderne (l'ASM) : une forme un peu différenciée de l'arabe classique, et qui constitue la langue écrite de tous les pays arabophones. L'ASM reste le langage de la presse, de la littérature et de la correspondance formelle, alors que l'arabe classique appartient au domaine religieux et est pratiqué par les membres du clergé
- les dialectes arabes : malgré l'existence d'une langue commune, chaque pays a développé son propre dialecte. Issus de l'arabe classique, leurs systèmes grammaticaux respectifs affichent de nettes divergences avec celui de l'ASM. On peut regrouper ces dialectes en quatre grands groupes :
  - les dialectes arabes, parlés dans la Péninsule Arabique.
  - les dialectes maghrébins
  - les dialectes proche-orientaux : égyptien, soudanais, syro-libano-palestinien, irakien
  - la langue maltaise est également considérée comme un dialecte arabe.

I.7.2. Particularité de la langue arabe

L'arabe est une langue sémitique qui diffère de langues indoeuropéennes syntaxiquement, morphologiquement et sémantiquement. Sur le plan de l'écrit, on trouve dans la langue arabe des voyelles diacritiques qui sont écrites au-dessus ou au-dessous une consonne pour lui donner le son et le sens désiré. Mais les arabophones utilisent rarement ces voyelles dans leur écriture, puisque pour eux le contexte suffit à leur faire comprendre le sens exact du mot. Cela peut cependant aboutir à de nombreuses ambiguïtés lors du traitement automatique de la langue. (Abdulhay, 2006)

En fait, dans la langue arabe on utilise différents types de voyelles qui déterminent la prononciation des lettres :

I.7.2.1. Voyelles brèves :

Ces voyelles sont notées avec des signes diacritiques qui permettent de savoir comment une lettre est prononcée dans un mot donné. L'écriture de ces voyelles est facultative et concerne peu de textes (Coran et textes didactiques), ce qui cause beaucoup d'ambiguïté au niveau de la prononciation et la compréhension de textes courants.

Ces voyelles ne s'écrivent pas dans le corps du mot, mais ils sont ajoutés sur ou sous la consonne à laquelle ils se réfèrent. La graisse FATHA et le KASRA sont représentés par un tiret mis respectivement au-dessus et au-dessous de la consonne à laquelle ils se réfèrent. Le DAMMA ressemble à un petit WAW et il est écrit au-dessus de sa consonne.

Les voyelles courtes en arabe sont :

Signe	Translittération	Nom arabe	Nom de voyelle
◌َ	A	فَتْحة	FATHA
◌ِ	I	كَسْرة	KASRA
◌ُ	U	ضَمَة	DAMMA

Tableau I.2 *: Voyelles courtes arabes*

I.7.2.2. Voyelles longues

Les voyelles longues sont la version allongée des courtes. Pour représenter une voyelle longue, la langue arabe adopte une méthode particulière la lettre d'allongement: on ajoute une lettre juste après la consonne et sa voyelle courte. Chaque voyelle courte à son signe d'allongement propre.

Voyelle courte	Exemple de sons courts	Prononciation selon API	Elongation de signe	Résultat	Exemple	Prononciation selon API	Sens
◌َ	بـ	/ba/	ا	بَا	بَاب	/bāb/	Port
◌ِ	بـ	/bi/	ي	بِي	كَبِير	/kabīr/	Grand
◌ُ	سـ	/su/	و	سُو	سُوق	/sūq/	marché

Tableau I.3 : *Voyelles longues arabes*

I.7.2.3. TANWIN ((ALFATHA, ALKASRA, ALDAMMA)

C'est le doublement de voyelle courte : TANWIN ALFATHA et ALDAMMA se trouvent sur la dernière lettre d'un mot mais TANWIN ALKASRA se trouve sous la dernière lettre. Ces types de TANWIN sont utilisés pour indiquer que la prononciation de la dernière lettre est le son de cette lettre influencée par une des voyelles courtes, une FATHA , DAMMA ou KASRA , selon le type de TANWIN trouvé, et la deuxième FATHA , DAMMA ou KASRA désigne le son.

Voici les différents types de TANWIN :

- TANWIN ALFATHA : (◌ً)
- TANWIN ALKASRA :(◌ٍ)
- TANWIN ALDAMMA :(◌ٌ)

SHADDA (◌ّ) : Elle désigne un doublement de la lettre sur laquelle elle se trouve.

SUKUN (ْ) : Il indique que la prononciation de la lettre où se trouve le SUKUN devrait être claire et aigu.

Par ailleurs, la langue arabe n'utilise pas de lettres majuscules (pour les noms propres ou au début d'une phrase) ce qui rend la tâche de désambigüisation plus difficile.

Le mot au sens graphique (suite de graphèmes entre deux blancs) se compose souvent en arabe de plusieurs formes collées : on parle alors d'agglutination. La structure du mot arabe est décomposable en cinq éléments : proclitique, préfixe, base, suffixe et enclitique. Cette caractéristique peut être source de difficultés pour l'analyse morphosyntaxique du texte arabe, car la distinction entre les formes agglutinées et les lettres radicales d'un mot est très difficile à faire.

En ce qui concerne *la morphologie*, l'arabe se compose essentiellement de trois sous-ensembles : le nom, le verbe, les particules qui se subdivisent elles-mêmes en différentes sous-catégories : préposition, conjonction, pronom, article, interjection et adverbe.

I.7.3. Ressources linguistiques

Les ressources linguistiques (RL) jouent un rôle essentiel dans les applications de la technologie des langues. Ainsi, d'une part les RL alimentent les différents processus des systèmes de TAL, d'autre part, elles sont de plus en plus exploitées pour accompagner le travail de modélisation linguistique par des méthodes statistiques (L. ROMARY, 2000).

Nous donnons un aperçu des ressources linguistiques existantes pour l'arabe. Nous nous limitons à celles utiles pour l'analyse automatique des corpus textuels. Nous abordons successivement donc les lexiques, les corpus de textes monolingues bruts et annotés.

I.7.3.1. Lexiques

Un lexique se constitue d'une liste d'entrées lexicales auxquelles peuvent être associées des informations linguistiques relevant la morphologie, la syntaxe, ou la sémantique ainsi que sa fréquence d'usage, des exemples d'emploi, etc.

Nous présentons quelques-uns des lexiques électroniques de l'arabe, en abordant d'une part les lexiques monolingues et d'autre part les lexiques multilingues.

I.7.3.1. Lexique monolingue :

- DIINAR1 : DIctionnaire INformatisé de l'ARabe Version 1, est l'une des bases lexicales les plus importantes dans le domaine du TALN traitant la langue arabe. Elle a été conçue et réalisée à l'ENSSIB par (M.GHENIMA, 1998), (D.HARMAN, 1987). Elle comprend un nombre total de 119 693 lemmes, entièrement voyelles, répartis comme suit : 29 534 noms, 19 457 verbes, 70 702 déverbaux

I.7.3.2. Lexiques multilingues

- DIINAR-MBC : (DIctionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus) est l'extension multilingue de DIINAR1 vers anglais et français.
- DixAF : Le DixAF (Dictionnaire bilingue français arabe, arabe français), copropriété CNRS/ENS lettres et sciences humaines développé par Fathi Debili, est constitué de près de 125 000 liens binaires établis entre 43 800 entrées françaises environ et près de 35 000 entrées arabes.

- AGROVOC : AGROVOC est un vocabulaire plurilingue de spécialité (regroupant environ 36 000 termes uniques) conçu pour couvrir l'ensemble des domaines ayant trait à l'agriculture, aux forêts, à la pêche, à l'alimentation et à d'autres domaines apparentés (comme l'environnement).AGROVOC en arabe contient actuellement plus de 24 699 descripteurs et plus de 1247 non descripteurs (synonymes). Chaque descripteur à son équivalent dans d'autres langues.
- UNBIS : La base multilingue UNBIS est élaborée, au sein du Département de l'information, par la Bibliothèque Dag Hammarskjöld. Il contient la terminologie utile pour l'analyse thématique des documents et autres publications relatifs aux programmes et activités de l'Organisation des Nations Unies. UNBIS en arabe contient actuellement 7002 descripteurs et 2183 non-descripteurs (synonymes). Chaque descripteur à son équivalent dans d'autres langues.

I.7.3.2. Corpus

Le corpus est défini par (E. LAPORTE, 2000) comme « l'ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

Parmi la ressource de TREC, une collection de large échelle, avec un volume de 884 MOctes, est disponible pour les utilisateurs y compris les arabophones. Ce corpus est constitué de 383 883 documents arabes, Il a été encode en utilisant le SGML et a été transcodé a Unicode (UTF-8). Il inclut des articles journalistiques provenant d'Arabic Newswire de l'AFP (Agence France Presse) de 13 mai 1994 au 20 décembre 2000 avec approximativement 76 millions d'unités lexicales (DELEKH, 2011)**.**

I.7.4. Outils de traitement automatique de la langue arabe

Les outils de traitement automatique de la langue arabe sont l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication. Notre objectif dans cette section est de recenser les principaux outils de TAL en langue arabe (Boulaknadel, 2008).

I.7.4.1. Analyseurs morphologiques

L'utilisation des analyseurs morphologiques est nécessaire pour segmenter les unités lexicales, repérer les différents composants et attester leur appartenance à la langue (ALI KHOUDJA, 2015). Nous citons :

- *Buckwalter*

L'analyseur de buckwalter développé par LDC (Linguistic Data Consortium) permet de segmenter chaque unité lexicale en une séquence du type préfixe-stem-suffixe. Le préfixe est une combinaison de 0-4 caractères, le suffixe est composé de 0 à 6 caractères et le stem comprend un à plusieurs caractères. Il est constitué principalement de trois lexiques : préfixes (548 entrées), suffixes (906 entrées), et stem (78839 entrées). Les lexiques sont complétés par trois tables de compatibilité utilisés pour couvrir toutes les possibilités de combinaisons préfixe-stem (2435 entrées), suffixe-stem (1612 entrées) et préfixe-suffixe (1138 entrées). Ainsi, l'analyseur donne en sortie l'unité lexicale, sa catégorie morphosyntaxique et sa traduction anglaise.

- *Aramorph*

L'analyseur morphologique Aramorph (R.ABBES, 2004)segmente les unités lexicales, repère les différents composants et atteste son appartenance à la langue. Pour cela, le système est assisté par le lexique DINAAR.1 pour éviter les analyses théoriquement possibles et inexistantes dans la langue. Par la suite, l'analyseur donne une liste des traits associés à l'unité lexicale en entrée. Il offre deux types d'options. Le premier vise les traits morphosyntaxiques, le second concerne l'analyse des préfixes et suffixes. En plus des étiquettes morphosyntaxiques, il donne en sortie d'autres informations comme la base, l'unité lexicale minimale vocalisé ou non ainsi que la forme complété supposée vocalisée ou non. Analyser les préfixes revient à décrire ses découpages possibles et d'examiner les compositions des clitiques. Ceci amène le système à faire la distinction entre les clitiques ayant la même forme mais appartenant à des catégories syntaxiques différentes.

- *Xerox*

L'analyseur morphologique de Xerox (R.BENVENISTE, 2001) est basé sur l'approche de transducteur à états finis. La segmentation de la phrase en unités lexicales est réalisée par un transducteur à états finis. Ce transducteur découpe la chaîne d'entrée en une séquence d'unités lexicales qui peuvent correspondre à une forme fléchie, une marque de ponctuation, etc. La deuxième étape est l'analyse morphologique des unités lexicales produites par la segmentation de la phrase. Cette étape est aussi réalisée par un transducteur qui relie la forme fléchie à la forme lexicale (et vice-versa). La forme lexicale est une séquence comprenant la représentation canonique de l'unité lexicale (le lemme), un ensemble d'étiquettes représentant le comportement morphologique de l'unité lexicale, et sa catégorie syntaxique.

- *ASVM*

L'analyseur de Diab (ASVM) est un logiciel libre, développé en Perl par l'équipe de Mona Diab à la Leland Stanford Junior University en 2004. Il s'agit d'une adaptation à l'arabe du système anglais YamCha basé sur les Support Vector Machines. Les données probabilistes ont été acquises pendant une phase d'entrainement sur le corpus annoté Arabic TreeBank. Ci-dessous la description des fichiers d'entrée et de sortie de l'analyseur.

ENTREE

Le texte à analyser doit être encodé en Buckwalter, qui est une table de correspondance biunivoque entre les caractères arabes et l'ASCII. Voici un exemple :

«ولم يحتسب الحكم المجري ساندور بول ركلة جزاء صحيحة اثر عرقلة داخل المنطقة من قبل اليساندرو»

« wlm yHtsb AlHkm Almjry sAndwr bwl rklp jzA' SHyHp Avr Erqlp dAxl AlmnTqp mn qbl AlysAndrw. »

SORTIE

Dans le fichier de sortie, chaque unité lexicale étant suivie d'un slash et de sa catégorie. Les clitiques s'écrivant attachés à leur hôte comme les conjonctions de coordination (fa-) et (wa-), la préposition (bi-) etc. sont étiquetés indépendamment.

« w/CC lm/RP yHtsb/VBP Al/DT Hkm/NN Al/DT mjry/JJ sAndwr/NNP bwl/NNP rklp/NN jzA'/NN SHyHp/JJ Avr/IN Erqlp/NN dAxl/IN Al/DT mnTqp/NN mn/IN qbl/NN Al/DT ysAndrw/NNP ./PUNC »

I.7.4.2 Racineurs

Le Racineur consiste à révéler la racine d'une unité lexicale, Les algorithmes de racinisation : (S.Leah, Ballesteros, E.Margaret, & E.Connell, 2005) et Khoja (S.Khoja, 1999)

- *Racineur de larkey*

Une analyse morphologique assouplie. Elle consiste à essayer de déceler les préfixes et les suffixes ajoutés à l'unité lexicale : par exemple le duel (ان) dans (معلمَان, deux professeurs), le pluriel des noms masculins (ون, ين) dans (معلمون, des professeurs) et féminins (ات) dans (مسلمات, musulmanes), la forme possessive (كم,هم,نَا) dans (كتَابهم, ses livres) et les préfixes dans les articles définis (وَال,بال,كال,فال,ال).

- *Racineur de Khoja*

Le racineur de Shereen khoja développé au sein de l'université de Lancaster, a été utilisé dans le cadre d'un système de recherche d'information développé à l'Université du Massachusetts. L'approche de Khoja (S.Khoja, 1999) consiste à détecter la racine d'une unité lexicale, d'une part, il faut connaître le schème par lequel elle a été dérivée et supprimer les éléments flexionnels (préfixes et suffixes) qui ont été ajoutés, d'autre part comparer la racine extraite avec une liste des racines préalablement conçue.

I.7.4.3 Lemmatiseurs

Le traitement morphologique, articulément la lemmatisation, devient un des traitements les plus importants dans le domaine de la recherche d'information, notamment pour la langue arabe qui renferme un certain nombre de complexités, et qui vise à déterminer une forme appropriée d'index aux mots, Ainsi, l'utilisation des lemmatiseurs lors du traitement des documents arabes dérivent une nécessité.

I.7.5.Propriétés morphologiques et difficultés de l'arabe

Les recherches sur le traitement automatique de l'arabe ont débuté vers les années 1970, et les premiers travaux concernaient notamment les lexiques et la morphologie. Les statistiques prouvent que depuis 1995 quand le premier journal arabe fut lancé en ligne sur www.asharqalawsat.com, le nombre de sites Web arabes s'était développé exponentiellement. Et dès l'année 2000, plus de 20 mille sites arabes sont introduits sur le Web, dont plus de 7% édités (Abdelali, 2004).

I.7.5.1. L'alphabet arabe

L'alphabet de la langue arabe compte 28 consonnes appelées « Huruf al_Hija » (tableau III.1). Au contraire de nombreuses autres langues, l'arabe s'écrit et se lit de droite à gauche. Des petits points noirs sont utilisés comme marques de différenciation entre des lettres qui partagent une forme identique, et ils sont placés au-dessus ou au-dessous de la lettre en un, deux ou trois.

Lettre	Prononciation	Lettre	Prononciation	Lettre	Prononciation	Lettre	Prononciation
أ	Alif	د	Dal	ض	Dad	ك	Kaf
ب	Ba'	ذ	Thal	ط	Tha	ل	Lam
ت	Ta'	ر	Ra	ظ	Zha	م	Mim
ث	Tha'	ز	Zayn	ع	Ayn	ن	Nun
ج	Jim'	س	Sin	غ	Ghayn	ه	Ha

ح	Hha'	ش	Chin	ف	Fa	و	Waw
خ	Kha'	ص	Sad	ق	Qaf	ي	Ya

Tableau III.1 : *Les 28 Lettres de L'Arabe*

I.7.5.2. Morphologie arabe

Selon la théorie grammaticale arabe ancienne, le lexique de la langue arabe comprend trois catégories de mots : les *verbes*, les *noms* et les *particules*, qui se subdivisent elles-mêmes en différentes sous catégories : préposition, conjonction, pronom, article, interjection et adverbe (EL-Dahadeh, 1996). La morphologie arabe se caractérise par le phénomène de la possibilité de générer un ensemble de mots, référant à une même représentation sémantique, à partir d'une seule unité linguistique appelée 'racine', à l'aide de différents patrons ou modèles appelés 'schèmes'. La plupart des racines se composent de 3 consonnes, et rarement de 4 ou de 5 consonnes.

De ces racines, nous pouvons produire des dérivés nominaux et verbaux par l'application des schèmes (patrons). On peut produire jusqu'à 30 mots d'une racine de 3 consonnes ; le tableau III. 2montre un exemple de certains schèmes appliqués au mot « كتب» (tuer) à partir duquel nous pouvons produire plusieurs mots :

Schéma	كتب	QTL	Sens
فَاعِل	كاتِب	Kateb	écrivain
فَعَلَ	كَتَب	Kataba	A écrit
مَفْعَل	مَكتب	Maktab	bureau
فُعِلَ	كُتِبَ	Keteba	A été écrit
…			

Tableau III.2 : *Exemple de modèles pour le mot « كتب»*

I.7.5.3.Structure d'un mot

La définition du mot du point de vue du traitement automatique se heurte à des considérations syntaxiques et sémantiques. Les mots sont le résultat d'une transformation appelée segmentation qui produit des unités appelés segments. Dans l'arabe, ces séquences de lettres sont appelées les mots graphiques. Les mots peuvent avoir une structure composée.

Les mots sont séparés par des espaces et d'autres signes de ponctuation. Plusieurs types d'affixes sont agglutinés au début et à la fin des mots masquant les limites entre eux : antéfixes, préfixes, suffixes et post fixes. La représentation suivante schématise une structure possible d'un mot complexe. Notons bien que la lecture se fait de droite à gauche.

Post fixe	Suffixe	Noyau	Préfixe	Antéfixe

Tableau III.3 : *Structure d'un mot Arabe*

- Les antéfixes sont des prépositions ou des conjonctions
- Les préfixes et suffixes expriment des traits grammaticaux, tels que les fonctions de noms, le mode du verbe, le nombre, le genre, la personne….
- Les post fixes sont des pronoms personnels.
- Le noyau représente la racine du mot.

I.7.5.4. Les catégories des mots

En ce qui concerne la langue arabe, on distingue trois catégories principales de mots :

- Le verbe : entité qui exprime un sens dépendant du temps.
- Le nom : l'élément désignant un être ou un objet qui exprime un sens indépendant du temps.
- Les particules : entités qui servent à situer les événements et les objets par rapport au temps.

III.2.2.4.1. Le Verbe

Les verbes arabes peuvent être classés selon le nombre ou la nature des consonnes de leurs racines, ou bien selon leurs schèmes (EL-Dahadeh, 1996). Si nous les classifions selon le premier critère, nous aurons soit des verbes trilitères qui ont trois consonnes, soit des verbes quadrilatères, peu nombreux, qui ont quatre consonnes.

Selon les premiers et troisièmes critères, nous avons soit des verbes nus « مجَّرد» qui sont composés seulement par les consonnes de leurs racines et des voyelles courtes, soit des verbes augmentés ou dérivés « مزيد»qui est dérivés de trois consonnes de la racine par modification des voyelles, par redoublement de la deuxième lettre de la racine, par adjonction et même par intercalation d'affixes.

III.2.2.4.2. Le Nom

Les noms arabes, regroupant les substantifs, les adjectifs et les pronoms, ainsi que d'autres noms invariables (R.Blachére & Domombynes, 1975), peuvent être divisés en deux classes : la première contient ceux qui sont dérivés à partir d'une racine, elle regroupe les noms conjugables ou semi-conjugables qui peuvent avoir la forme duelle, plurielle, etc. La deuxième classe regroupe les noms non-conjugables qui gardent la même forme quel que soit le contexte, comme les noms étrangers et certains noms fixes.

La morphologie des noms arabes dépend de diverses fonctions :

- Agent : celui qui fait l'action.
- Objet : celui qui subit l'action.
- Instrument : signifiant l'instrument de l'action.
- Lieu : qui désigne en général un endroit.
- Nom d'action : désigne l'action
- Etc.

III.2.2.4.3. Les Particules

Les particules sont des lemmes invariables et en nombre limité. Ils indiquent l'articulation de la phrase et ils servent à préciser les modalités des prépositions verbales et nominales (EL-Dahadeh, 1996), (R.Blachére & Domombynes, 1975). Pour notre traitement, ces particules font partie du dictionnaire qui regroupe les mots vides. Elles sont classées selon leur champ sémantique et leur fonction dans la phrase ; on en distingue plusieurs types :

- *Préposition :* ب, ك, ل, عَنْ, حَتَّى
- *Particules de coordination :* وَ, فَ, ثُمَّ, أو
- *Particules interrogatives :* أ, هَل, مَا
- *Particules d'affirmation :* نَعَم, هَل, مَا
- *Particules de négation :* لا لنْ, لمْ
- *Particules distinctive :* أيّ
- *Particules relatives :* ما
- *Particules de futur :* سَ, سَوفَ

- *Particules conditionnelles :* إنْ, لَو

Certaines particules peuvent également porter des affixes, ce qui complique leur identification. Les affixes peuvent être classés par catégorie selon leur rôle syntaxique.

I.7.6. Problèmes de traitement automatique de l'arabe

En Arabe écrit, les voyelles (signes diacritiques) sont omises et le résultat de cette omission est que les mots tendent à avoir un haut niveau d'ambiguïté. Cette ambiguïté constitue un problème crucial dans la recherche d'information, en fait un mot arabe peut avoir plusieurs significations. Par exemple, le mot « على» sans voyelles peut signifier le nom propre (Ali) ou la préposition (à). Un autre exemple, une ambiguïté vient du mot « العلم» qui signifie la science ou drapeau alors que voyelle on aura «العلم» pour la science et «العَلَمْ» pour le drapeau.

En plus du phénomène d'ambiguïté, il y a un autre problème de la forme plurielle des noms irréguliers, également appelée le pluriel cassé. Dans ce cas-ci, un nom en pluriel prend une autre forme morphologique différente de sa forme initiale dans le singulier. Par exemple, le mot « امرأة»en singulier prend la forme « نسوة» en pluriel. L'absence d'un dictionnaire pour ces noms irréguliers rend difficile de concevoir un algorithme basé sur les règles pour transformer ce genre de pluriel à la forme singulière.

De plus, les mots sont séparés par des espaces et d'autres signes de ponctuation. Néanmoins, des prépositions sont agglutinées au mot, faisant des limites invisibles entre le mot et la préposition. Par exemple, «أعلمكم».

Ce bref tour d'horizon de la morphologie arabe montre la richesse, la complexité des phénomènes mis en jeu, et les différences de ses propres caractéristiques par rapport aux autres langues au niveau lexical, syntaxique, sémantique et pragmatique. Les problèmes dus au système dérivationnel très riche que possède la langue arabe, font qu'elle reste une langue particulièrement difficile à traiter.(Abdulhay, 2006)

I.8 Conclusion

Dans ce chapitre, nous avons présenté les notions de base de la RI classique. En particulier, nous avons exposé les différentes étapes d'un processus de recherche d'information, et explicité les bases de l'indexation, la reformulation de requête, ainsi que, l'appariement et les modèles de base de la recherche d'information, on passe par les approches d'évaluation des SRI et les outils de la RI sur le web, Et nous avons terminé par présenter le RI arabe par sa particularité, ces ressources linguistique et ces outils de TALN.

La RI classique est basée sur l'indexation par des mots-clés et l'appariement lexical. Ces techniques présentent des insuffisances du fait de l'utilisation d'entités lexicales, les mots-clés, qui sont par nature ambigües. Pour pallier ces insuffisances, plusieurs approches d'indexation ont été proposés dans l'objectif principal est de représenter les documents et les requêtes par des concepts (sens des mots), qui sont par nature non ambigus. C'est l'objet de l'indexation sémantique que nous présentons dans le chapitre suivant, vis-à-vis aux déférent notion et méthodes de l'apprentissage non supervisée.

CHAPITRE II :

INDEXATION SÉMANTIQUE & CLUSTERING

Plan de Travail

II.1. Introduction 37
II.2. Notations et définitions préliminaires 37
II.3. La Recherche d'Information basée sur une ressource externe 38
II.3.1. Les ressources externes 38
II.3.2. Utilisation des ressources externes dans la RI 39
II.4. Indexation Sémantique et conceptuelle 39
II.4.1. Indexation sémantique 39
II.4.2. Indexation conceptuelle 43
II.4.3. Similarité sémantique 44
II.5. Apprentissage Automatique 47
II.5.1. Apprentissage supervisé 47
II.5.2 Apprentissage non supervisé 47
II.6 Conclusion 58

II.1. Introduction

L'indexation est une étape primordiale dans tout processus de recherche d'information. Sa qualité dépend de sa capacité à mieux représenter l'information portée par le contenu d'un corpus textuel. Dans les systèmes de recherche d'information classiques ; l'indexation d'un document, est réalisé par l'ensemble des mots-clés »s qu'il contient. Cette représentation, appelé aussi indexation classique, ne prend pas en considération la sémantique des mots pour décrive avec précision la thématique abordée dans le document, ce qui implique des résultats non pertinents lors de la recherche. C'est ainsi que les travaux récents en recherche d'information proposent l'indexation sémantique, en se basant sur la représentation des documents par les sens des mots (ou concepts), plutôt que par les mots eux-mêmes.

Dans le contexte d'indexation sémantique le manque des ressources électroniques arabe pos un problème lors de l'utilisation de ces méthodes et approches, qui impose l'obligation de passer vers un autre domaine qui est l'apprentissage automatique (DataMining), qui offre des méthodes non supervisée applicable pour la recherche d'information et l'extraction des relations entre ces différents composants (termes, documents, …)

Dans ce chapitre, nous présentons les notions fondamentales de la RI sémantique, et après cette introduction, nous définissons les différents concepts de la recherche d'information basés sur les ressources externes, Nous décrivons ensuite les approches les plus connues d'indexation sémantique et conceptuelle et en passant par les mesures de similarité sémantique. Finalement on introduite les notions de l'apprentissage automatique et on passe par la suite sur l'apprentissage non supervisé (Clustering) par la discussion de son processus et ses algorithmes

II.2. Notations et définitions préliminaires

- Mots clés

En recherche d'information, les mots clés sont les mots qui décrivent mieux le contenu d'un document ou d'un corpus. Les mots clés sont souvent des noms, des verbes ou des adjectifs, par opposition aux mots outils comme les prépositions, les déterminants ou les prénoms (BOUCHAM, 2009).

- Termes

Le terme est défini de la maniérée suivante : Elément d'une terminologie. Combinaison indissociable d'un concept et d'une dénomination (désignation).

Le Terme est donc un élément construit, dont le statut est différent des autre mots de la langue car il répond à un besoin de normalisation sémantique.

- Unités de sens : Concepts

Un concept est la représentation mentale d'un semble d'objets différents, mais considérés comme équivalents d'un certain point de vue (nom identique, action commune, etc.). Les concepts ne se trouent pas directement dans les textes. En effet comme le constate très justement C.Roche «Il n'y a pas de concepts dans un texte, mais uniquement des traces linguistiques de leur usages».

- Similarité

Gilles Bisson définit la similarité comme étant l'opérateur qui permet d'évaluer les ressemblances et les dissemblances qui existent au sein d'un ensemble de données (Gilles , 2000). Il subdivise la similarité en deux grandes familles, les similarités numériques qui quantifient les ressemblances sous la forme d'une valeur dans l'intervalle [0,1] et les similarités symboliques qui permettent de caractériser les ressemblances

- Distance

Les notions de similarité et de distance sont équivalentes et sont utilisées indifféremment pour calculer le degré de similarité entre deux objets, et donc de deux cas. En effet, pour deux cas donnés, plus sont similaires moins ils sont distants. L'évaluation de la similarité entre deux cas, décrits par un ensemble d'attributs, est souvent évaluée par la mesure de la distance entre ces deux cas à travers leurs ensembles d'attributs. Si deux individus sont séparés par une faible distance, alors ils se ressemblent.

II.3. La Recherche d'Information basée sur une ressource externe

Grâce à une grande disponibilité des ressources externes (générales ou spécifiques), les systèmes à base de connaissances externes se développent. Dans le domaine de la RI, les ressources externes, i.e. les ressources hors de la collection des documents et de la requête, jouent un rôle important. Ces ressources offrent des connaissances sur la corrélation entre des mots et des termes, donc une meilleure compréhension du texte. La compréhension du texte évolue alors du niveau linguistique vers le niveau sémantique. De ce fait, dans la RI, la correspondance entre un besoin d'information et des documents peut évoluer du niveau de correspondance entre sacs de mots vers la correspondance sémantique (Thi Hoang, 2010).

II.3.1. Les ressources externes

Il existe une grande variété de ressources externes. Certains types de ressources sont populaires, nous reprenons leurs définitions issues du grand dictionnaire dans le domaine des technologies de l'information (Thi Hoang, 2010).

Les ressources externes exploitées en WSD (Word Sense Disambiguation) sont classées en deux catégories (R.Navigli, 2009) : ressources structurées et ressources non structurées**.**

II.3.1.1. Les ressources structurées

- Les dictionnaires informatisés : qui représentaient des sources de connaissances très populaires dans les années 80 pour les différentes disciplines du domaine du traitement automatique de la langue. Dans un dictionnaire informatisé, un mot de la langue possède un ou plusieurs sens qui sont définis par leurs glossaires. Le glossaire d'un sens décrit le sens du mot par une définition, des commentaires et/ou des exemples d'utilisation courante.
- Vocabulaire contrôlé : « Dans un domaine préalablement défini (d'ordre scientifique, technique, professionnel ou autre, et en général pour une langue donnée), le choix de termes sélectionnés, classés et indexés en vue de faciliter l'indexation, le stockage et la recherche des publications traitant des concepts apparentés à ces termes. »
- Taxonomie : « Construction d'un plan de classification de concepts utilisant des classes disjointes de concepts agrégés. »

- Thésaurus : « Vocabulaire contrôlé et dynamique de termes ayant entre eux des relations sémantiques et génériques, et qui s'applique à un domaine particulier de la connaissance. »
- Ontologie : « Ensemble d'informations dans lequel sont définis les concepts utilisés dans un langage donné et qui décrit les relations logiques qu'ils entretiennent entre eux ».

Le vocabulaire contrôlé ne contient que des termes d'indexation pour faciliter l'accès et l'indexation des documents, alors que la taxonomie classifie, organise les termes et possiblement ajoute des relations hiérarchiques entre groupes de termes. Le thésaurus possède en plus des relations sémantiques entre termes du type de la causalité, l'association,... L'ontologie, quant à elle, est plutôt la représentation formelle des informations qui sont définies sous la forme des concepts avec leurs relations.

II.3.1.1. Les ressources non structurées

- Les corpus d'apprentissage : Ce sont de longs textes utilisés dans les techniques d'apprentissage pour construire la connaissance nécessaire pour la désambiguïsation des sens des mots (WSD). Ces corpus peuvent être étiquetés manuellement avec les sens des mots.
- Les corpus de collocations : Ce sont des ensembles de collocations de mots qui ont une tendance de se produire ensemble régulièrement.

II.3.2. Utilisation des ressources externes dans la RI

L'utilisation de ressources externes aux documents dans la RI a déjà été largement explorée. Les applications principales concernent :

- L'extraction de concepts et l'indexation conceptuelle : il s'agit d'identifier les concepts correspondant aux termes et de les utiliser dans l'indexation.
- L'expansion de requêtes et de documents consiste à ajouter les concepts sémantiquement liés aux concepts des requêtes ou des documents.
- Le calcul de similarité sémantique entre concepts via les informations sur les concepts et leurs relations dans les ressources externes.

II.4. Indexation Sémantique et conceptuelle

L'indexation est une étape primordiale dans tout processus de recherche d'information. Sa qualité dépend de sa capacité à mieux représenter l'information portée par le contenu d'un corpus textuel. Dans les systèmes de recherche d'information classique, l'indexation d'un document, est réalisée par l'ensemble des mots-clés qu'il contient. Cette représentation, appelée aussi indexation classique, qui ne prend pas en considération la sémantique des mots pour décrire avec précision la thématique abordée dans le document

II.4.1. Indexation sémantique

L'indexation sémantique en RI est née du problème de l'ambiguïté des mots de la langue naturelle utilisés classiquement pour représenter les documents et requêtes. L'indexation sémantique a pour objet de représenter les documents et requêtes par les sens des mots permettant ainsi de lever toute ambiguïté, ce qui a pour conséquence d'améliorer les résultats de la recherche (AZZOUG, 2013).

Pour retrouver les sens corrects des mots dans un document (ou dans un requête donnée), l'indexation sémantique requiert des techniques de désambiguïsation des sens des mots. Ces techniques de basent sur l'utilisation de ressources, telles que les corpus d'apprentissage ou les ressources terminologiques.

Nous décrivons, dans ce qui suit, les travaux les plus significatifs dans le domaine. Les approches d'indexation qui sont basées sur les techniques de désambiguïsation linguistique des mots seront aussi présentées.

II.4.1.1. Les approches de désambiguïsation des sens des mots (WSD)

L'objectif principal de la WSD est de retrouver les sens corrects des mots dans leur contexte d'utilisation. De nombreuses approches de désambiguïsation sémantique des mots existent. Ces approches peuvent être divisées en : approches basées sur les ressources linguistiques externes et approches basées sur les corpus d'apprentissage.

Les approches basées sur les ressources linguistiques utilisent les ressources terminologiques structurées comme sources d'évidence pour la définition des sens des mots. Les approches de désambiguïsation basées sur les corpus d'apprentissage s'appuient plutôt sur de gros textes pour construire la connaissance nécessaire pour cela.

II.4.1.1.1. Les approches basées sur les ressources linguistiques

Ces approches se basent sur les dictionnaires informatisés, les thésaurus ou les ontologies pour désambiguïser un mot ambigu.

- Les approches basées sur les dictionnaires informatisés

Lesk (M.E.Lesk, 1986)a développé l'un des premiers systèmes de désambiguïsation basé sur un MRD. Dans ce système, les sens possibles d'un mot cible sont d'abord identifiés à partir d'un dictionnaire. Un score est ensuite associé à chacun de ces sens. Le sens qui maximise ce score est retenu comme sens correct du mot cible. Formellement :

$$Score_{Lesk}(S_i) = \arg\max \sum_{i \neq j} \left| gloss(S_i) \cap gloss(S_j) \right| \qquad \text{(II.1)}$$

- gloss(S_i) est l'ensemble des mots significatifs non vides, appartenant au glossaire du sens S_i associé au mot cible m_i.

- gloss(S_j) est l'ensemble des mots significatifs du glossaire de S_j associé à un mot m_j appartenant au contexte de m_i.

Cowie et al. (J.Cowie, J.Guthrie, & L.Guthrie, 1992) Proposent d'enrichir chaque définition d'un sens d'un mot ambigu par son code de domaine (ou catégorie de sujet), qui lui est attribué dans le LDOCE. Ce code est traité dans cette méthode comme un mot faisant partie de la définition. La désambiguïsation d'un mot repose alors sur le nombre de mots et de codes de domaine en commun entre les définitions des sens de ce mot et celles des sens des autres mots appartenant à son contexte. Le sens ayant le plus grand nombre est retenu comme sens de ce mot dans son contexte d'apparition.

Bien que les dictionnaires informatisés aient été exploités largement dans la désambiguïsation des mots, leur qualité est souvent remise en question. En effet, ces

dictionnaires, développés pour usage humain, ne donnent pas une description suffisante des liens sémantiques entre mots qui peuvent être utilisés dans un cadre automatique.

Ils présentent en outre, une divergence au niveau de la représentation des sens lexicaux et des relations entretenues entre eux (M.Apidianaki, 2008).

- Les approches basées sur un thésaurus

Ces approches reposent sur le vocabulaire contrôlé d'un thésaurus comme source de connaissances pour déterminer les sens exacts des mots dans leur contexte d'apparition. Le vocabulaire d'un thésaurus fournit une description sémantique des associations entre mots et classent les sens des mots liés sémantiquement dans des catégories sémantiques (catégories de domaines).

Yarowsky (D.Yarowsky, 1992)se base sur les catégories sémantiques[18] du thésaurus Roget (Roget, 1995), pour retrouver les sens adéquats des différents mots qui apparaissent dans l'encyclopédie « Grolier multimédia » (Grolier, 1990). Il propose une approche de désambiguïsation à deux étapes : la première consiste à retrouver parmi les catégories sémantiques du Roget, celle qui correspond au mot à désambiguïser dans son contexte d'utilisation. La seconde permet alors d'assigner le sens correct du mot ambigu dans la catégorie ainsi identifiée.

Dans une approche différente, Mohammad et Hirst (S.Mohammed & G.Hirst, 2006)utilisent le thésaurus pour la désambiguïsation des sens des mots. L'idée de base de leur approche est que la majorité des occurrences d'un mot dans un corpus textuel, portent un même sens qui représente le sens prédominant du mot. Ce sens désigne son sens approprié dans le corpus.

Quoique l'utilisation des thésaurus dans les travaux de désambiguïsation des mots ait permis de réduire le problème de l'ambiguïté de la langue, en offrant une catégorisation sémantique des mots représentant leurs degrés d'associations, ils restent cependant une ressource de connaissances peu exploitée par les chercheurs à cause de leur formalisation qui s'appuie sur la notion de termes plutôt que celle de concepts (R.Mizoguchi, 2004).

- Les approches basées sur une ontologie

Ces approches se basent sur les concepts des mots dans l'ontologie ainsi que leurs relations sémantiques conceptuelles, telles que l'équivalence, la synonymie, la relation d'hyponymie/ hyperonymie, et … autres, pour définir le sens correspondant du mot dans son contexte d'utilisation. Parmi les ressources ontologiques les plus exploitées dans la désambiguïsation des mots de la langue on retrouve WordNet

Sussna (M.Sunssa, 1993) défini une approche de désambiguïsation basée sur les liens sémantiques entre synsets de noms dans l'ontologie WordNet. Il propose dans un premier temps d'assigner un poids à chaque relation sémantique entre deux synsets dans WordNet, exprimant la proximité sémantique entre eux. Il associe le poids le plus fort à la synonymie et le poids le plus faible à la relation d'antonymie. Pour désambiguïser un mot (nom) dans son contexte d'utilisation, cette approche affecte à chacun de ses synsets dans l'ontologie WordNet un score égal à la somme de ses distances sémantiques minimales avec les synsets des mots co-occurrents de son contexte. La distance sémantique minimale entre deux synsets est calculée par la somme des poids des relations sur le chemin le plus court entre eux. Le synset qui maximise ce score est alors retenu comme sens adéquat du mot dans son contexte.

Dans une autre approche intéressante, Kolte et al. (S.G.Kolte & S.G.Bhirud, 2008); proposent une désambiguïsation à deux niveaux : d'abord désambiguïser le domaine correct d'un mot dans son contexte local (i.e. la phrase où il apparait), puis désambiguïser ce mot dans le domaine ainsi identifié. Le domaine correct d'un mot est celui qui maximise ses occurrences dans le contexte du mot. Pour identifier le sens du mot dans le domaine choisi, Kolte et al. Proposent d'exploiter les relations sémantiques entre synsets dans les taxonomies de WordNet (Hypernym, Meronymy/Holonymy,…).

II.4.1.1.2. Les approches basées sur les corpus d'apprentissage

Ces approches se basent sur l'utilisation d'un corpus composé d'un grand nombre de contextes de mots polysémiques, dans le but d'apprendre les connaissances utiles sur le sens d'usage des mots. Cette phase d'identification automatique des connaissances est appelée apprentissage. A l'issue de cette phase, l'algorithme de désambiguïsation est capable d'assigner le sens adéquat aux mots apparaissant dans une nouvelle phrase en s'appuyant sur les connaissances acquises durant la phase d'apprentissage.

Les approches de désambiguïsation basées sur les corpus d'apprentissage se distinguent en approches supervisées et approches non supervisées.

- Les approches supervisées

Weiss (F.Weiss, 1973)est l'un des premiers à s'intéresser à l'apprentissage des règles de désambiguïsation à partir d'un corpus étiqueté. En analysant les 20 phrases contenant des occurrences de cinq mots ambigus dans le corpus étiqueté ADI[22], il construit manuellement deux types de règles : des règles générales de contexte et des règles de modèle.

Une règle générale de contexte établit la relation structurelle entre une occurrence d'un mot ambigu avec les mots co-occurrents dans son contexte, permettant ainsi de définir son sens.

Une règle de modèle établit la relation de position d'un mot co-occurrent par rapport une occurrence d'un mot ambigu, afin de pouvoir définir son sens.

Dans une autre approche de désambiguïsation, Yarowsky (D.Yarowsky, 2000)repose sur des connaissances basées sur un arbre de listes de décision hiérarchiques pour identifier le sens adéquat d'une occurrence d'un mot ambigu. Chaque liste est établie à partir d'une classification de succession de règles conditionnelles ordonnées de type if….then, en observant les différentes caractéristiques des occurrences d'un mot dans une partie du corpus annoté SENSEVAL. Ces caractéristiques sont traduites par des attributs (sa localisation, son type (lemme, catégorie syntaxique,…) et la valeur du Token) de patterns permettant de déclencher les règles de décision pour obtenir le sens attendu du mot ambigu ou se pointer vers une autre liste de décision. Dans une liste, les règles sont ordonnées par le ratio de log vraisemblance.

- Les approches non supervisées

Ces approches se basent sur des corpus non annotés pour construire la connaissance nécessaire à la désambiguïsation. L'apprentissage est basé sur l'idée que les occurrences d'un mot qui ont même sens possèdent souvent des mots co-occurrents similaires. Il s'agit alors de regrouper en clusters, les mots voisins des différentes occurrences similaires d'un mot dans une partie du corpus. Les ensembles des clusters obtenus représentent ainsi les sens possibles de ce mot, qui seront utilisés pour la désambiguïsation des autres occurrences se trouvant dans l'autre partie du corpus.

Schütze (H.Schutze, 1998)propose une technique de désambiguïsation non supervisée basée sur le modèle vectoriel. Dans sa phase d'apprentissage, cette approche représente les contextes des différentes occurrences d'un mot par des vecteurs de mots co-occurrents dérivés à partir de leurs contenus. Le contexte d'une occurrence est défini par une fenêtre de cinquante mots autour de l'occurrence en question. Ces vecteurs contextes sont ensuite regroupés en clusters en fonction de leur degré de similitude calculé sur la base de la mesure de cosinus. Pour chaque cluster obtenu, Schütze calcule le centroïde (ou le barycentre) de l'ensemble de vecteurs dans ce cluster est lui affecte le sens du mot qu'il lui correspond. Par conséquent, les ensembles de clusters caractérisés par leurs centroïdes décrivent les sens possibles (ou usages possibles) du mot. La désambiguïsation d'une nouvelle occurrence de ce mot est alors basée sur la distance entre le vecteur contexte construit pour cette occurrence et le centroïde de chaque cluster associé à ce mot dans la phase d'apprentissage. Le cluster ayant son centroïde le plus proche du vecteur contexte de l'occurrence est alors considéré comme sens adéquat de cette occurrence dans son contexte d'utilisation.

Dans une autre approche, Pantel et al. (P.Pantel & D.Lin, 2002) Proposent un algorithme de clustérisassions dit Clustering by Committee. Cet algorithme consiste dans un premier temps à retrouver pour chaque occurrence m_i d'un mot m dans un corpus, ses k mots voisins les plus proches. Pour ce faire, les mots co-occurrents de m_i sont représentés par des vecteurs traits représentants leurs caractéristiques syntaxiques ou sémantiques. Par exemple, (John, SUBJ_V, found) est défini comme un vecteur trait du mot John, désignant que John est le sujet du verbe found. Un score est attribué à chacun de ces mots co-occurrents, basé sur une mesure de l'information mutuelle entre les vecteurs traits associés respectivement à l'occurrence m_i et au mot co-occurrent appartenant à son voisinage. Les k (valeur prédéfinie) premiers mots voisins ayants les scores les plus élevés sont les plus proches sémantiquement à l'occurrence m_i en question. Ces mots sont représentés dans un vecteur définissant le vecteur contexte de m_i. Les vecteurs contextes associés aux différentes occurrences d'un mot dans le corpus d'entrainement, sont ensuite regroupés en clusters (ou comités) en fonction de leur degré de similitude. Chaque comité défini un sens possible du mot cible. Pour désambiguïser une nouvelle occurrence de ce mot, l'algorithme calcule la distance entre son vecteur contexte (construit à partir des k premiers mots co-occurrents appartenant à son voisinage) et le centroïde de chaque comité qui correspond à ce mot. Le centroïde du comité le plus proche du vecteur contexte de l'occurrence est considéré comme sens approprié de l'occurrence du mot à désambiguïser.

II.4.2. Indexation conceptuelle

L'indexation conceptuelle consiste à indexer un document, non plus avec les termes des documents, mais avec les concepts d'une base de connaissances (Mark A & WJ, 1997). Dans ces approches, il faut une "liste" de concepts cibles (qui exprime le sens des termes possibles) pour pouvoir transformer le terme en concept. Les concepts sont tirés d'un vocabulaire contrôlé : les dictionnaires de synonymes, les ontologies, les thésaurus, les taxonomies, etc. Parmi les approches qui s'inscrivent dans cet état d'esprit :

II.4.2.1. La méthode de Woods

Woods (William A, 1997)propose d'indexer conceptuellement les documents en utilisant un réseau ad-hoc, comportant des liens syntaxiques, sémantiques et morphologiques. Il a évalué

son approche sur de petites collections (pages du manuel Unix) et les résultats sont de l'ordre de 60% de précision.

II.4.2.2. La méthode de Aggarwal and al

Aggarwal and al. (C.Aggarwal & S.Philip, 2001) Construisent des "chaînes de mots conceptuelles" (conceptual word-chains) à partir de la taxonomie de concepts de Yahoo! Ils utilisent cette nouvelle représentation comme alternative à la représentation par fichier inverse. Cette approche permet d'avoir une représentation condensée de documents, ce qui diminue l'effet du bruit dû à l'ambiguïté, à la redondance et à l'indépendance des termes dans les documents.

II.4.2.3. La méthode de Baziz

Dans (BAZIZ, 2005), il a été proposé de reformuler une requête d'un utilisateur en s'appuyant sur une ontologie. L'approche consiste à projeter une requête sur une ontologie, identifier les nœuds (concepts) de l'ontologie qui représentent au mieux le contenu de la requête en utilisant différentes relations sémantiques, puis de la réécrire dans le format de départ (mots simples) avant de l'envoyer au SRI.

II.4.3. Similarité sémantique

La similarité sémantique entre les concepts est une méthode de mesure de la distance sémantique entre deux concepts en fonction d'une ontologie. La similarité sémantique est utilisée pour identifier les concepts ayant des «caractéristiques» communes. Les méthodes de similarité sémantique sont intensivement utilisées dans la plupart des applications des systèmes à base de connaissances, dans la recherche sémantique d'information (identifier la meilleure combinaison de sens entre les termes de la requête), et dans la désambiguïsation du sens (ABDELLI, 2016).

II.4.3.1. Pondération des attributs

Les attributs décrivant un problème n'ont pas la même importance. Cette importance variable doit se refléter dans le calcul de la similarité entre deux cas en attribuant des poids aux attributs pour décrire leur influence respective sur le résultat de la comparaison. Les attributs qui sont considérés comme les plus importants dans la résolution du problème doivent avoir des poids plus forts.

La détermination des poids des attributs dans la mesure de similarité des cas a une influence significative dans l'efficacité et la justesse de la phase de remémoration (IDRI, 2003). Dans la plupart des cas, ces poids sont attribués, manuellement, par l'utilisateur ou l'expert ce qui ne garantit pas la remémoration des cas les plus utiles. Toutefois, plusieurs méthodes ou techniques ont été proposées telles que les méthodes statistiques, les algorithmes génétiques ou les réseaux de neurones...

Quel que soit la méthode utilisée, ces poids ne sont pas statiques et doivent être continuellement, au fur et à mesure de l'utilisation du système, ajustés jusqu'à l'obtention des valeurs optimales.

II.4.3.2. Types de mesure de similarité sémantique

Les mesures de similarité peuvent être affectées par les caractéristiques communes des concepts comparés. Les différences entre les concepts entrainent la diminution ou l'augmentation de la distance sémantique entre ces concepts.

En outre, les mesures de similarité sont liées à la taxonomie, la similarité est affectée par la position des concepts dans la taxonomie et le nombre de liens hiérarchiques. On peut regrouper les méthodes de calculs de similarité en trois catégories :

II.4.3.2.1. Approches basées sur les arcs

La mesure de similarité la plus intuitive des objets dans une ontologie est leurs distances. Evidemment, un objet X est plus similaire à un objet Y qu'un objet Z. Cette similarité est évaluée par la distance qui sépare les objets dans l'ontologie.

a. Mesure de Wu & Palmer

Etant donnée une ontologie Ω formée par un ensemble de nœuds et un nœud racine (R). Soit X et Y deux éléments de l'ontologie dont nous allons calculer la similarité. Le principe de calcul de similarité est basé sur les distances (N1 et N2) qui séparent les nœuds X et Y du nœud racine et la distance qui sépare le concept subsumant (CS) de X et de Y du nœud R

La mesure de Wu et Palmer est définie par la formule suivante :

$$SimC(X,Y) = \frac{2 * N}{N1 + N2} \qquad \text{(II.2)}$$

b. Mesure de Rada et al

Cette mesure est adoptée dans un réseau sémantique et elle est fondée sur le fait qu'on peut calculer la similarité en se basant sur les liens hiérarchiques «issa».

Pour calculer la similarité de deux concepts dans une ontologie, on doit calculer le nombre des arcs minimums qui les séparent. Cette mesure, basée sur le calcul de la distance entre les nœuds par le chemin le plus court, présente un moyen des plus évidents pour évaluer la similarité sémantique dans une ontologie hiérarchique.

II.4.3.2.2. Approches basées sur les nœuds

Ces approches adoptent une nouvelle mesure en termes de la mesure entropique de la théorie de l'information. La probabilité « P » pour l'identification de l'utilisation d'une classe ou de ses descendants dans un corpus désigne l'information de la classe.

a. Mesure de Reni

Pour évaluer la pertinence d'un objet il faut calculer le contenu informationnel. Le contenu informationnel est obtenu en calculant la fréquence de l'objet dans le corpus (WordNet). La formule proposée par Resnik est définie par :

$$Sim(X,Y) = Max\big[E\big(CS(X,Y)\big)\big] = Max[-\log\big(p\big(CS(X,Y)\big)\big] \qquad \text{(II.3)}$$

Ou CS(X,Y) représente le concept le plus spécifique (qui maximise la valeur de similarité) qui subsume (situé à un niveau hiérarchique plus élevé) les deux concepts X et Y dans l'ontologie.

b. Mesure de Hirst et ONGE

L'idée de cette mesure est que deux concepts lexicalisés sont sémantiquement étroits si leurs ensembles synonymes (synsets) de WordNet sont reliés par un chemin qui n'est pas trop long et qui ne change pas la direction trop souvent. Avec cette mesure, toutes les relations contenues dans un réseau WordNet sont prises en considération les auteurs ont classé la direction des liens en lien haut (superclasse), lien bas (sous-classe) et lien horizontal (antonyme). Le calcul de la similarité, avec cette méthode, s'effectue entre objets (mots) par le poids du plus court chemin allant d'un terme à un autre, en plus des classifications qui indiquent les changements de direction. La force du rapport est donnée par : Simht = T-PCC-K*nd (T et K constantes, PCC distance du plus court chemin, nd nombre de changement de direction)

II.4.3.2.3. Approches basées sur l'espace vectoriel

Ces approches utilisent un vecteur caractéristique, dans un espace dimensionnel, pour représenter chaque objet et calculent la similarité se basant sur la mesure de cosine ou la distance euclidienne.

a. Similarité de Jaccard

Définie par le nombre des objets communs divisé par le nombre total des objets moins le nombre d'objets communs :

$$SimJ = \frac{x \times y}{\|x\|_2^2 + \|y\|_{2-x\times y}^2} \quad \text{(II.4)}$$

tels que x et y sont des vecteurs extraits à partir des concepts X et Y

b. Similarité de Cosine

La fréquence des objets (mots). Deux objets (documents) sont similaires si leurs vecteurs sont confondus. Si deux objets ne sont pas similaires, leurs vecteurs forment un angle (X, Y) dont le cosinus représente la valeur de la similarité

$$Simc(X,Y) = \cos(X,Y) = \frac{x \times y}{\|x\|_2 \times \|y\|_2} \quad \text{(II.5)}$$

c. Similarité de Dice

définie par le nombre des objets communs multipliés par 2 sur le nombre total d'objets.

$$Simd(X,Y) = \frac{2 \times x \times y}{\|x\|_2^2 + \|y\|_2^2} \quad \text{(II.6)}$$

II.4.3.3. Les applications de la similarité

Si on peut mesurer la similarité ou la dissimilaire, alors : (BRAHIMI , 2011)

- On peut distinguer un objet d'un autre ;
- On peut grouper les objets en se basant sur la similarité (k-means) ;
- Si on peut grouper les objets dans des groupes, on peut comprendre les caractéristiques de chaque groupe ;

- On peut expliquer le comportement d'une classe ;
- On peut classer un nouvel objet ;
- On peut prédire le comportement d'un nouvel objet ;
- On peut prendre une décision basée sur la prédiction des données.

II.5. Apprentissage Automatique

La notion d'apprentissage automatique fait référence au développement, à l'analyse et à l'implémentation des méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage et de remplir des taches dont il est difficile ou impossible de remplir par des moyens algorithmiques classiques (R.Herbrich, 2001).

L'apprentissage automatique englobe tout méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complétement le modèle, Il existe deux tendances en apprentissage, celle issue de l'intelligence artificielle, et celle issue des statistiques.

Nous distinguons deux types d'apprentissage : apprentissage non supervisé et apprentissage supervisé

II.5.1. Apprentissage supervisé

Appelée aussi classement ou classification inductive Il relève d'une démarche inductive consistant à construire automatiquement un classifier qui apprend, à partir de l'exemple déjà classé (ou étiquetés), les caractéristiques et les propriétés des catégories cibles. Ce type d'apprentissage est dit supervisé par ce que la fonction de classification s'entraine sur les catégories (ou classe) ainsi que sur leurs caractéristiques.

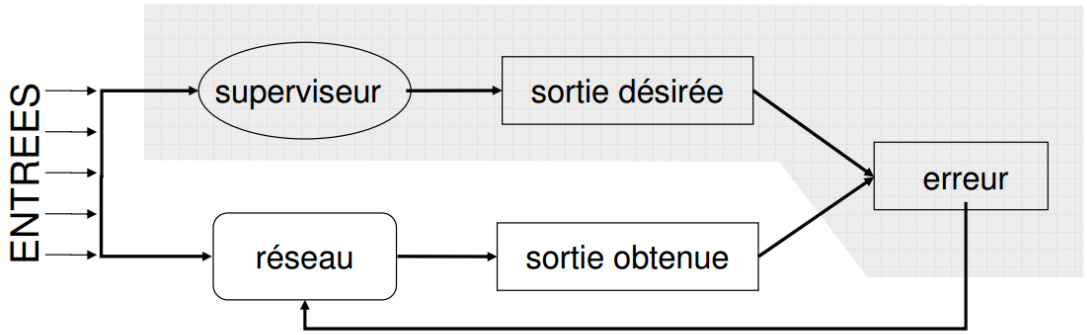


Figure II.1 *: Processus d'apprentissage supervisé*

II.5.2 Apprentissage non supervisé

Appelé aussi apprentissage par observation ou par découverte Dans ce type d'apprentissage, il n'existe pas de classes prédéfinies, le but est d'effectuer les meilleurs regroupements possibles, entre les objets dans lesquels, les observations différent très peu, ou regard de ses valeurs.

Le plus connu des problèmes non-supervisés est la classification non-supervisés ou clustering. Les classes qu'on appellera clusters, sont formées par regroupement des données qui on certaines caractéristique en commun.

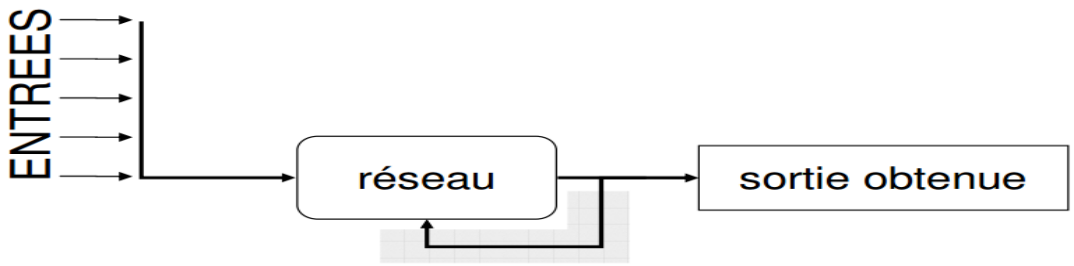


Figure II.2 : *Processus d'apprentissage non supervisé*

Pour construire un regroupement de ces données, nous avons trois choix à faire :

- Choisir une mesure de ressemblance (ou similarité) entre données.
- Choisir le type de structures que nous voulons obtenir : partition, hiérarchie, arbre…
- Choisir la méthode permettant d'obtenir la structure désirée.

II.5.2.1. Méthode de Sélection des caractéristiques en classification

La sélection de caractéristiques consiste à choisir un sous-ensemble minimum de P caractéristiques à partir d'un ensemble original en contenant Q (P < Q), de sorte que l'espace de caractéristiques soit réduit de façon optimale selon certains critères d'évaluation.

L'objectif principal est la réduction du nombre de caractéristiques utilisées tout en essayant de maintenir ou d'améliorer les performances de classification du système.

La nécessité d'avoir une bonne sélection de caractéristique est une chose indubitable, en ce sens qu'une bonne classification dépend largement des attributs utilisés pour représenter les documents. Il a été souligné dans la littérature qu'une caractéristique non pertinente n'apporte aucune information permettant de discriminer les classes entre elles. De même qu'une caractéristique bruitée porte des informations incorrectes et pouvant fausser la classification. Des problèmes de corrélations entre des caractéristiques peuvent être identifiés également si cette sélection n'a pu être faite avec une bonne stratégie. Dans ce cas-ci une même information sera redondante et cela impliquera une saisie des données sur de nombreuses caractéristiques dont le coût en termes de temps de calcul sera très élevé. Il est donc nécessaire de déterminer quelles sont les caractéristiques indispensables pour classifier les données. C'est pourquoi certains critères de sélection de caractéristique ont été mis en place par des chercheurs.

Parmi les méthodes les plus souvent utilisées citons par (Pederson & A.Weigend, 1995) (D.Lewis, 1992) (M.Moulinier & J.G.Ganascia, 1996)

- le calcul de l'information mutuelle.
- la méthode du chi-2.
- le gain d'information.

Nous décrivons brièvement ces méthodes de sélections dans les sous sections suivantes.

II.5.2.1.1. Information mutuelle

Cette méthode est basée sur le nombre de fois qu'un mot apparaît dans une certaine catégorie.

En effet plus un mot va apparaître dans une catégorie, plus l'information mutuelle entre ce mot et cette catégorie sera jugée élevée

$$IM(w, clas) = \frac{P(w, clas)}{P(w) * P(clas)} \quad \text{(II.7)}$$

- P (w, clas) : probabilité jointe de la classe clas et de la caractéristique w (la proportion des documents dans la collection qui appartiennent à la classe et qui contiennent le mot)

- P(w) : la probabilité de w (e DF document frequency relative de la caractéristique w)

- P (clas) : la probabilité de la classe (la proportion de documents de classe clas)

II.5.2.1.2. Le gain d'information

Le gain d'information est utilisé pour déterminer en quelque sorte le pouvoir discriminatoire d'un mot. Elle mesure la diminution de l'entropie, ou la quantité de connaissance gagnée si une caractéristique est présente ou absente. Mathématiquement le gain d'information (GI) pour une caractéristique w est défini par :

$$GI(w) = \sum_{X \in \{w, \overline{w}\} clas \in \{c_i\}} \sum_{clas \in \{c_i\}} P(X, clas) \log\left(\frac{P(X, clas)}{P(X) * P(clas)}\right) \quad \text{(II.8)}$$

Où :

- P (w, clas) est la proportion de documents dans la collection appartenant à la classe clas et où la caractéristique X est présente ;

- P(X) est le DF (Document-frequency) relative qui représente le nombre de document dans lequel la caractéristique X apparaît par rapport au nombre total de documents.

- P (clas) est la proportion de documents de classe clas dans le corpus.

II.5.2.1.3. Méthode du chi2 (X^2)

Il s'agit d'une mesure statistique bien connue, qui s'adapte bien à la sélection de caractéristiques. Elle évalue le manque d'indépendance entre un mot et une classe. Elle utilise les mêmes notions de cooccurrence mot/catégorie que l'information mutuelle. Une différence importante est qu'elle est soumise à une normalisation, qui rend plus comparable les termes entre eux. Elle perd quand même de la pertinence pour les termes peu fréquents

Le calcul de Chi2 nécessite de construire le tableau de contingence (2×2)

	Caractéristique w présente	Caractéristique w absent	
Classe présente	A	C	ac
Classe absente	B	D	bd
	a+b	c+d	N=a+b+c+d

Tableau II.1 : *Table de contingence de Chi2*

$$X^2(w, clas) = \frac{N(ad - cd)^2}{(a + c)(b + d)(a + b)(c + d)} \quad \text{(II.9)}$$

II.5.2.2 Applications du Clustering

Il existe de nombreuses applications possibles au Clustering, que l'on peut classer en trois groupes principaux (P.Berkhin, 2002).

1. La segmentation.
2. La classification.
3. Et l'extraction de connaissance

II.5.2.2.1. La segmentation

A pour objectif principal de réduire la taille de l'ensemble des données d'une base de données considérée. On parle dans ce cas de « Condensation » ou de « Compression des données». Elle est utilisée également pour « discrétiser » une base de données, c'est-à-dire transformer la description complexe des objets par un unique attribut caractérisant leur appartenance à une classe identifiée automatiquement. Les applications principales de telle technique concernent : l'identification des différents sous-espaces homogènes pouvant exister dans des bases de données pour aider à l'indexation efficace de telles bases, la segmentation d'images, afin d'identifier les différentes zones homogènes de l'espace décrit (les maisons, routes, fleuves, etc.).

II.5.2.2.2. La classification

Concerne l'identification des sous populations ayant des caractéristiques proches. Dans les bases de données clients, par exemple, l'objectif principal est d'établir des profils de clients aux comportements similaires, afin de comprendre leurs attentes et leurs besoins en fonction de leur profil. Les applications principales de telles techniques concernent donc la gestion de la relation client, que ce soit dans la grande distribution, les banques, les assurances, les fournisseurs d'énergie, les opérateurs téléphoniques, les transports, les médias, les navigateurs web, etc. Cela peut également être appliqué aux cas des journaux en ligne ou des forums par exemple.

II.5.2.2.3. L'extraction de connaissance

Contrairement aux deux méthodes précédentes, l'utilisation du Clustering n'a pas d'objectif préétabli. Il s'agit dans ce cas d'aider à la compréhension de la structure des données en les organisant en groupes homogènes et en faisant ainsi émerger certains sous concepts dans les données. On parle dans ce cas de génération d'hypothèses ou de modélisation prédictive, par l'inférence de règles caractérisant les données, et la suggestion de modèles sous-jacents à ces données. L'intérêt de telles techniques est par exemple avéré pour effectuer des diagnostics médicaux sur des bases de données de patients, en ciblant par exemple différents groupes de patients atteints d'une maladie donnée, reflétant ainsi différentes causes possibles pour la maladie. De même, cette technique peut être utilisée en analyse du web, en analyse des données textuelles, en analyse de ventes, en bio-informatique, en analyse génétique, etc.

II.5.2.3 Les principales étapes du Clustering

La classification, bien que primordiale dans toutes les applications citées précédemment, n'est pas la seule phase à considérer avec attention lors du développement d'un outil d'extraction et d'analyse de l'information. En effet, le processus de clustering divise en trois étapes principales : (A) La préparation des données, (B) Le choix de l'algorithme de clustering et (C) validation et interprétation des résultats (G.Cleuziou, 2004).

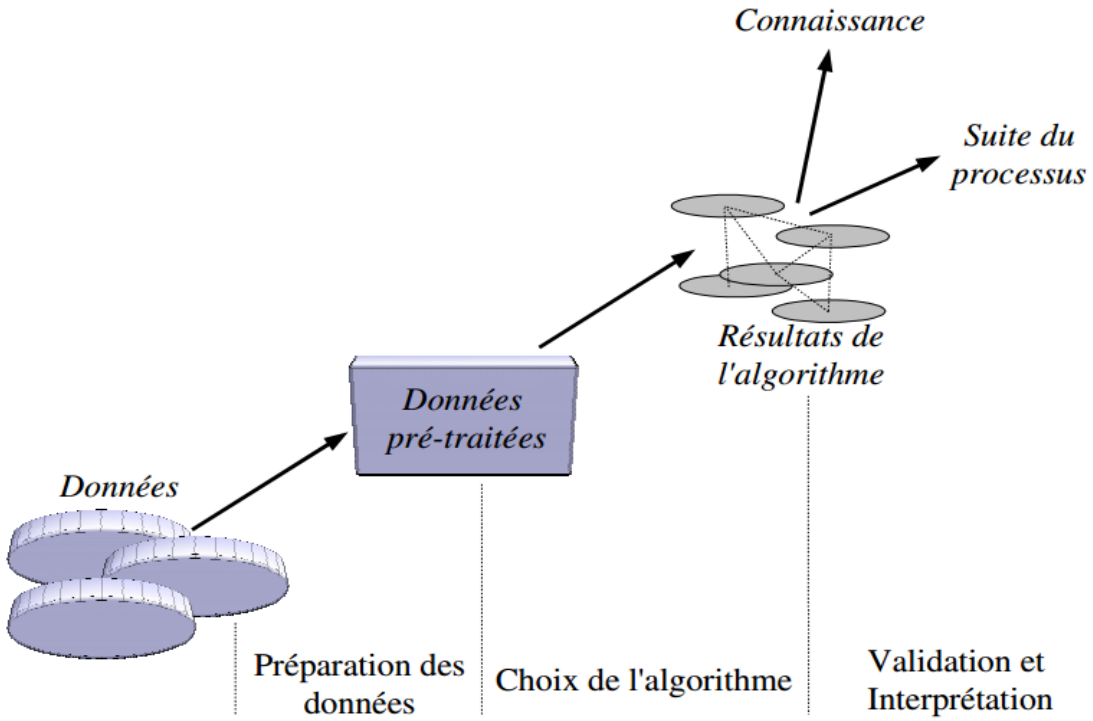


Figure II.3 *: Les différentes étapes du processus de clustering.*

II.5.2.3.1. La préparation des données

L'étape de prétraitement consiste à filtrer les informations, ne conserver que ce qui est pertinent dans le contexte de l'étude, et coder ces données pour ne retenir que les paramètres de description les plus discriminants. Ce dé bruitage est nécessaire pour pouvoir donner un sens aux résultats obtenus à l'issue de la classification des données.

- Variables et sélections

Les objets sont décrit par des variables $\{v_j\}$ j=1…p, aussi appelées attributs, descripteurs ou traits. Ces variables sont de différentes natures :

Variables *quantitatives* : continues (ex. la taille d'une personne), discrètes (ex. le nombre de personnes) ou sous forme d'intervalles (ex. la période de vie d'une personne),

Variables *structurées* : exemple la forme d'un objet (polygone, parallélogramme, rectangle, etc.)

- Distances et similarité

La plupart des algorithmes de clustering utilisent une mesure de proximité entre les objets à traiter. Cette notion proximité est formalisée à l'aide d'une mesure (ou indice) de similarité, dissimilarité ou encore par une distance. Le choix ou la construction de cette mesure est déterminant pour la suite du processus. En effet, deux mesures différentes peuvent induire deux schémas de classification différente. Finalement, chaque domaine d'application possédant ses propres données, il possédé également sa propre notion de proximité ; il faut concevoir alors une mesure différente pour chaque domaine d'application, permettant de retranscrire au mieux les différences (entre les objets) qui semblent importantes pour un problème donnée.

II.5.2.3.2. Le choix de l'algorithme

Le choix de l'algorithme de clustering doit donner lieu à une analyse globale du problème : quelle est la nature (qualitative et quantitative) des données ? Quelle est la nature des clusters attendus (nombre, forme, densité, etc.) ? L'algorithme doit alors être choisi de manière à ce que ses caractéristiques répondent convenablement à ces deux dernières questions. Les critères de décision peuvent être : les connaissances à priori, la présentation des résultats, complexité, déterministe, incrémental, any time, hard, prise en compte du contexte, tolérance au bruit, tolérance à l'effet de chaîne, tolérance aux clusters de tailles variées, tolérance aux clusters de forme quelconque, tolérance aux clusters concentriques (L.Candillier, 2006).

- La taille des données

La quantité d'objets à traiter est un premier facteur de décision. En effet pour des données de très grande taille, les algorithmes de complexité plus que linéaires (*O (a.n)*) sont quasiment prohibes. Ainsi que des méthodes telles que l'algorithme des *K-moyennes*, ou plus généralement la méthode des nuées dynamiques, étant de complexité linéaire sont très souvent utilisés.

- La nature des données

Les algorithmes de clustering s'appuient sur une matrice de similarité ou dissimilarité. Le plus souvent, cette matrice est obtenue à partir de la description des données. La nature de ces description (variables qualitatives et/ou quantitatives), détermine alors le choix de mesure de (dis) similarité utilisée.

- La forme des clusters

Certaines méthodes de clustering aboutissent à des clusters de formes spécifiques. Par exemple, les méthodes hiérarchiques le plus souvent à des clusters de forme convexe.

La question est alors de savoir s'il est raisonnable de formuler a priori, une hypothèse sur la forme des clusters.

Les variations de taille et densité ainsi que la gestion des 'outliers' sont également à prendre en compte dans le choix de l'algorithme de clustering. On entend, par variation de taille la capacité d'un algorithme à obtenir à la fois des clusters contenant beaucoup d'objets, et des clusters contenant peu voire très peu d'objets. De même, la prise en compte de la densité ou non d'obtenir des clusters contenant des objets plus ou moins 'proches', au sens de la mesure de similarité établie.

- Le type de résultats attendus

La sortie d'un algorithme de clustering peut être, par exemple, une partition, une fonction ou encore un dendrogramme (arbre hiérarchique). De même, chaque cluster obtenu peut être défini soit par l'ensemble des objets qui le composent, soit par une description relative aux variables initiales.

II.5.2.3.3. Validation et interprétation des résultats

Cette phase est dépendante de l'ensemble de la chaîne en amont. Plus les sources de données sont de bonne qualité et plus il sera aisé d'interpréter les résultats de la classification. De la même manière, ces résultats seront d'autant moins fiables ou crédibles que les informations à la base étaient bruitées ou corrompues. Deux situations sont possibles : soit la tâche de

clustering s'inscrit dans un traitement global d'apprentissage, soit les clusters générés par le clustering constituent un résultat final. Dans le premier cas, l'analyse des clusters obtenus (mesures statistiques de qualité) peut aider à orienter le traitement suivant et une description des clusters n'est pas nécessaire dans cette situation. En revanche, dans le cas où le clustering constitue à lui seul un processus global de découverte de classe, l'exploitation des clusters pour une application donnée passe par une description de ces derniers.

II.5.2.4 Les méthodes du Clustering

Dans cette partie, nous dressons un panorama des différentes méthodes de clustering rencontrées dans la littérature, nous ne prétendons pas faire une liste exhaustive de l'ensemble des notions et méthodes existant dans le cadre du clustering, mais plutôt de donner un aperçu général du fonctionnement des différentes méthodes avec chacune ses atouts et ses limites. Il est ainsi admis dans la communauté travaillant sur le clustering qu'aucun critère ni aucune méthode ne sont intrinsèquement meilleurs que d'autres sur l'ensemble des problématiques envisageables. Par contre, on considère que certaines méthodes sont plus adaptées que d'autres en fonction de l'application ciblée.

On distingue classiquement deux grandes familles de méthodes en clustering :

- Les méthodes hiérarchiques
- Les méthodes par partition

Dans le premier cas, une hiérarchie de clusters est formée de telle manière que plus on descend dans la hiérarchie, plus les clusters sont spécifiques à un certain nombre d'objets considérés comme similaires. Au contraire, dans le second cas, le résultat fourni est une partition de l'espace des objets, c'est-à-dire que chaque objet est associé à un unique cluster.

II.5.2.4.1. Le clustering hiérarchiques

Le fondement du clustering hiérarchique est de construire une hiérarchie de clusters (ou dendrogramme) tel que présenté en figure II.4 et formalisé par la définition suivante :

- La racine de l'arbre est formée par le cluster X contenant l'ensemble des objets ;
- Chaque nœud de l'arbre constitue un cluster$C_i \subset X$;
- Les feuilles de l'arbre correspondent aux singletons $\{X_i\}, \dots, \{X_N\}$;
- L'union des objets contenus dans les fils d'un nœud donné correspondent aux objets présents dans ce nœud ;
- Les « Paliers » sont indicés relativement à l'ordre de construction.

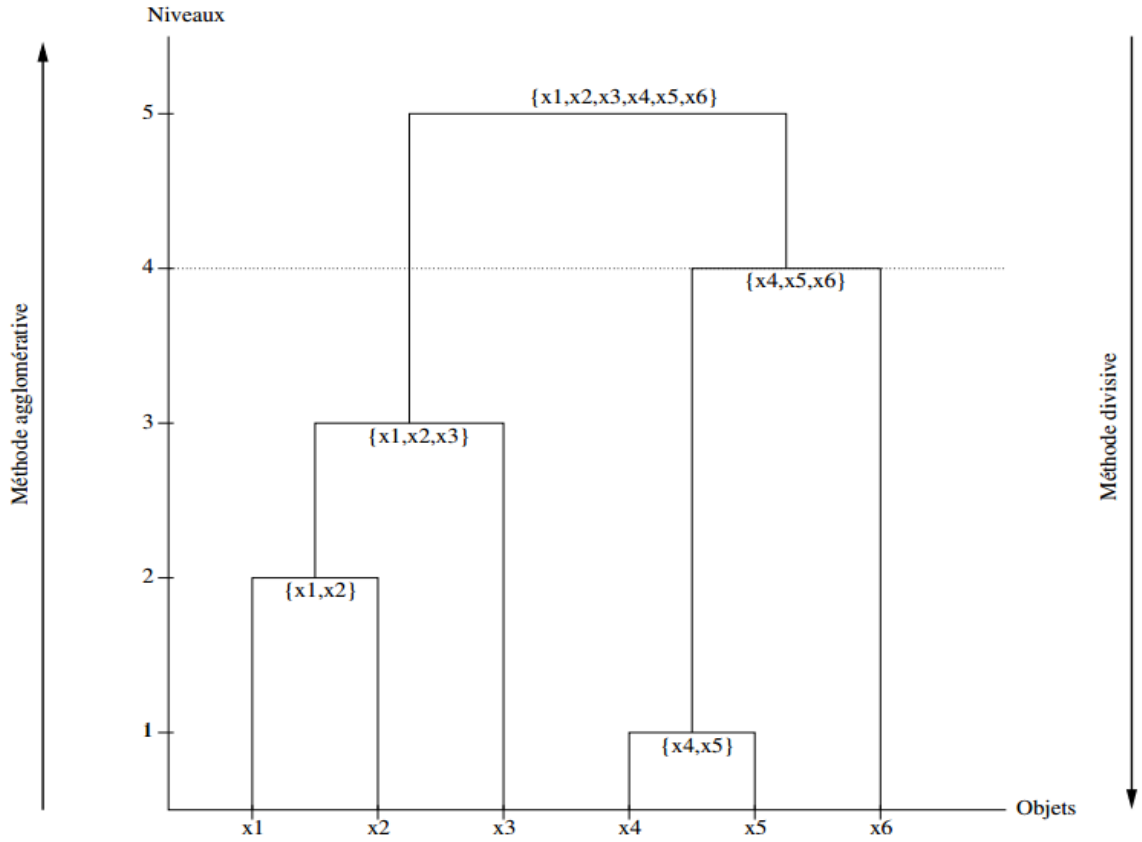


Figure II.4 *: Exemple d'une hiérarchie de clusters (dendrogramme)*

Afin de parvenir à un tel arbre hiérarchique, il existe deux méthodes principales :

a. La méthode ascendante (agglomerative ou Bottom up) :

Construit l'arbre du « bas » vers le « haut » en démarrant avec autant de clusters que d'objets initiaux dans la base, puis fusionnant successivement les clusters considérés comme les plus similaires, jusqu'à obtenir un unique cluster « racine », contenant l'ensemble des objets.

b. La méthode descendante (divisive, Top down) :

Construit l'arbre du « haut » vers le « bas » en démarrant avec un unique cluster contenant l'ensemble des objets de la base, puis divisant successivement les clusters de manière à ce que les clusters résultants soient les plus différents possibles, et ce jusqu'à obtenir des singletons (autant de clusters que d'objets dans la base). (Khatir, 2012)

Ensuite, une fois la hiérarchie formée, une étape optionnelle peut être ajoutée pour affiner le résultat. Il s'agit de déterminer le niveau L de coupure le plus approprié à appliquer dans l'arbre pour un regroupement des données aussi pertinent que possible. Par exemple, le choix du niveau de coupure (L=4) dans le dendrogramme de la figure 6 renvoie le partitionnement suivant :

$C = \{\{X_1, X_2, X_3\}, \{X_4, X_5, X_6\}\}$. Ce dernier paramètre « L » peut être choisi relativement au nombre de clusters désirés ou à l'aide d'une analyse statistique de la qualité des différentes partitions que l'on peut extraire de l'arbre.

Les critères d'agrégation

La classification hiérarchique initialise un système de classe comme un ensemble de singletons (dans le cas agglomerative) ou comme un seul cluster qui contient toutes les données (dans le cas divisive) et procède itérativement par fusion ou éclatement des clusters les plus adéquats jusqu'à ce qu'un critère d'arrêt soit satisfait. La concordance d'un cluster pour la fusion/éclatement dépend de la (dis) similarité des éléments du cluster. Cela reflète l'hypothèse générale que les clusters contiennent des objets similaires. Pour fusionner/éclater des ensembles d'objets plutôt que des objets à part, on doit généraliser la distance entre des objets individuels à la distance entre sous-ensembles. De telles mesures de proximité désirées, sont appelées critères d'agrégation ou critères du lien (en anglais linkage metrics).

De nombreux critères d'agrégation ont été proposés, la difficulté du choix du critère d'agrégation réside dans le fait que ces critères peuvent déboucher sur des résultats différents. Les plus connus sont :

- Le critère du saut minimal : (Single link)

Pour single link, la distance entre deux clusters est le minimum des distances entre toutes les paires d'objets appartenant à des clusters différents. La distance entre 2 clusters C1 et C2 est définie par la plus courte distance séparant un individu de C1 et un individu de C2 :

$$D(C_1, C_2) = min(\{d(x, y)\}, x \in C_1, y \in C_2) \quad \text{(II.10)}$$

- Le critère du saut maximal : (Complete link)

Pour la méthode complète link, on utilise le maximum de ces distances, la distance entre 2 classes C1 et C2 est définie par la plus grande distance séparant un individu de C1 et un individu de C2 :

$$D(C_1, C_2) = max(\{d(x, y)\}, x \in C_1, y \in C_2) \quad \text{(II.11)}$$

- Le critère de la moyenne : (Average link)

Pour la méthode average link, on utilise la moyenne de ces distances ; ce critère consiste à calculer la distance moyenne entre tous les éléments de C1 et tous les éléments de C2 :

$$D(C_1, C_2) = \frac{1}{n_{C_1}\, n_{C_2}} \sum_{x \in C_1} \sum_{y \in C_2} d(x, y) \quad \text{(II.12)}$$

- Le critère des centres de gravité

La distance entre deux classes C1 et C2 est définie par la distance entre leurs centres de gravité comme suit :

$$D(C_1, C_2) = d(gC_1, gC_2) \quad \text{(II.13)}$$

Avec : gC_1 *est le centre de gravité de* C_1 *et* gC_2 *le centre de gravté de* C_2

Avantage des méthodes hiérarchiques (Khatir, 2012):

- Une flexibilité concernant le niveau de granularité ;
- La facilité de manipulation de toute forme de similarité ou de distance entre objets ;
- Applicable aux différents types d'attributs.

Inconvénients des méthodes hiérarchiques :

- Bien que ces méthodes soient largement utilisées, elles deviennent difficilement utilisables face à de larges bases de données, car la complexité est quadratique en fonction du nombre d'objets de la base, puisque les distances entre toutes les paires d'objets possibles doivent être calculées ;
- Elles ne peuvent pas être utilisées de façon incrémentale, puisque les distances entre toutes les paires d'objets possibles doivent être calculées ;
- Elles ne peuvent pas être utilisées de façon incrémentale, puisque ces méthodes ne peuvent jamais défaire ce qu'elles ont déjà fait auparavant ;
- Elles souffrent du problème de « l'effet de chaîne », c'est-à-dire que les clusters proches mais distincts peuvent être fusionnés s'il existe une chaîne d'objets qui les relie ;
- Le choix du critère d'arrêt qui reste vague.

II.5.2.4.2. Le clustering par partitionnement

Contrairement aux méthodes de classification hiérarchiques précédentes, qui construisent les classes progressivement, les algorithmes de partitionnement construisent directement, en sortie, une partition de l'espace des objets en k classes. Le principe général, par définition d'une partition, cela signifie que chaque classe doit contenir au moins un objet, et que chaque objet doit appartenir à une classe unique. Pour ce faire, étant donné le nombre de classes k requises, ces algorithmes génèrent une partition initiale, puis cherchent à l'améliorer en réattribuant les objets d'une classe à une autre. En pratique il est impossible de générer toutes les partitions de clustering pour des raisons évidentes de complexité. On cherche alors une « bonne » partition correspondant à un optimum local en optimisant une fonction objectif qui traduit que les objets doivent être « similaires » au sein d'une même classe, et « dissimilaires » d'une classe à une autre. Cet optimum est obtenu de façon itérative, en améliorant la partition initiale choisie plus ou moins aléatoirement, par réallocation des objets autour de centres mobiles. Les classes sont représentées par leur « centrioles », qui correspond à la moyenne de l'ensemble des objets contenus dans la classe. La méthode la plus classique et qui reste très utilisée est celle des k-means et les nombreuses variations qui en découlent.

a. Les méthodes des k-moyennes :

L'algorithme k-means, ou l'algorithme du k-moyennes été introduit par J.MacQuenn. Il est sans aucun doute la méthode de partitionnement la plus connue et la plus utilisée dans divers domaines d'application scientifiques et industrielles. Ce succès est dû au fait que cet algorithme présent un rapport coût/efficacité avantageux. Dans sa version classique, l'algorithme consiste à sélectionner aléatoirement k objets qui représente les centrioles initiaux. Un objet est assigné au cluster pour lequel la distance entre l'objet et le Centroïde est minimale. Les centrioles sont alors recalculés et l'on passe à l'itération suivante. La fonction objective traditionnellement utilisée est :

$$F = \sum_{r=1}^{k} \sum_{x_i \in c_r} (x_i - g_r)^2 \qquad \text{(II.14)}$$

Où :

- C_r est la classe numéro r

- X_i est un objet dans une classe
- g_r est le centre de classe C_r

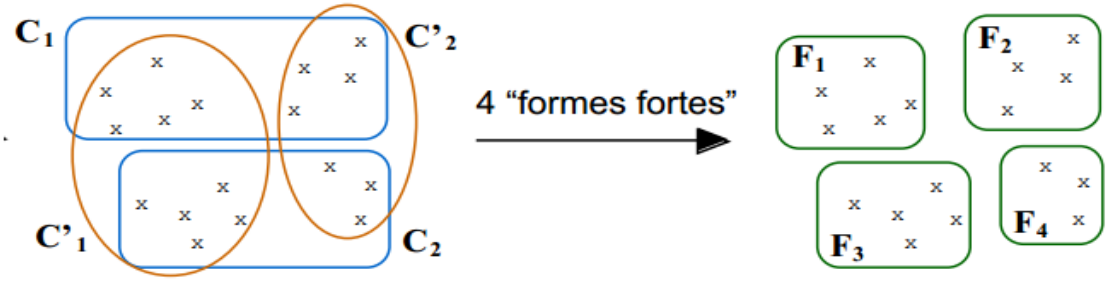


Figure II.5 *: Exemple de K-moyennes Clustering Process*

L'algorithme général de ces méthodes :

Données : k le nombre maximum de classes désiré.

Début

(1) Choisir k objets au hasard (comme centre des classes initiales)

(2) Affecter chaque objet au centre le plus proche

(3) Recalculer le centre de chacune de ces classes

(4) Répéter l'étape (2) et (3) jusqu'à stabilité des centres

(5) Editer la partition obtenue.

FIN

L'algorithme des k-means ne peut être utilisé que sur des données décrites par des attributs numériques permettant ainsi le calcul des centrioles. Dans le cas d'attributs symboliques (ou catégoriels) plusieurs variantes ont été proposées : méthodes des centres mobiles et méthodes des nuées dynamiques.

b. L'algorithme des centres mobiles :

Le principe de l'algorithme consiste à construire une partition en k classes en sélectionnant k objets au hasard de l'ensemble des objets et les considérer comme centres des classes. Après cette sélection, on affecte chaque objet au centre le plus proche en créant k classes, les centres des classes seront remplacés par des centres de gravité. Ainsi des nouvelles classes seront créées. Généralement la partition obtenue est localement optimale car elle dépend du choix initial des centres. Pour cette raison, les résultats entre deux exécutions de l'algorithme sont significativement variés (E.Forgy, 1965).

Données : k le nombre maximum de classes désiré.

Début

(1) Choisir k objets au hasard (comme centre des classes initiales)

(2) Affecter chaque objet au centre le plus proche, ce qui

Donne une partition en k classes P1= {C_1,..., C_k}

(3) On calcule les centres de gravité de chaque classe de P1, ce qui donne k nouveaux centres de classes

(4) Répéter l'étape 2 et 3 jusqu'à ce que deux itérations successives donnent la même partition

(5) Editer la partition obtenue.

FIN

Avantage des k-moyennes :

- Grande simplicité.
- Compréhensible.
- Les classes sont facilement interprétables et représentées naturellement par les centrioles.
- Il s'adapte bien pour des populations de grandes tailles.
- Complexité algorithmique est intéressante, puisqu'elle est en o (nkt) où t est le nombre d'itérations, n le nombre d'objets et k le nombre de classes.
- L'algorithme converge, généralement, avec k et t suffisamment inférieur à n (k, t<<n) Inconvénients des k-moyennes :

Parmi les inconvénients de cet algorithme, on cite :

- Le nombre de classes doit être fixé au départ ;
- Le résultat dépend du tirage initial des centres des classes ;
- Il ne détecte pas les données bruitées (isolées) ;
- Il converge souvent vers un optimum local.

II.6 Conclusion

Dans ce chapitre, nous avons passé en revue les différentes approches d'indexation sémantique et conceptuelle proposées en recherche d'information. Ces approches ont apporté la preuve que la représentation des documents et requêtes par les sens (ou concepts) de leur mots bénéfique dans un processus de recherche, permettant ainsi de résoudre les problèmes causés par les SRI classique. Ces sens sont le plus souvent identifiés par des approches de désambiguïsation utilisant des ressources linguistiques externes, Ces ressource qui peut être également remplace par des méthodes d'apprentissage automatique non supervisée applique à la collection des documents.

En basent sur ces principes, nous présentons dans le chapitre suivant nos contributions à la définition d'un novelle approche d'indexation automatique de texte arabe basée sur la similarité des termes.

PARTIE II

CONTRIBUTION

CHAPITRE III

VER UNE NOUVELLE APPROCHE D'INDEXATION AUTOMATIQUE BASE SUR LA SIMILARITÉ DES TERMES

Plan de Travail

III.1.Introduction 59
III.2. Approche proposée 59
III.2.1. Présentation de l'architecture proposée 59
III.2.2. Schéma général de l'approche 60
III.3. Conception détaillée 63
III.3.1. L'indexation 63
III.3.2. Exploitation de L'index 70
III.4. Conclusion 72

III.1.Introduction

Les SRI sont des outils informatiques qui ont pour but la mise en relation des informations contenues dans le corpus documentaire d'une part, et les besoins de l'utilisateur d'autre part. Le défit est de pouvoir retourner le maximum de documents pertinents tout en limitant le bruit et le silence documentaire, c'est-à-dire de ne pas restituer des documents qui ne répondent pas au besoin, et en même temps restituer le maximum des documents pertinents.

Dans ce chapitre nous présentons une contribution traduisant deux points de vue relatifs à l'amélioration des systèmes de recherche d'information :

Classification et catégorisation des documents : qui consiste à classer de façon automatique des ressources documentaires, généralement en provenance d'un corpus.

Extraction des termes similaires, afin de capturer tout le sens voulu par le besoin d'information, qui n'est pas exprimée par la requête.

Ce chapitre est organisé de la manière suivante. Dans la section suivante, nous discutons les problèmes généraux lors du traitement et de l'utilisation de la langue arabe dans la domaine de la recherche d'information, Au cours de la Section III.3, nous proposons l'architecture conceptuelle de notre système et nous expliquons les étapes de notre approche d'une manière générale, Ensuite dans la section III.4, nous détaillons notre approche en utilisant des schémas et en présentant les différents composants de notre système d'indexation. Nous conclusions ce chapitre par une discussion sur l'approche proposée.

III.2. Approche proposée

Dans le but d'améliorer la recherche d'information arabe, et suite au différents études et approches proposées dans ce domaine, en considérant les particularités de la langue arabe, et en se basant sur le processus d'indexation automatique et la structure interne de l'index, nous proposons une nouvelle approche d'indexation automatique de texte arabe basée sur les méthode d'apprentissage non supervisée (clustering) pour évite l'utilisation des ressources externes.

III.2.1. Présentation de l'architecture proposée

Notre approche d'indexation automatique basée sur les méthodes d'apprentissage non supervise se repose principalement sur deux axes :

III.2.1.1. l'Axe des Documents

Un mot a plusieurs sens par rapport ou domaine d'utilisation « accord de prononciation avec un sens différent ». Ce qui introduit l'obligation de la définition de domaine, thème ou contexte d'un document, donc la catégorisation de texte qui consiste à chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (étiquettes, classes).

Notre but est d'appliquer la catégorisation de texte sans utilisation des ressources externes (textes préalablement étiquetés) à cause du manque de ces dernières dans la RI arabe. Pour une requête, nous faisons la recherche sur une seule catégorie de documents, afin de minimiser les documents hors contexte et, par conséquent, limiter le bruit.

Le principe de cette approche consiste, dans un premier temps, à construire une matrice des documents, dons laquelle chaque terme et représente par son poids TF-IDF, Ensuite en applique la méthode de clustering (k-means) sure cette représentation des documents. Le résultat est un ensemble de clusters où chaque cluster est considéré comme une catégorie.

Le choix de K (Nombre de catégories à construire à partir du corpus) est relié au corpus. Certains observations sur l'ensemble des documents de l'index peuvent nous aidons pour cette décision, ou dans un cas général, on peut se basé sur certains catégorisations qui existent déjà telles que : (k=10) qui est retenu à partir de la structure du portail thématique de Wikipédia arabe «جغرافيا – تاريخ – ثقافة – تقنية – مجتمع – طب – أديان - رياضة – علوم - سياسة»[1]

Hypothèse :

- Un document ne peut appartenir qu'à une seule catégorie ;
- Une catégorie contient 1 ou N documents.

III.2.1.2. l'Axe des Termes

La catégorisation répond à la relation entre les documents, mais elle ne donne pas le lien entre les termes similaires, car plusieurs mots peuvent avoir le même sens, dans son contexte d'utilisation, mais certains termes sont moins utilisés qu'autres, pour certains cas

- un terme plus complet que l'autre (بكي / انتحب) ;
- un terme plus puissance que l'autre (أنهك / تعب) ;
- un terme plus célèbre que l'autre (فؤاد / قلب) ;
- un terme plus d'affiliation à la langue que l'autre (حكم ذاتي/ استقلال) ;
- un terme plus spécialisés que l'autre (بكي / انتحب) ;
- un terme plus sophistiqué dans la langue que l'autre (أم فلان/والدة فلان).

Pour cette raison, on va appliquer un processus de regroupement (clustering) des mots (termes), à partir d'une matrice des termes où chaque terme est représenté par un vecteur qui désigne son poids dans les documents par son fréquence, et chaque document représentent une dimension.

Ce processus est appliqué à chaque ensemble des termes d'une catégorie obtenue de la phase de catégorisation, afin d'affecter le bon sens de chaque terme dans son domaine d'utilisation.

Hypothèse :

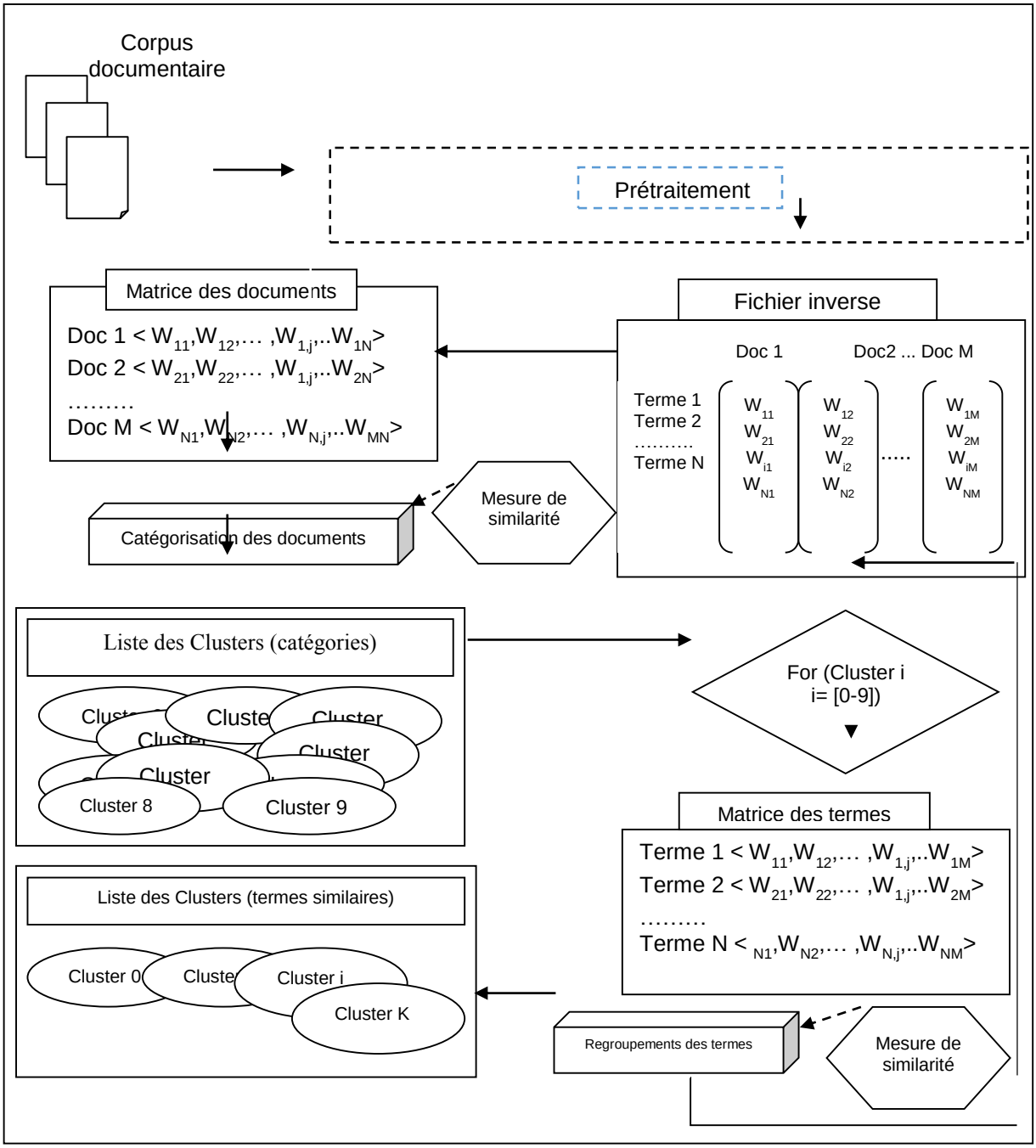
- Chaque cluster est créé à partir d'un terme racine et 7 termes similaire ;
- Un terme appartenir à 1 ou N clusters dans l'ensemble des catégories i, i= [0-9].

III.2.2. Schéma général de l'approche

La figure ci-dessus représente la structure générale proposée pour notre approche. Elle détaille les différentes étapes d'indexation automatique des documents arabes afin de créer un nouvel index basé sur les méthodes de clustering.

Ce schéma représente le passage d'un index ordinaire (prétraitement de texte, pondérations des attributs et génération de fichier inverse), à un index évolué qui repose sur le sens d'un terme dans les différents domaines, et sur la présence de ses termes similaires.

[1]https://ar.wikipedia.org/wiki/تصفح:بوابة


Corpus documentaire
Prétraitement
Matrice des documents
Doc 1 < W11,W12,… ,W1,j,..W1N>
Doc 2 < W21,W22,… ,W1,j,..W2N>
……
Doc M < WN1,WN2,… ,WN,j,..WMN>
Fichier inverse
Doc 1
Doc2 ... Doc M
Terme 1
Terme 2
……
Terme N
W11 W21 Wi1 WN1
W12 W22 Wi2 WN2
W1M W2M WiM WNM
Mesure de similarité
Catégorisation des documents
Liste des Clusters (catégories)
Cluster 8
Cluster 9
For (Cluster i i= [0-9])
Matrice des termes
Terme 1 < W11,W12,… ,W1,j,..W1M>
Terme 2 < W21,W22,… ,W1,j,..W2M>
……
Terme N < N1,WN2,… ,WN,j,..WNM>
Liste des Clusters (termes similaires)
Cluster 0
Cluster i
Cluster K
Regroupements des termes
Mesure de similarité

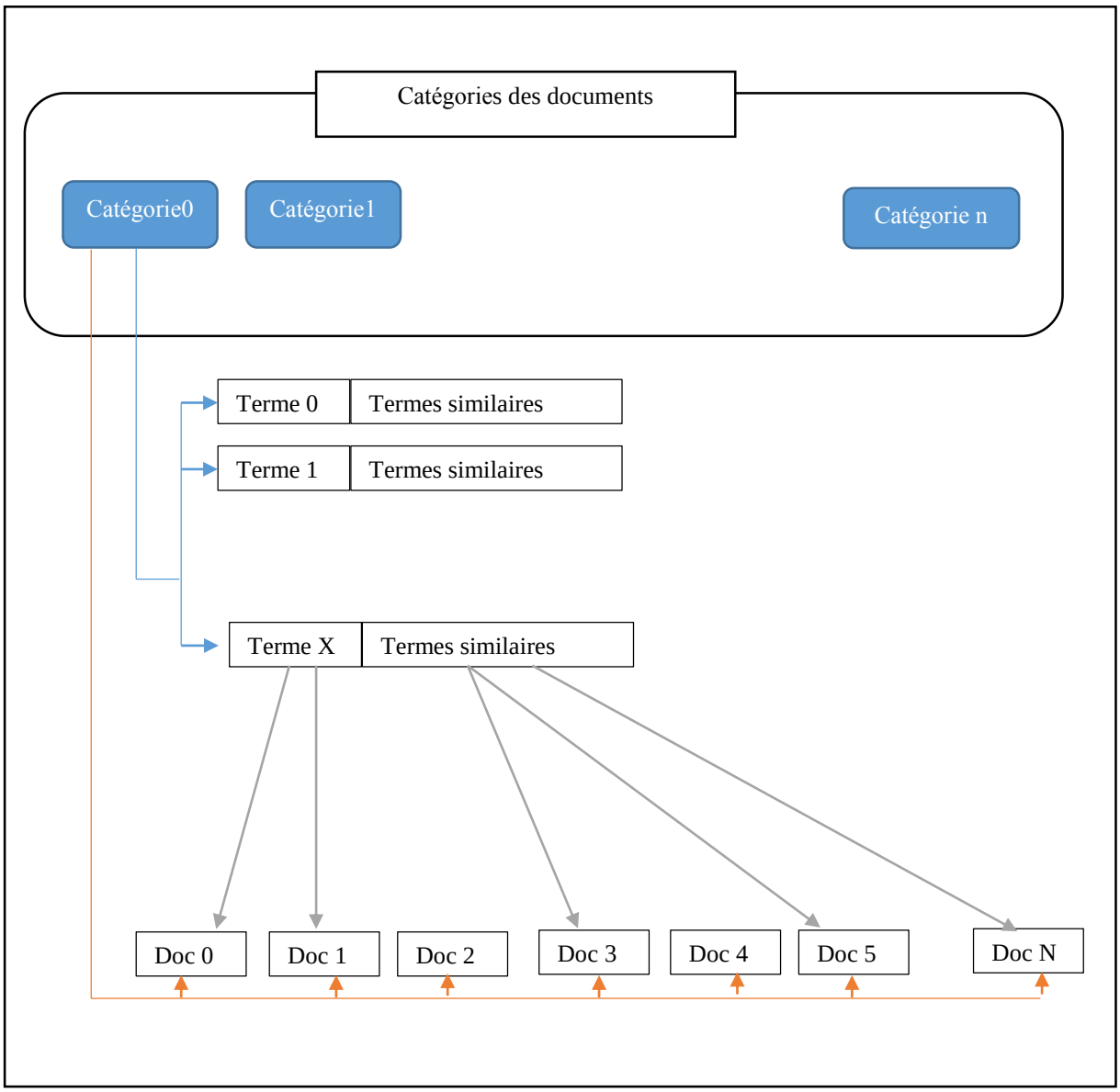


Figure III.1 : *Schéma général de l'approche proposée*

III.3. Conception détaillée

Afin de clarifier tous les points concernant notre approche, on va expliquer toutes ses étapes dans cette section, à partir de prétraitement initial du corpus, jusqu'à la création de l'index final et l'établissement des différents liens entre ses composants.

III.3.1. L'indexation

La création de l'index passe par 4 étapes, à partir de prétraitement de texte et la construction de Fichier Inverse, en passant par la catégorisation des documents, ainsi le clustering et le regroupement des termes pour chaque classe ou catégorie, jusqu' à la création des liens entre les différents termes, clusters, catégories et documents, afin de produire un nouvel index.

III.3.1.1. Prétraitement

Cette étape a pour but de caractériser les textes par leurs mots significatifs (noms, verbes et adjectifs, en éliminant les noms propres, les abréviations et certains verbes). Chaque texte doit être retenu dans sa forme canonique. Cette opération est réalisée par une chaîne de traitement illustrée par la figure III.2.

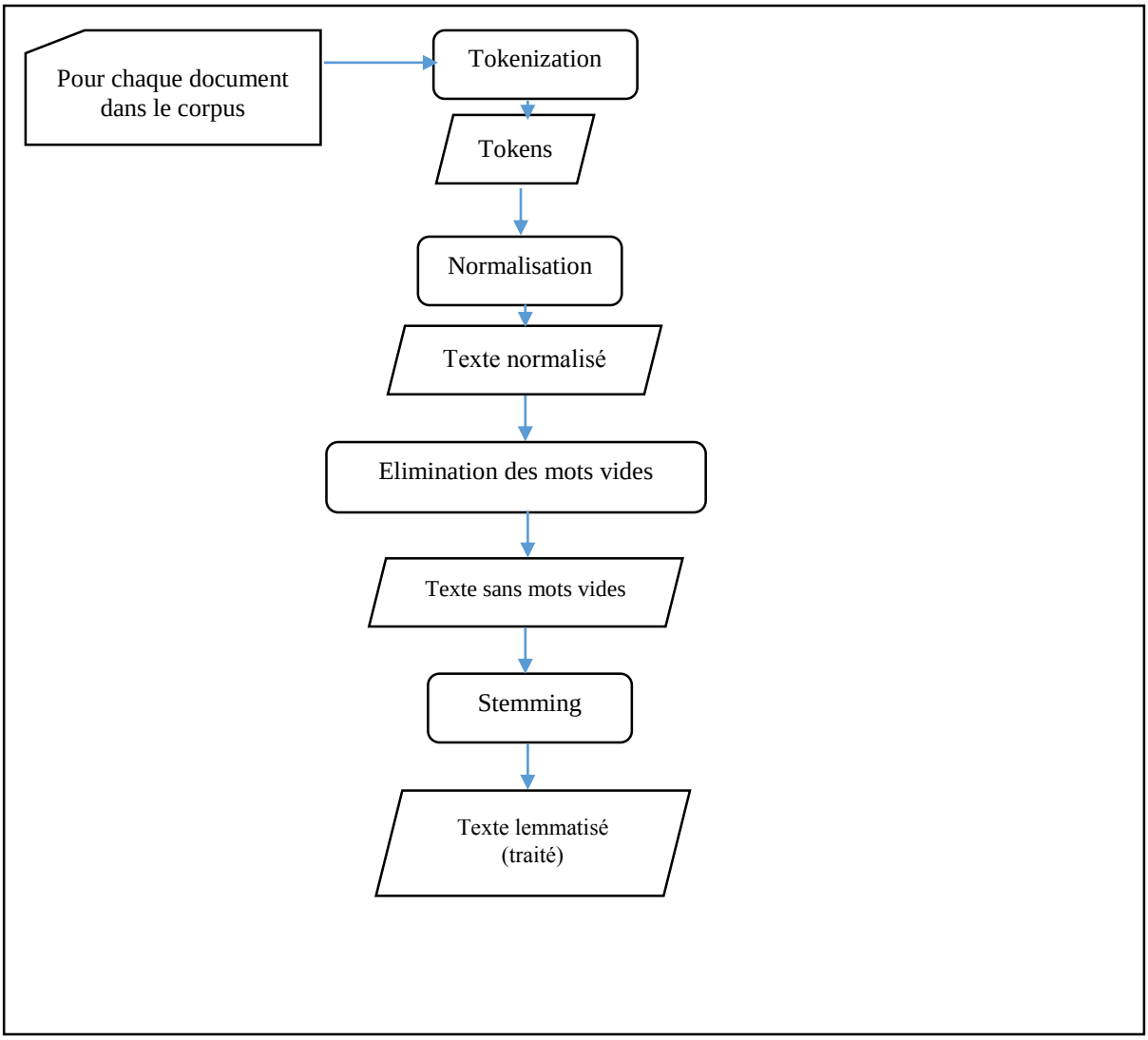


Figure III.3. *Processus du Prétraitement des documents*

Afin de rendre les documents de corpus comparables, nous utilisons le format d'encodage UTF-8.

III.3.1.1.1. Segmentation de Texte (Tokenisation)

L'étape de Tokenisation consiste à segmenter le texte aux mots par la suppression des espaces, des marques de ponctuation et des signes mathématiques. Après la segmentation, chaque mot obtenu est traité à part, si c'est un mot arabe (par l'utilisation des expression régulier) alors on passe à l'étape suivante sinon on le rejette.

III.3.1.1.2. Normalisation

Afin de manipuler les variations du texte de la requête correspondante écrite en arabe, nous avons appliqué plusieurs genres de normalisation sur les documents

Pour chaque Token arabe en entrée, la normalisation consiste à :

- Retirer les signes diacritiques ;
- Retirer les caractères spéciaux ;
- Retirer les lettres étrangères ;
- Retirer les chiffres ;
- Remplacer le إ ou le أ initial par l'alif nu ا;
- Remplacer le آ par le ا;
- Remplacer le ىء d'ordre par le ئ;
- Remplacer le ي final par le ى;
- Remplacer le ة final par le ه.

III.3.1.1.3. Elimination des Mots Vides

Les mots vides sont les mots qui occurrent plus dans le corpus, et qui n'ont pas une valeur importante. Ils sont tellement communs qu'il est inutile de les indexer ou de les utiliser dans une recherche, ils sont souvent ignorés. La stratégie générale pour déterminer une liste des mots vides est d'identifier les termes les plus fréquents, mais ce n'est pas vraiment efficace, nous avons souvent besoin de les filtrer manuellement suivant leur contenus sémantiques et leurs relations avec les documents du corpus. L'utilisation d'une liste des mots vides réduit de manière significative le nombre des termes qu'un système d'indexation et dans le but de réduction de la matrice initiale des occurrences.

ان بعد ضد يلي الى في من حتى وهو يكون به وليس احد على وكان تلك كذلك التي وبين فيها
عليها ان وعلى لكن عن مساء ليس منذ الذي اما حين ومن لا ليسب وكانت اي ما عنه حول دون
مع لكنه ولكن له هذا والتي فقط ثم هذه انه تكون قد بين جدا لن نحو كان لهم لان اليوم لم هؤلاء
فان فيه ذلك لو عند اللذين كل بد لدى وثي ان ومع فقد بل هو عنها منه بها وفي فهو تحت لها او
اذ علي عليه كما كيف هنا وقد كانت لذلك امام هناك قبل معه يوم منها الى إذا هل حيث هي اذا او
و ما لا الي الي مازال لازال لايزال مايزال اصبح اصبح امسى امسى اضحى اضحى ظل مابرح
مافتى ماانفك بات صار ليس ان كان ليت لعل لاسيما ولايزال الحالي ضمن اول وله ذات اي بدلا
اليها انه الذين فانه وان والذي وهذا لهذا الا فكان ستكون مما ابو بان الذي اليه يمكن بهذا لدي
وان وهي وابو ال الذي هن الذى

Figure III.3 : *Les mots vides arabes*

III.3.1.1.4. Lemmatisation (Streaming)

L'objectif de la Lemmatisation est de trouver la forme représentative d'index d'un mot à partir de sa forme représentée dans le document.

La segmentation des Token arabes en préfixes, suffixes et stems est une étape essentielle de l'analyse, elle suit les contraintes suivantes:

- La longueur du préfixe ne dépasse pas 5 lettres ;
- La longueur du suffixe ne dépasse pas 6 lettres ;
- La longueur du stem est au minimum 2 lettres.

III.3.1.1.5. Pondération et Création de Fichier Inverse

La phase de pondération consiste à effectuer un poids à chaque terme dans le document. Ce poids est obtenue par la mesure TF-IDF *(Section I.3.1.3-4)* Afin de construire le Fichier Inverse *(Section I.3.1.3-5)*, qui est le composent fondamental de la recherche d'information classique

Dans notre approche, ce fichier représente la pierre de fondation pour l'étape suivante.

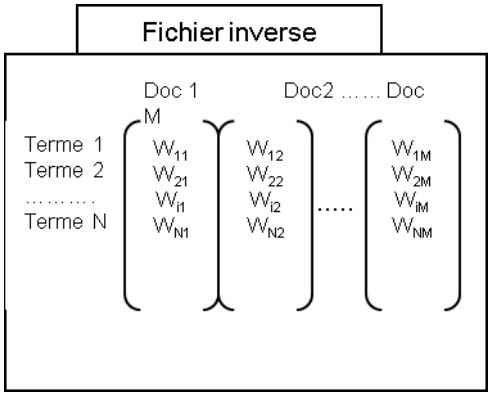


Figure III.4 : *Structure de Fichier Inverse*

III.3.1.2. Catégorisation des Documents

La catégorisation de texte consiste à chercher une liaison fonctionnelle entre un ensemble de textes (ou termes) et un ensemble de catégories (étiquettes, classes). Cette liaison fonctionnelle, que l'on appelle également modèle de prédiction, est estimée par un apprentissage automatique (traduction de machine learning method). Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement étiquetés, dit ensemble d'apprentissage, à partir duquel nous estimons les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire le modèle qui produit le moins d'erreur en prédiction (JALAM, 2003).

Dans notre approche, on évite d'utiliser les ressources externes (textes préalablement étiquetés), et on passe au processus d'apprentissage non supervisée « Clustering ».

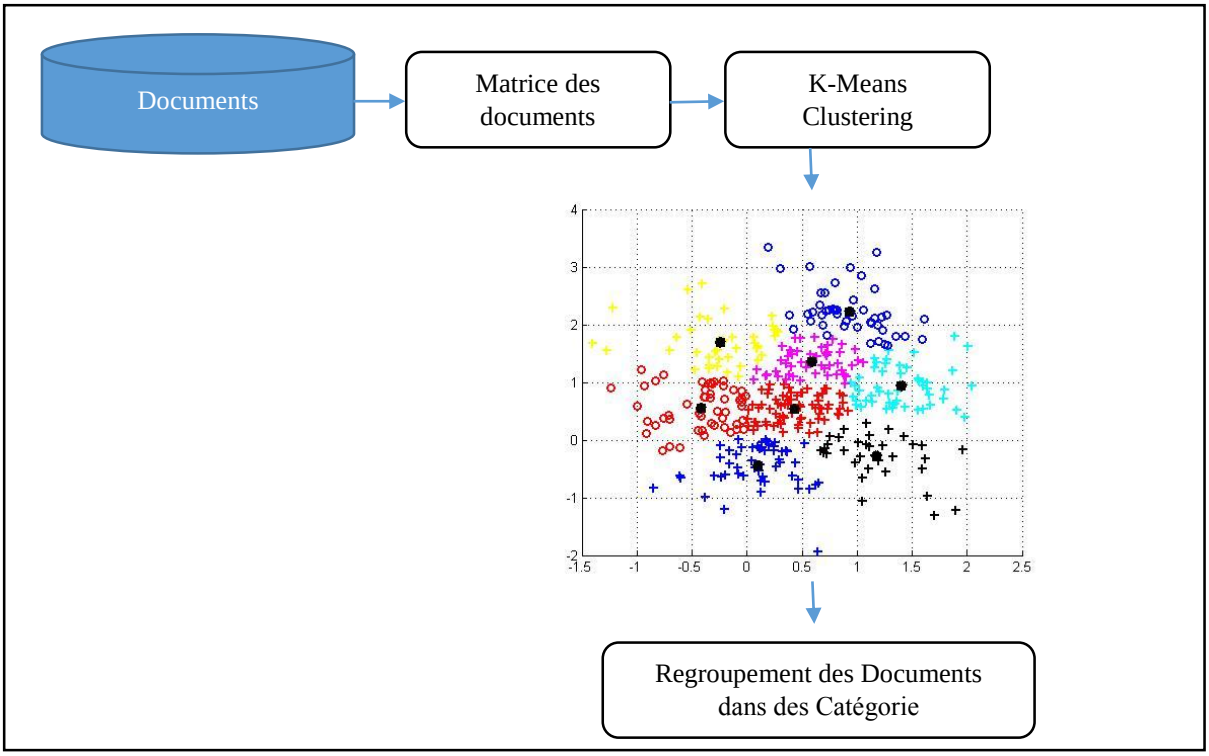


Figure III.5 *: Processus de Catégorisations des documents*

Ce processus et basé sur les 4 étapes ci-dessous :

III.3.1.2.1.Constriction de Matrice des Document :

Le Fichier Inverse Construit dans la phase de prétraitement est considéré comme une matrice des documents dont laquelle chaque document et représenté par l'ensemble des poids des termes (TF-IDF), et chaque terme a une dimension dans l'espace des documents. .

Matrice des Documents

Doc 1 < W_{11},W_{12},… ,$W_{1,j}$,..W_{1N}>
Doc 2 < W_{21},W_{22},… ,$W_{1,j}$,..W_{2N}>
……
Doc M < W_{N1},W_{N2},… ,$W_{N,j}$,..W_{MN}>

Figure III.6 *: Matrice des Documents*

III.3.1.2.2.Choix de mesure de similarité :

On fait le choix de la mesure de similarité à partir des différents mesures présentées dans la (section II.4.3.1), afin d'obtenir les liaisons entre les différents documents.

Dans un cas général, deux documents contiennent le même terme sont considérés comme des documents similaires

Sur la base de l'étude (Guadalupe J.Torres, July 2008), Nous choisirent le mesure « Euclidian distance » Afin de détermine la similarité entre deux document dans le processus de clustering.

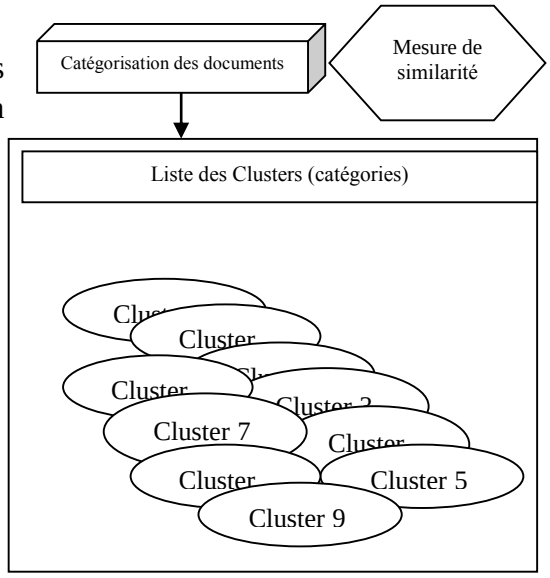


Figure III.7 *: Processus de catégorisation des documents*

III.3.1.2.3.Choix de l'Algorithme de Clustering :

Dans notre approche, la quantité des données est énorme, le choix de l'algorithme de clustering doit prendre en considération cette hypothèse. Pour cela, on fait le choix de l'algorithme k-means qui a une complexité liniaire.

III.3.1.2.4.Le Choix de K

Le choix de K=10 (nombre de cluster) est obtenu à partir de la structure des portails thématiques de Wikipédia arabe «جغرافيا – تاريخ – ثقافة – تقنية – مجتمع – طب– أديان - رياضة – علوم - سياسة», qui donne une structure général sur le nombre des catégories pouvant être eu dans une collection des documents arabes..

III.3.1.3. Le regroupement des Termes (Clustering)

Dans La phase précédente, on a construit un ensemble des clusters, dans lesquels chaque cluster contient un ensemble des documents.

Pour chaque cluster, on applique un processus de clustering basé sur les termes. Pour cela, on passe par les étapes suivantes :

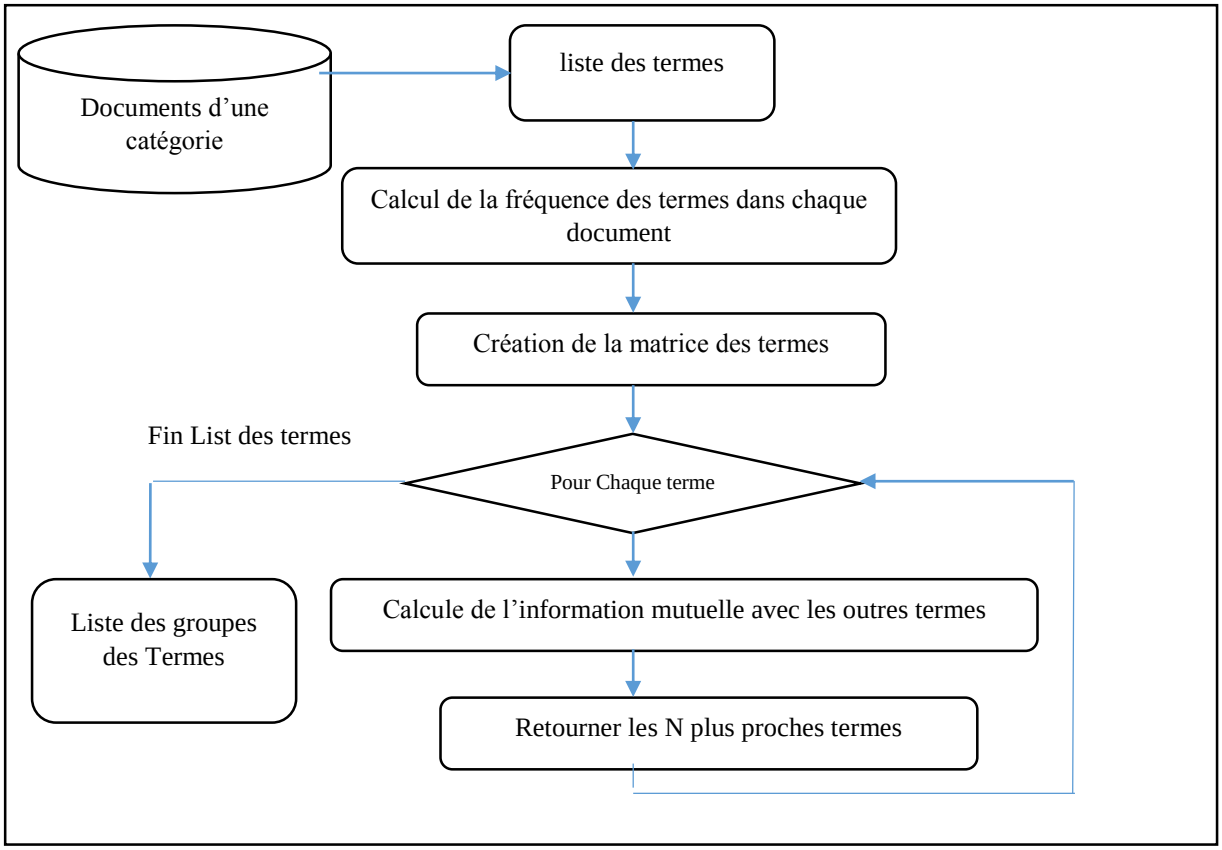


Figure III.8 : *Processus de Clustering (regroupement) des termes*

III.3.1.3.1.Constriction de Matrix des Termes :

A partir des documents dans un cluster i où i= [0-9], on construit une matrice des termes, dans laquelle chaque terme est représenté par un vecteur qui désigne son poids dans les différents documents d'une catégorie i, par son fréquence, et chaque document est une dimension dans l'espace des termes.

Matrice des Termes

Terme 1 < W_{11},W_{12},… ,$W_{1,j}$,..W_{1M}>
Terme 2 < W_{21},W_{22},… ,$W_{1,j}$,..W_{2M}>
……
Terme N < W_{N1},W_{N2},… ,$W_{N,j}$,..W_{NM}>

Figure III.9 *: Matrice des Termes*

Certains linguistes estiment qu'on peut classifier un mot selon les mots avec lesquels il cooccurrence (Turenne, 2001). Étudier les cooccurrences permet d'avoir beaucoup d'informations sur les mots qui « accompagnent » le terme recherché. Nous pouvons donc connaitre sa signification et obtenir des renseignements sur son utilisation en ayant le contexte dans lequel il est utilisé.

III.3.1.3.2.Choix de mesure de similarité :

Pour le regroupement des termes, on utilise la mesure d'information mutuelle (IM) en se basant sur le nombre de fois qu'un mot apparaît dans une certaine catégorie.

Cette mesure provient de la théorie de l'information où elle est définie sur deux variables aléatoires. Cependant, dans le cas des cooccurrences, elle a été adaptée afin d'être définie sur les valeurs de variables aléatoires. L'information mutuelle reliée à la réalisation de deux

évènements particuliers x et y, dans notre cas entre deux mots particuliers. La formule de cette mesure a été mentionnée dans la section (II.5.2.1.1.).

L'information mutuelle mesure la quantité d'information que la présence d'un certain mot va apporter au sujet de la présence d'un autre mot (équation II.7).

Il est facile de voir que, plus deux mots ont tendance à apparaître ensemble, plus le score sera élevé. En effet, si les deux mots étaient totalement indépendants, le rapport des probabilités serait de 1, donnant donc une valeur finale de 0. À l'opposé, si les deux mots ont une forte tendance à apparaître ensemble, alors la probabilité au numérateur dépassera l'autre, faisant croître le score d'information mutuelle (Miry, 2007).

III.4.1.3.3.Choix de l'Algorithme de Clustering :

Pour la phase de Clustering des termes, on se basé sur la structure de l'ontologie Arabic WordNet pour déterminer le nombre des termes dans chaque cluster (Concept), Le nombre moyen des objets d'un concept est égale à 7, pour cela, chaque terme est regroupé avec les 7 termes ayons maximisé la mutuelle information.

L'Algorithme suivante représente les étapes de la création des clusters (concept)

Données : Liste des Termes d'une catégorie.

Début

(1) récupère un terme

(2) calcule la mutuelle Information avec tous les autre terme

(3) retourner les 7 termes qui maximise la valeur de MI

(4) Répéter l' étape (1),(2) et (3) jusqu' à la fine de liste des termes

FIN

III.3.1.4 La création de l'Index

La création de l'index nécessite d'établir des liens entre ses différents composants, d'une manière qui facilite la recherche des documents. Pour cela, deux index sont crées

III.6.1.4.1. Index initial

Le premier index ou l'index initial a pour but de définir le sujet de la requête, ou dans notre cas la catégorie, pour cela, chaque terme et représenté par sa fréquence totale dans chaque catégorie, le tableau suivant représente la structure générale de cet index.

Catégorie	0	1	2	3	4	5	6	7	8	9
Terme 1	P(1,0)	P(1,1)	P(1,2)	P(1,3)	P(1,4)	P(1,5)	P (1,6)	P(1,7)	P(1,8)	P(1,9)
Terme 2	P(2,0)	P(2,1)	P(2 ,2)	P(2,3)	P(2,4)	P(2,5)	P(2,6)	P(2,7)	P(2,8)	P(2,9)
……	……	…….	…….	…….	…….	…….	……..	……..	…….	…….
Terme n	P(n ,0)	P(n ,1)	P(n,2)	P(n,3)	P(n,4)	P(n,5)	P(n,6)	P(n,7)	P(n,8)	P(n,9)

Tableau III.4 : *Index initial (terme - Catégorie)*

Le calcule du poids entre un terme et une catégorie est réalisé par une des deux méthodes suivantes :

1- IDF Inverse

L'IDF a pour but de mesurer la représentativité globale du terme vis-à-vis de la collection des documents. En effet, un terme fréquent dans la collection a moins d'importance qu'un terme moins fréquent.

Donc nous inversons la formule d'IDF Pour obtenir le poids d'un terme dans une collection des documents, ou dans notre cas, dans une catégorie

$$IDF_i = (\frac{n_i}{N}) \quad \text{(II.1)}$$

Où N représente le nombre de documents de la collection et n_i est le nombre de documents dans lesquels le terme t_i apparaît.

2- Centroïde

La catégorisation des documents par l'algorithme K-means a pour résultat supplémentaire, la liste des centroïdes. En d'autres mots, chaque catégorie a un document central, et comme le processus de catégorisation est basé également sur les termes, chaque terme à un poids relié à ce centroïde.

En effet ce poids peut être utilisé pour indiquer la catégorie où un terme peut appartenir.

Le choix de la catégorie est effectué par la formule suivante :

$$\boldsymbol{Catégorie(q) = MAX\left(\sum_{terme\ i \in q} poids(i, Categorie\ j)\right)\ j = [0-9]} \quad \text{(III.2)}$$

III.6.1.4.2. Index des Termes Similaire

Le deuxième index qui exploite le clustering des termes, est utilisé pour la détection des termes associés à ceux de la requête, chaque terme dans une catégorie appartient à un cluster, et l'ensemble des termes d'un cluster représente un groupe des termes similaires. Par ce principe, la recherche d'un terme i dans une catégorie j nécessite la recherche de tous les termes présents dans le même cluster de terme i.

En fin, dans cet index, chaque terme est représenté par son cluster, documents et ses poids TF-IDF dans ces documents

	Cluster	Document	TFIDF	Document	TFIDF	……
Terme 1	1	Doc 1	TFIDF(Terme1,Doc1)	Doc 5	TFIDF(Term1,Doc5)	……
Terme 2	1	Doc 1	TFIDF(Terme2,Doc1)	Doc 40	TFIDF(Terme2,Doc40)	……
…….	……	……	……	……	……	……
Terme K	x	Doc m	TFIDF(Terme x,Doc m)	Doc m+1	TFIDF(Terme x,Doc m+1)	……

Tableau III.5 : *Index Terme Cluster*

III.3.2. Exploitation de L'index

L'index créé a une structure unique qui impose l'obligation de la définition de son utilisation pour répondre à un besoin d'information.

Comme nous avons déjà mentionné, une requête cible une catégorie spécifique ou un domaine unique, pour cela, on commence par identifier la catégorie désiré, une fois ça est effectué, on passe vers la liste des clusters des termes pour récupérer les termes similaires, pour capturer tous les sens voulus par le besoin d'information qui n'est par exprimé dans la requête. Une dernière étape consiste à retourner la liste des documents triée selon la mesure TF-IDF.

La figure ci-dessous représente notre processus de sélection de la catégorie d'une requête.

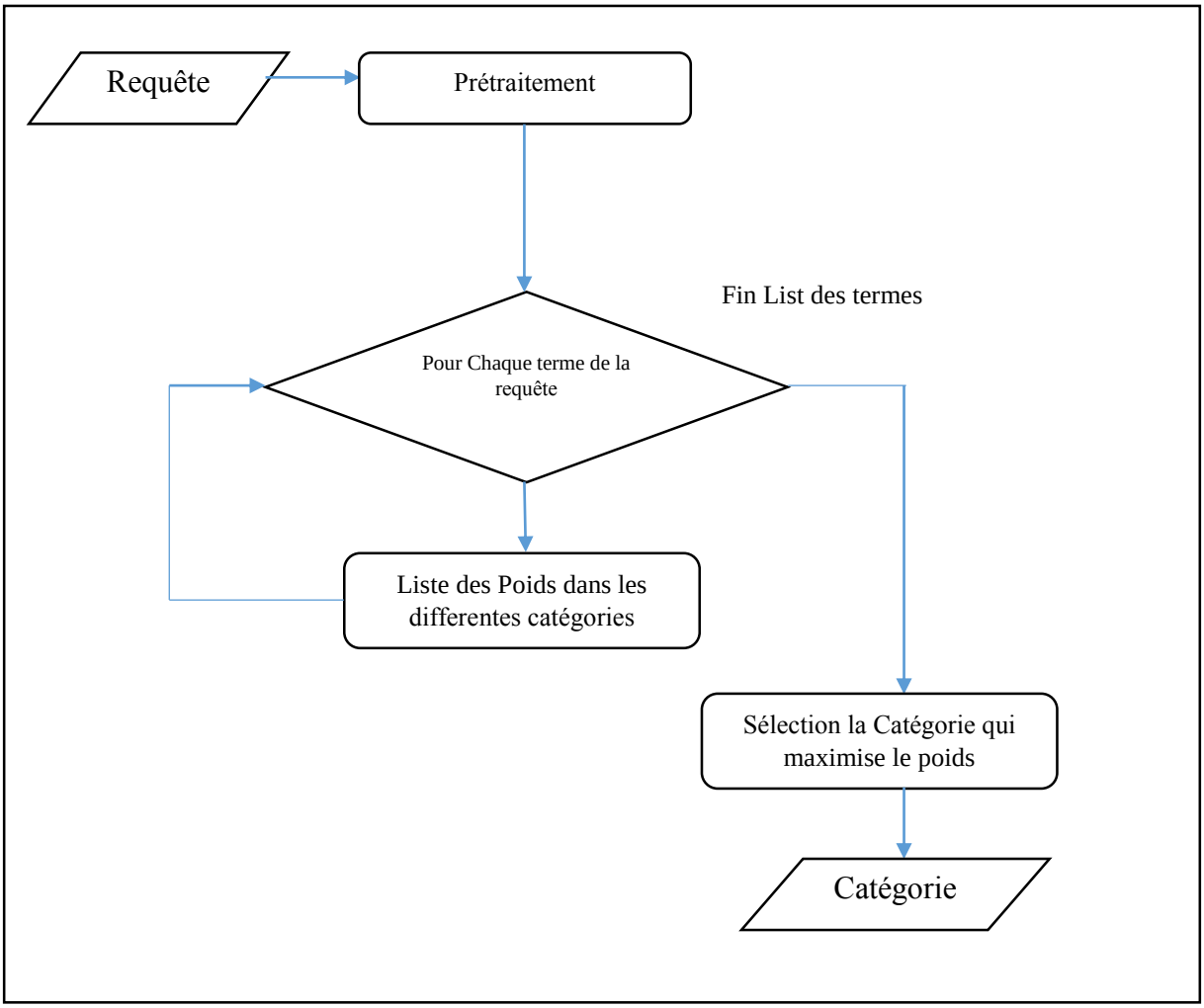


Figure III.10 : *Processus de sélection de la catégorie d'un Requête*

III.4. Conclusion

Dans cette étude, nous avons proposé une nouvelle approche d'indexation automatique pour améliorer la recherche d'information arabe. L'idée se bas sur la proposition qu'un besoin d'information cible un domaine d'étude spécifique, et que l'utilisateur ne peut pas exprimer son besoin par sa requête.

Les méthodes actuelles basées sur les ressources externes, ne répondent pas aux problèmes de la langue arabe dans la recherche d'information.

Par l'utilisation des méthodes de l'exploitation des données (Data Mining), nous essayons de résoudre ce problème, en se basent sur l'apprentissage non supervisée.

Nous consacrons maintenant le chapitre suivant à la réalisation et l'implémentation des différentes procédures de notre approche, ainsi à l'évaluation des résultats obtenus.

CHAPITRE IV

IMPLÉMENTATION ET EXPÉRIMENTATION

Plan de Travail

IV.1. Introduction 73
IV.2. Environnement des expérimentations 73
IV.2.1. Matériel 73
IV.2.2. Logiciels 73
IV.3. Présentation de l'application 75
IV.3.1. RapidMiner Catégorisation des documents et traitement de texte 75
IV.3.2 Java Création et exploitation de l'index 80
IV.3.3 Interface d'évaluation 85
IV.4 Expérimentation 86
IV.4.1 Contexte d'expérimentation 86
IV.4.2 Description de Corpus de Test 86
IV.4.2 Méthodologie de test 86
IV.4.3 Test et Résultats 86
IV.4.4. Interprétation 97
IV.5 Conclusion 99

IV.1. Introduction

Après avoir exposé les différents aspects théoriques relatifs à notre approche d'indexation automatique de texte arabe citées précédemment, nous présentons, dans ce chapitre, une mise en œuvre de notre contribution, afin de juger nos propositions et d'évaluer l'efficacité de notre système d'indexation automatique, en présentant l'environnement et les différents outils utilisé pour l'implémentation de notre approche proposée.

IV.2. Environnement des expérimentations

IV.2.1. Matériel

Notre application est exécuté sur un ordinateur de type TOSHIBA d'un :

- Processeur Intel® Core™ i3, vitesse 2.40 GHz.
- Espace mémoire de 4 GB, type DDR3.
- Disque dur de 500 GB.
- Système d'exploitation : Windows 7 Home Premium, 64bits.

IV.2.2. Logiciels

Nous avons mis en place un prototype en utilisant le langage de programmation java, afin de démontrer la faisabilité de nos contributions

IV.2.2.1. Environnement de développement

- *Langage de Programmation Java*

Java est un langage de programmation orienté objet développé par Sun Microsystems (Sun) au début des années 1990.

Java a de multiples particularités. La première d'entre elle se situe dans son mode de distribution, car en effet, son code source est accessible en Open Source. De plus, son accès est libre et gratuit, L'autre aspect est celui d'être multi-plateforme, c'est à dire que les programmes écrits en java peuvent s'exécuter de la même manière sur une plateforme Windows, MacOS, Linux ou Solaris. Ceci est dû à son architecture faisant appel à une Machine Virtuelle (ou VM en anglais : Virtual Machine) (Java, 2017).

- *Java Development Kit (JDK 1.8.0_121)*

Le Java Development Kit (JDK) désigne un ensemble de bibliothèques logicielles de base du langage de programmation Java, ainsi que les outils avec lesquels le code Java peuvent être compilé, transformé en byte code destiné à la machine virtuelle Java (Java Development Kit (JDK), 2017).

- *Environnement de développement NetBeans 8.1*

NetBeans (NetBeans IDE 8.2 Information, 2017) est un environnement de développement intégré (IDE) pour Java, placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, XML et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages web), NetBeans est disponible

sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et Open VMS, NetBeans est lui-même développé en Java, ce qui peut le rendre assez lent et gourmand en ressources mémoires.(NetBeans, 2017)

IV.2.2.2. Outils utilisés

- *RapidMiner Studio 7.4*

RapidMiner est à la fois un logiciel open source gratuit, et un produit commercial, destiné au mining (de texte, d'images ou autres) en général.

RapidMiner propose des fonctionnalités de "data mining" et de "machine learning" dont : chargement de données et transformation (ETL), prétraitement et visualisation, modélisation, évaluation et déploiement de données. Les processus de "data mining" peuvent être construits à partir d'opérateurs "intégrables les uns aux autres", décrits dans des fichiers XML et créé dans l'interface graphique (GUI) de RapidMiner. Le logiciel a été écrit en Java. Il intègre également des schèmes d'apprentissage et des évaluateurs d'attributs de l'environnement de "machine learning" Weka ainsi que des schèmes de modélisation statistique du R-Project (RapidMiner Studio, 2017).

- *RapidMiner Text Processing Extension*

L'extension de texte ajoute tous les opérateurs nécessaires à l'analyse de texte statistique. Elle permet de charger des textes à partir de différentes sources de données, les transformer par un énorme ensemble de différentes techniques de filtrage et enfin analyser les données textuelles.

Les extensions de texte prennent en charge plusieurs formats de texte, y compris le texte brut, HTML ou PDF, ainsi que d'autres sources de données. Il fournit des filtres standards pour la Tokenisation, la suppression, le filtrage de mots d'ordre ou la génération de n-gramme pour fournir tout ce qui est nécessaire pour préparer et analyser des textes (Text Processing, 2017).

- *Lucene 3.0.3*

Lucene est une API libre écrite en Java qui permet de construire des moteurs classiques et performants de recherche d'informations au sein de documents textuels.

Lucene n'est pas une application compète mais un ensemble de classes java (packages), Cette bibliothèque offre la possibilité d'indexer et la rechercher de texte, et basé sur le modèle vectoriel et la pondération du terme TF-IDF. (Michael , Erik, & Otis, 2010)

- *Microsoft Excel 2013*

C'est un logiciel tableur de la suite bureautique Microsoft Office, développée et distribuée par l'éditeur Microsoft. Microsoft Excel connaît de nombreuses évolutions et voit aujourd'hui ses possibilités s'étendre grâce à Power Bi au Big data (Excel, 2017).

- *Apache POI 3.16*

C'est un projet de l'Apache Software Fondation permettant de manipuler avec la technologie Java divers types de fichiers créés par Microsoft Office. POI est constitué de plusieurs composants parmi les quelles, XSSF (XML Spreadsheet Format ou Format de feuilles de calcul XML) : fichiers Excel (XLSX) (Apache POI, 2017).

- *JFreeChart-1.0.0*

C'est une API Java permettant de créer des graphiques et des diagrammes de très bonne qualité. Cette API est open source et sous licence LGPL (Welcome To JFreeChart, 2017).

- *JAVAMI*

C'est un outil qui s'adresse aux travaux sur des ensembles de données discrets pour la classification. Toutes les fonctions attendent des entrées discrètes. Il fournit des implémentations des fonctions de la théorie de l'information de Shannon. La boîte à outils a été développée pour soutenir notre recherche sur les algorithmes de sélection de fonctionnalités (JavaMI, 2017).

Contient des fonctions pour :

- Calcul de l'entropie, H (X)
- Calcul de l'entropie conditionnelle, H (X | Y)
- Calcul de l'information mutuelle, I (X; Y)
- Calculer l'information mutuelle conditionnelle, I (X; Y | Z)
- Générer une variable aléatoire commune

IV.3. Présentation de l'application

L'implémentation de notre approche est divisée en deux parties :

1. Prétraitement et Clustering : qui se basent principalement sur le logiciel RapidMiner, qui serve à la réalisation des différentes étapes de clustering et prétraitement de texte ;
2. Création et Exploitation de l'index : à partir des différentes sorties de l'étape précédente, une application java à été mise en œuvre pour réaliser les liens entre les documents et les termes grâce à divers liens des clusters et catégories, et adapte au processus de recherche et d'évaluation, afin de tester la performance de notre approche.

IV.3.1. RapidMiner Catégorisation des documents et traitement de texte

Le prétraitement de texte et les différents processus de catégorisation (Clustering) des documents, se fait avec le logiciel RapidMiner Studio 7.4, à cause de la difficulté de les implémenter, et la gestion de consommation de la mémoire, qui pose un grand problème dans les opérations d'apprentissages.

Les résultats délivrés par ce logiciel en format Excel (XLSX) sont traités par notre application afin de les utiliser pour la création de l'index et les liens entre ses différents composantes (termes-clusters-catégories-documents)

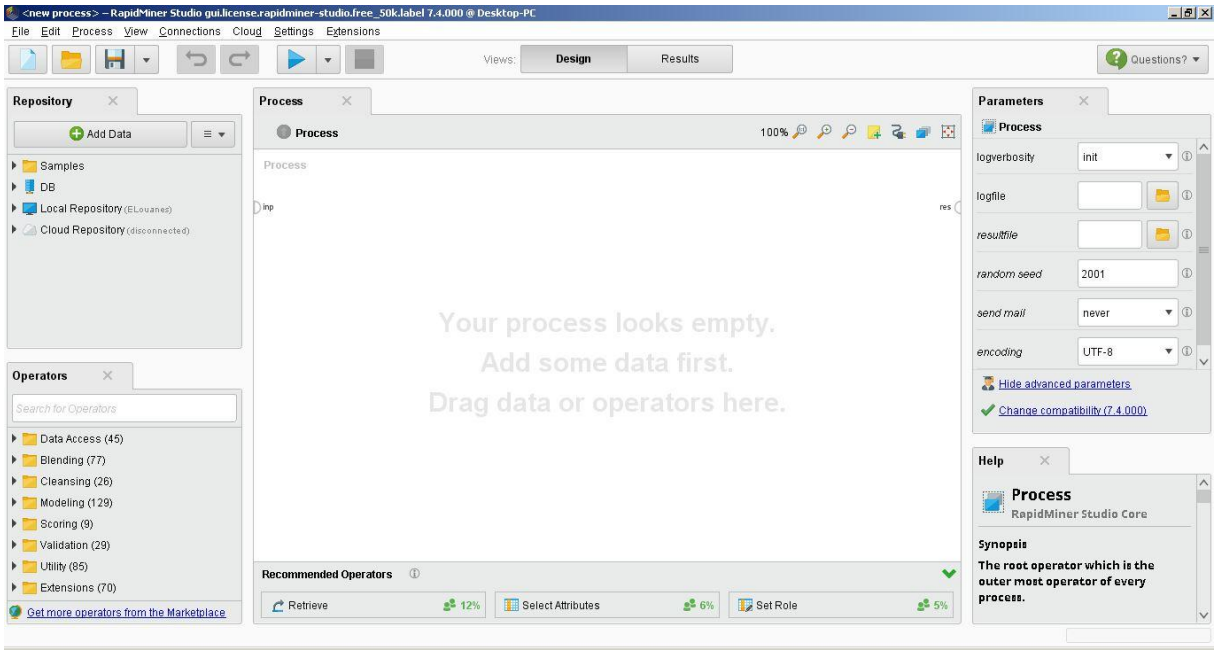


Figure IV.1 : *Interface de RapidMiner Studio 7.4*

Les différentes étapes réalisées à l'aide de ce logiciel sont représentées comme suit :

IV.3.1.1 Prétraitement des documents (Corpus)

Le prétraitement de texte illustré dans la figure IV.2 et passée par les 4 phases suivantes :

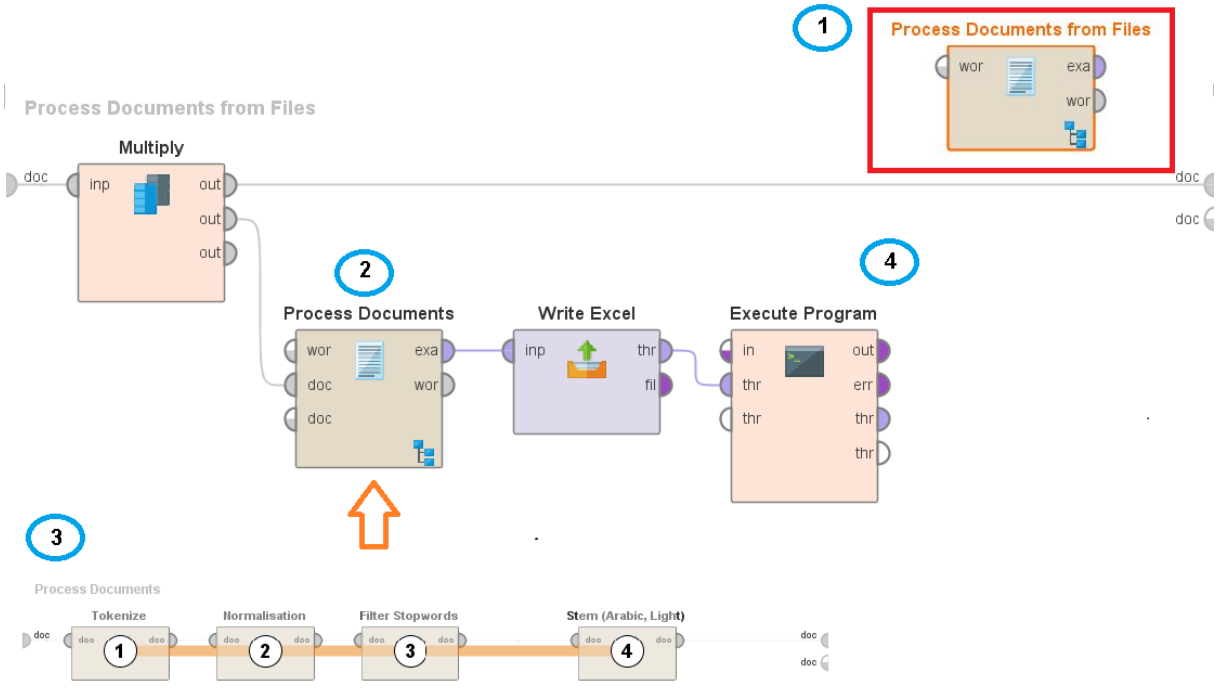


Figure IV.2 : *Processus de Prétraitement des documents*

1. Process Documents from Files: ce processus sert à lire tous les documents dans un dossier spécifié, avec Unicode utf-8, et délivre ces derniers un par un.

2. Chaque document délivré dans le port « doc » de 'Process Documents passe par des opérations de prétraitement, puis il sera sauvegardé dans un fichier Excel.

3. Le prétraitement consiste à appliquer à chaque document en entrée la Tokenisation, la Normalisation, l'Elimination des StopWords et le Stemming (Arabic light stemmer).



Figure IV.3 : *Exemple de Processus de Prétraitement*

Dans Notre Implémentation, on se base sur la racinisation légère (light stemming).

Lemmatisation assouplie (Light stemming) : La racinisation légère (ou light stemming en anglais) est un processus d'enlèvement de préfixes et/ou suffixes qui génère une pseudo-racine (stem en anglais), sans se préoccuper des infixes ou de reconnaître les schèmes (patterns en anglais) pour retrouver la racine (Larkey, Ballesteros, & E.Connell, 2007). Elle se base sur une liste prédéterminée contenant un ensemble de préfixes et suffixes censés être enlevés au cas où ils sont retrouvés attachés au mot dont on cherche la racine (**Tableau** III.1), Cette liste a été construite dans la plupart des cas en se basant sur des calculs statistiques et la connaissance de la langue arabe.

Préfixes	Suffixes
وال, فال, بال, بت, يت, لت, مت, وت, ست, نت, بم, لم, وم. كم, قم, ال, لل, وي, لي, في, وا, فا, لا ,با	ات, وا, ون, وه, ان, تي, ته, تم, كم, هم, هن, ها, ية, تك, نا, ين, يه,ة, ي, ا

Tableau IV.1 : *Liste des préfixes et suffixes dans la racinisation légère (Larkey, Ballesteros, & E.Connell, 2007)*

4. Une application java est exécutée à la fin de la création du fichier Excel pour créer un fichier texte prétraité relié au premier.



Figure IV.4 : *Exemple d'output de processus de Prétraitement*

IV.3.1.2 Extraction de List des termes

A partir du corpus prétraité, l'opération représentée par la figure ci-dessous génère un fichier Excel contenant la liste de tous les termes existants dans notre corpus.

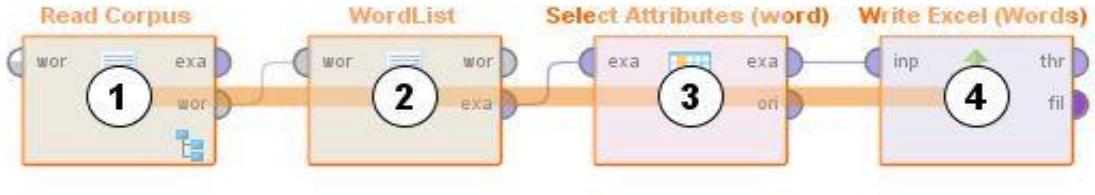


Figure IV.5 : *Processus d'Extraction de la liste des termes*

IV.3.1.3 Catégorisations des documents (documents clustering)

La catégorisation des documents nécessite une opération de clustering, elle sera appliquée sur l'ensemble des documents du corpus par la méthode « k-Means » avec K=7(par l'observation des documents de notre corpus), et la mesure de similarité (Euclidienne), sur une matrice de documents générée grâce au « Processus Read Corpus », basé sur TF-IDF.

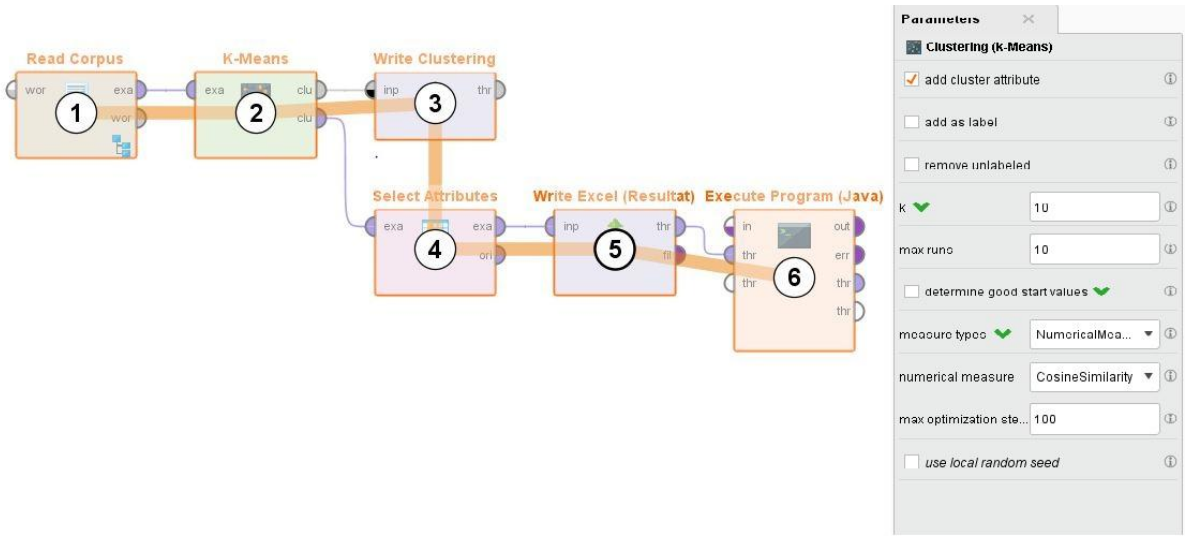


Figure IV.6 : *Processus de Catégorisation des documents*

A la fin de cette opération, un fichier Excel va être généré. Dans ce dernier, chaque document est affecté à sa propre catégorie (0-6). Une application java est exécutée pour allouer chaque document à un dossier qui signifie sa catégorie.

	A	B	C	D	E	F
1	metadata_file	metadata_	label	metadata_path	id	cluster
2	Doc0000001.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000001.txt	1.0	cluster_5
3	Doc0000002.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000002.txt	2.0	cluster_5
4	Doc0000003.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000003.txt	3.0	cluster_5
5	Doc0000004.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000004.txt	4.0	cluster_3
6	Doc0000005.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000005.txt	5.0	cluster_6
7	Doc0000006.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000006.txt	6.0	cluster_3
8	Doc0000007.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000007.txt	7.0	cluster_5
9	Doc0000008.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000008.txt	8.0	cluster_5
10	Doc0000009.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000009.txt	9.0	cluster_5
11	Doc0000010.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000010.txt	10.0	cluster_5
12	Doc0000011.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000011.txt	11.0	cluster_5
13	Doc0000012.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000012.txt	12.0	cluster_5
14	Doc0000013.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000013.txt	13.0	cluster_6
15	Doc0000014.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000014.txt	14.0	cluster_6
16	Doc0000015.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000015.txt	15.0	cluster_5
17	Doc0000016.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000016.txt	16.0	cluster_3
18	Doc0000017.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000017.txt	17.0	cluster_6
19	Doc0000018.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000018.txt	18.0	cluster_5
20	Doc0000019.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000019.txt	19.0	cluster_6
21	Doc0000020.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000020.txt	20.0	cluster_5
22	Doc0000021.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000021.txt	21.0	cluster_1
23	Doc0000022.txt	########	corpus	D:\Cluster_index\StreamCorpus\Doc0000022.txt	22.0	cluster_6

Figure IV.7 : *Exemple d'output de Processus de Catégorisation des documents.*

IV.3.1.4 Construction de matrix des termes

A partir de l'ensemble des documents de chaque catégorie, on applique un processus qui va générer une matrice des termes basée sur la fréquence de terme-document afin de préparer les données pour l'étape suivante.

La figure ci-dessous présente un processus RapidMiner qui sert à la définition de la fréquence de chaque terme dans les différentes catégories.



Figure IV.8 : *Processus de Construction de matrix des termes*

Le fichier Excel généré contient la liste des termes inclus dans chaque catégorie. Une application java sera exécutée pour affecter les poids à chaque vecteur de terme en se basant sur sa fréquence dans le document i ∈ [1, n] avec n : le nombre des documents de la catégorie concernée.

	A	B	C	D	E	F	G	H	I	J
1	in documents	total	in class (C0)	in class (C1)	in class (C2)	in class (C3)	in class (C4)	in class (C5)	in class (C6)	word
2	1	1	0	0	1	0	0	0	0	وكد
3	3	3	0	0	2	0	0	1	0	ومل
4	1	1	0	0	0	0	0	1	0	ونفث
5	1	1	0	0	1	0	0	0	0	ةلثلاثاء
6	4	4	0	2	0	0	0	2	0	ةلم
7	2	3	0	0	0	0	0	2	1	ةل
8	2	2	0	0	0	0	0	2	0	ةم
9	1	2	0	2	0	0	0	0	0	ةبدا
10	2	2	2	0	0	0	0	0	0	ةبس
11	26	41	2	22	9	0	5	1	2	اء
12	1	1	0	0	0	0	0	0	1	اوجل
13	8	8	2	1	1	2	0	2	0	اوكد
14	2	2	0	0	0	1	0	1	0	اوكل
15	5	5	0	1	1	0	1	0	2	اومن
16	3	3	0	1	1	1	0	0	0	اويد
17	129	454	403	1	14	0	1	14	21	ائتلاف
18	8	9	9	0	0	0	0	0	0	ائتلافا
19	1	1	1	0	0	0	0	0	0	ائتلافات
20	1	2	2	0	0	0	0	0	0	ائتلافهم
21	1	2	2	0	0	0	0	0	0	ائتلف
22	32	48	0	27	0	0	1	15	5	ائتم

Figure IV.9 : *Exemple de résultat de processus de construction de la matrix des termes*

IV.3.2 Java Création et exploitation de l'index

Afin de faire une comparaison entre notre approche d'indexation, et l'indexation classique basée sur le fichier inverse, on réalise deux types d'index, le premier est basé sur notre approche et sur l'output des phases précédentes, le second est développé à l'aide de l'outil Lucene (section IV.2.2.2 - 2) pour une indexation classique basée sur le modèle vectoriel et la pondération des termes TF-IDF.

IV.3.2.1. Terme Clustering

Notre but est d'affecter à chaque terme ses termes similaires, pour cela, un processus de 3 étapes est mis en œuvre :

IV.3.2.1.1. Création de la matrice de l'information mutuelle

Nous exploitons le résultat du processus précédent pour créer une matrice, dans laquelle chaque terme est associé avec un ensemble de poids qui représentent l'information mutuelle entre ce terme et les différents termes de la même catégorie.

Attribut...	وَمَل	ابقاء	ابكر	ابو	اتجاهنا	اتحاد	اتخاذ	اتخذت	اتفاق
وَمَل	0.811	0.123	0.123	0.123	0.123	0.811	0.811	0.123	0.811
ابقاء	0.123	0.811	0.811	0.123	0.811	0.811	0.123	0.123	0.123
ابكر	0.123	0.811	0.811	0.123	0.811	0.811	0.123	0.123	0.123
ابو	0.123	0.123	0.123	0.811	0.123	0.311	0.123	0.811	0.123
اتجاهنا	0.123	0.811	0.811	0.123	0.811	0.811	0.123	0.123	0.123
اتحاد	0.811	0.811	0.811	0.311	0.811	1.500	0.811	0.311	0.811
اتخاذ	0.811	0.123	0.123	0.123	0.123	0.811	0.811	0.123	0.811
اتخذت	0.123	0.123	0.123	0.811	0.123	0.311	0.123	0.811	0.123
اتفاق	0.811	0.123	0.123	0.123	0.123	0.811	0.811	0.123	0.811
انفض	0.811	0.123	0.123	0.123	0.123	0.811	0.811	0.123	0.811
اثار	0.811	0.123	0.123	0.123	0.123	0.811	0.811	0.123	0.811

Figure IV.10 : *Exemple de Matrice de Mutuelle Information*

Cette matrice et calculée à l'aide de l'outil JavaMI, et l'utilisation des vecteurs des termes comme entrée.

IV.3.2.1.2. Extraction des Termes Similaires

Pour chaque terme dans une catégorie, on extrait les 7 plus proches termes ayant les valeurs maximales de l'information mutuelle.

Cette opération va introduire une liste des concepts dans laquelle chaque terme est associé avec l'ensemble de ses termes similaires.

IV.3.2.1.3. Injection des termes similaires

Dans cette phase, on injecte chaque terme dans les documents contenants ses termes similaires pour retourner ces derniers dans l'étape de la recherche.

IV.3.2.2. Création de l'index (Cluster Index)

Notre index est de deux composants, index (Terme-Catégorie) qui sert à détecter la catégorie de la requête, et index (Terme-Cluster) qui sert à extraire les termes associés (similaires - même cluster) afin de capturer tout le sens voulus par la requête.

1. Index (Terme-Catégorie)

Par le principe discuté dans le chapitre précédent, une requête a un domaine spécifique. Cet Index présente le poids de chaqu'un des termes de notre corpus, dans chaque catégorie, par l'une des deux mesure déjà mentionnées.

La figure suivante représente une interface pour chercher la catégorie d'une requête, par le choix de la mesure (IDF inverse ou Centroïde), et permet aussi d'afficher la liste des termes et leur document associé dans une catégorie donnée.

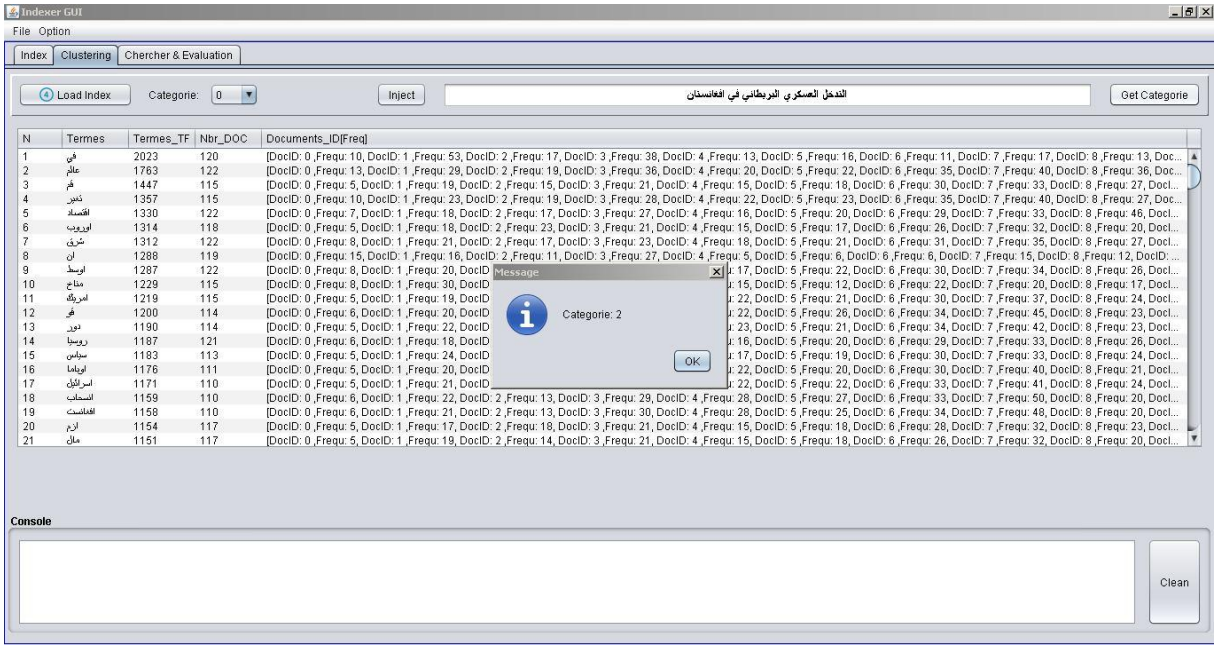


Figure IV.11 : *Interface de Détermination de la catégorie d'une requête*

2. Index (Terme-Cluster)

L'index Terme-Cluster est relié à la catégorie de la requête, en effet, à partir de la catégorie d'une requête, on accède à l'index Terme-Cluster pour définir l'ensemble des termes similaires au terme de notre requête, en pratique, tous les termes du même cluster sont considérés similaires.

La figure ci-dessous présente l'application de la recherche d'une requête. Dans un premier temps, le processus de recherche extrait la catégorie de la requête, par la suite, il se base sur le modèle vectoriel et la pondération TF-IDF pour le processus d'appariement document/requête, pour retourner la liste des documents répondants au besoin d'information lié aux termes de la requête et ses termes similaires.

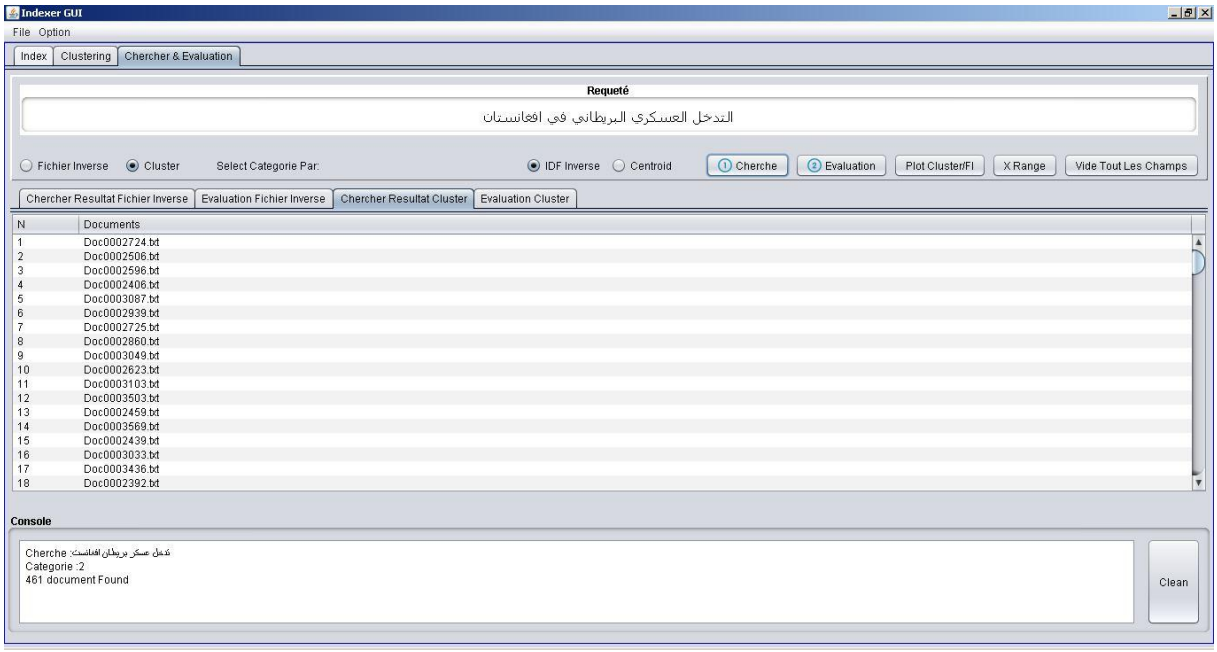


Figure IV.12 : *Interface de Recherche dans l'index base sur la clustering*

IV.3.2.2 Création de l'index (Lucene)

Toutes les requêtes de la recherche effectuée par Lucene se font sur un index. Il s'agit d'une collection de mots-clés et de propriétés identifiants des documents. Lors de la création d'un index, la bibliothèque crée un nouveau répertoire contenant plusieurs fichiers appelés des segments.

Les classes nécessaires à la création d'un nouvel index, ou à l'extension d'un index existant, se trouvent dans les paquetages org.apache.lucene.index et org.apache.lucene.analysis.standard. Le premier contient l'outil de création de l'index tandis que le second accueille les analyseurs de texte.

L'indexation de données met en œuvre 4 classes Lucene.

Index Writer	c'est la classe qui donne accès aux index en écriture (création, ajout de document, optimisation, …)
Analyzer	il s'agit d'un ensemble de classes qui ont pour but le découpage du texte en « token » (mot) et la normalisation du texte à indexer. Dons Notre implémentation le Corpus de texte est déjà traite grâce au RapidMiner
Document	Un Document représente une unité élémentaire d'information. Par exemple, indexer tous les fichiers Word d'un répertoire va ajouter dans l'index un Document Lucene par fichier. Ce sont des Documents qui sont retournés dans la liste de résultats d'une recherche. Comme cela a déjà été dit, un document est constitué de champs « Field » (nom / valeurs).
Field	Il s'agit d'un sous élément d'un document. Les champs les plus fréquents sont : titre, auteur, date de publication, url et bien sur le texte du fichier Word, PDF ou HTML.

Tableau IV.2 : *Listes des class d'indexation de Lucene*

A partir d'un index, notre application permet à l'utilisateur de saisir des requêtes pour lui retourner les résultats pertinents.

La figure ci-dessous représente le processus d'indexation par l'index Lucene. Dans un premier temps, on choisit le corpus à indexer (dans notre cas, le corpus est déjà traité par RapidMiner), en suite, on fait un appel à la classe « indexer.java » qui fait exécuter Lucene.

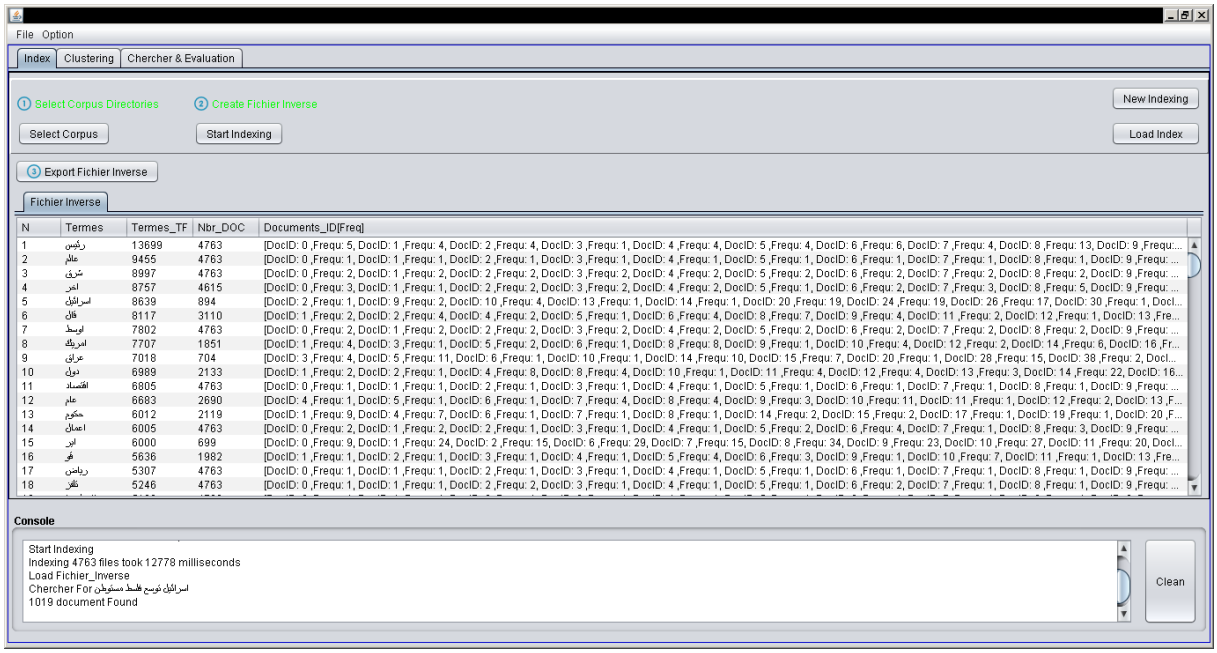


Figure IV.13 *Interface d'indexation par Lucene*

Les classes relatives à la recherche sont dans les paquetages org.apache.lucene.search et org.apache.lucene.queryParser. Avant tous, le développeur doit charger un index sur lequel les requêtes seront appliquées.

La recherche met en œuvre 6 classes Lucene.

IndexSearcher	c'est la classe qui donne accès aux index en recherche
Analyzer	Tout comme pour l'indexation les analyzer font partie du processus de recherche fin de normaliser les critères de recherche :
QuryParser	un parser de requête
Qury	un objet qui représente la requête de l'utilisateur et utilisé par un IndexSearcher.
Hits	Une collection d'éléments résultats de la recherche
Hit	Un élément de la collection des résultats
Document	Un document retrouvé et tel qu'il était lors de son ajout dans l'index (constitué des mêmes champs)

Tableau IV.3 : *Listes des class de Recherche de Lucene*

La figure ci-dessous représente le processus de recherche par l'index Lucene.

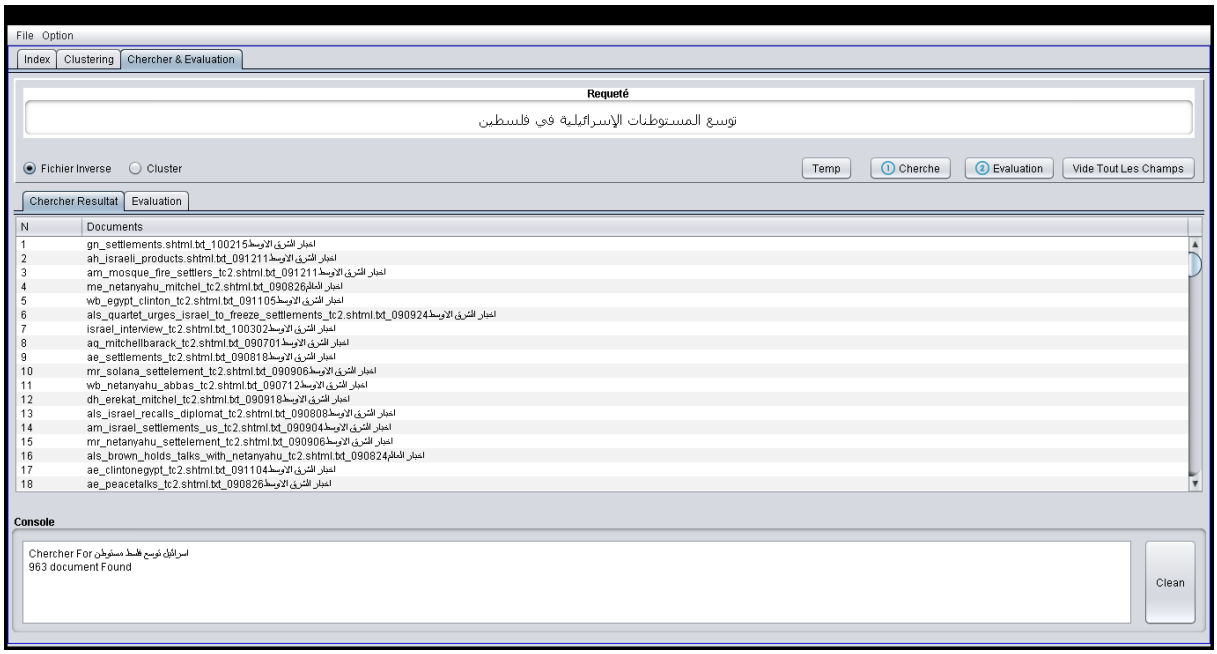


Figure IV.14 : *Interface de Recherche par Lucene*

IV.3.3 Interface d'évaluation

Afin d'effectuer une comparaison objective entre les deux index, une interface graphique est mise en place pour faire les différents calculs des mesures d'évaluation, et visualiser les résultats.

Une fois une recherche est effectuée, le bouton d'évaluation et activé pour permettre de sélectionner la liste des documents pertinents correspondante à la requête. Cette action lance une procédure de calcul pour les différents mesures d'évaluations (Précision, Rappel, F-Mesure, Précision Moyenne et Précision à n Pions), et établit le graphe Précision /Rappel), comme montré dans la figure ci-dessous.

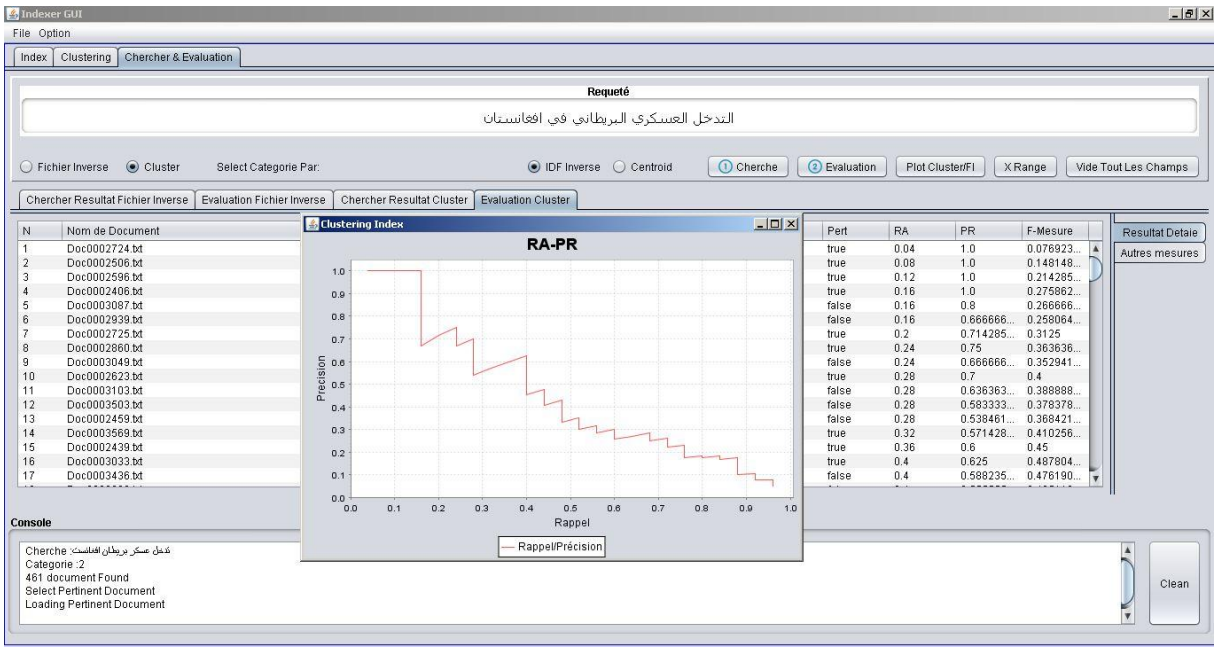


Figure IV.15 *: Interface d'Evaluation des Résultat de la Recherche*

Ces résultats sont les entrées de la procédure de test et d'évaluation dans la section suivante.

IV.4 Expérimentation

Dans cette section, nous détaillons les résultats obtenus après l'application de notre index générées à l'aide des méthodes de clustering, et nous évaluons lesquelles semblent avoir le meilleur potentiel, en fonction du rappel et de la précision par rapport à l'index classique.

IV.4.1 Contexte d'expérimentation

Notre critère principal pour trouver un environnement d'expérimentation était de disposer d'une grande base documentaire, avec des requêtes de test et leurs jugements de pertinence. En d'autres mots, nous utilisons une collection de test.

IV.4.2 Description de Corpus de Test

Pour une comparaison objective entre notre système et le système d'indexation classique basé sur le Fichier Inverse, nous utilisons le même corpus de texte arabe suivant :

BBC Arabic[1] : il est composé de 4763 documents textuels classés en sept catégories, et contenants 1, 860,786 millions de mots et 106, 733 mot-clé après la suppression des mots exclus (DR.Motaz, 2017).

Catégorie	Nombre de documents
أخبار الشرق الأوسط	2356
أخبار العالم	1489
اقتصاد و أعمال	296
رياضة	219
عرض الصحف	49
علوم و تكنولوجيا	232
منوعات	122

Tableau IV.4 : *Description de Corpus BBC Arabic*

IV.4.2 Méthodologie de test

IV.4.3 Test et Résultats

Nous avons choisi 25 requêtes de différentes catégories. Le tableau ci-dessous représente les requêtes utilisées pour le test et le nombre de documents pertinents pour chaque requête.

	Requetés	Nombre de documents pertinents
Q1	التدخل العسكري البريطاني في أفغانستان	25
Q2	نتائج وأحداث تصفيات كأس العالم 2010	45
Q3	العقوبات المفروضة على إيران بسبب النووي	61
Q4	قمة حول التغير المناخي في كوبنهاجن	18
Q5	الطعن في الانتخابات الرئاسية الارانية	75
Q6	المنتخب الجزائري لكرة القدم	31
Q7	ضحايا الهجمات و التفجيرات في العراق	34
Q8	اقتصاد الغاز و النفط في الشرق الاوسط	30

[1]http://www.bbc.com/arabic

Q9	نتائج الإنتخابات الرئاسية الإيرانية و عواقبها	40
Q10	توسع المستوطنات الإسرائيلية في فلسطين	34
Q11	المخططات الأوروبية لانقاذ اليونان من أزمة الديون المالية	31
Q12	العنف في مباريات كرة القدم العربية	30
Q13	تطوير محرك البحث جوجل لكسب مستخدمي الانترنيت و محاربة منافسيه	31
Q14	مباريات دوري أبطال أوروبا	37
Q15	مقتل الزعيم محسود من قبل الجيش الباكستاني	34
Q16	حركة التمرد في السودان	43
Q17	مناطق اكتشاف النفط و تطور أسعاره	32
Q18	حركة حماس الفلسطينية في قطاع غزة	41
Q19	التدخلات السياسية لباراك أوباما في العالم	106
Q20	إنسحاب القوات الأمريكية من العراق	35
Q21	سياسة دبي لتجاوز الأزمة الاقتصادية	38
Q22	الزلزال الذي ضرب هـايتي	23
Q23	احتفالات بذكرى سقوط جدار برلين	17
Q24	أزمة اللاجئين بدارفور	57
Q25	الهجوم الارهابي في روسيا	18

Tableau IV.5 *: Liste des Requêtes de test avec leurs documents pertinents*

Chaque requête passe par 3 étapes (prétraitement, sélection de catégorie, extraction des termes similaires)

- Prétraitement de la requête

Pour avoir une représentation de la requête adéquate à celle de l'index, le même processus de prétraitement appliqué aux documents de corpus sera appliqué à la requête, le tableau suivant représente le prétraitement de chaque requête.

	Requetés	Requête prétraité
Q1	التدخل العسكري البريطاني في أفغانستان	تدخل عسكر بريطان افغانست
Q2	نتائج وأحداث تصفيات كأس العالم 2010	نتائج احداث تصف كاس عالم 2010
Q3	العقوبات المفروضة على إيران بسبب النووي	عقوب مفروض ايرنوو
Q4	قمة حول التغير المناخي في كوبنهاجن	قم تغير مناخ كوبنهاجن
Q5	الطعن في الانتخابات الرئاسية الارانية	طعن انتخاب رئاس اران
Q6	المنتخب الجزائري لكرة القدم	منتخب جزائر لكر قدم
Q7	ضحايا الهجمات و التفجيرات في العراق	ضحايا هجم تفجير عراق
Q8	اقتصاد الغاز و النفط في الشرق الاوسط	اقتصاد غاز نفط شرق اوسط
Q9	نتائج الإنتخابات الرئاسية الإيرانية و عواقبها	نتائج انتخاب رئاس ايران عواقب
Q10	توسع المستوطنات الإسرائيلية في فلسطين	توسع مستوطن اسرائيل فلسط
Q11	المخططات الأوروبية لانقاذ اليونان من أزمة الديون المالية	مخطط اوروب لانقاذ يون ازم دين مال
Q12	العنف في مباريات كرة القدم العربية	عنف مبار كر قدم عرب
Q13	تطوير محرك البحث جوجل لكسب مستخدمي الانترنيت و محاربة منافسيه	تطوير محرك بحث جوجل لكسب مستخدم انترنيت محارب منافس
Q14	مباريات دوري أبطال أوروبا	مبار دور ابطال اوروبا
Q15	مقتل الزعيم محسود من قبل الجيش الباكستاني	مقتل زعيم محسود جيش باكستان
Q16	حركة التمرد في السودان	حرك تمرد سودان
Q17	مناطق اكتشاف النفط و تطور أسعاره	مناطق اكتشاف نفط تطور اسعار
Q18	حركة حماس الفلسطينية في قطاع غزة	حرك حماس فلسطين قطاع غز
Q19	التدخلات السياسية لباراك أوباما في العالم	تدخل سياس لباراك اوباما عالم
Q20	إنسحاب القوات الأمريكية من العراق	انسحاب قو امريك عراق

Q21	سياسة دبي لتجاوز الأزمة الاقتصادية	سياس دبي لتجاوز ازم اقتصاد
Q22	الزلزال الذي ضرب هـايتي	زلزال ضرب هايت
Q23	احتفالات بذكرى سقوط جدار برلين	احتفال بذكر سقوط جدار برل
Q24	أزمة اللاجئين بدارفور	ازم لاجئ بدارفور
Q25	الهجوم الارهابي في روسيا	هجوم ارهاب روسيا

Tableau IV.6 *: Liste des Requetés prétraite*

- Sélection de la catégorie de la requête

Chaque requête correspond à une catégorie dans laquelle on effectue la recherche et on extrait les termes similaires de ceux de la requête.

La table ci-dessous représente le résultat obtenu des deux méthodes de choix de la catégorie d'une requête.

	Requetés	**Catégorie**	
		IDF Inverse	Centroïde
Q1	التدخل العسكري البريطاني في أفغانستان	2	2
Q2	نتائج وأحداث تصفيات كأس العالم 2010	4	6
Q3	العقوبات المفروضة على إيران بسبب النووي	3	5
Q4	قمة حول التغير المناخي في كوبنهاجن	5	0
Q5	الطعن في الانتخابات الرئاسية الارانية	3	5
Q6	المنتخب الجزائري لكرة القدم	4	6
Q7	ضحايا الهجمات و التفجيرات في العراق	0	3
Q8	اقتصاد الغاز و النفط في الشرق الاوسط	5	3
Q9	نتائج الإنتخابات الرئاسية الإيرانية و عواقبها	3	5
Q10	توسع المستوطنات الإسرائيلية في فلسطين	6	1
Q11	المخططات الأوروبية لانقاذ اليونان من أزمة الديون المالية	5	0
Q12	العنف في مباريات كرة القدم العربية	4	6
Q13	تطوير محرك البحث جوجل لكسب مستخدمي الانترنيت و محاربة منافسيه	1	6
Q14	مباريات دوري أبطال أوروبا	4	6
Q15	مقتل الزعيم محسود من قبل الجيش الباكستاني	2	2
Q16	حركة التمرد في السودان	0	6
Q17	مناطق اكتشاف النفط و تطور أسعاره	2	3
Q18	حركة حماس الفلسطينية في قطاع غزة	4	1
Q19	التدخلات السياسية لباراك أوباما في العالم	3	6
Q20	إنسحاب القوات الأمريكية من العراق	2	3
Q21	سياسة دبي لتجاوز الأزمة الاقتصادية	3	6
Q22	الزلزال الذي ضرب هـايتي	1	6
Q23	احتفالات بذكرى سقوط جدار برلين	4	1
Q24	أزمة اللاجئين بدارفور	0	6
Q25	الهجوم الارهابي في روسيا	3	5

Tableau IV.7 *: Résultat de la sélection de la Catégorie des requêtes par les deux mesures*

- Extraction des termes similaires

Pour chaque terme dans notre corpus, on extrait ses termes similaires pour effectuer une recherche permettant de capturer tous ses sens.

La recherche d'un terme par l'utilisation de ces termes similaires retourne en résultat, tous les documents contenants ce terme ou un de ses termes similaires.

Le table ci-dessous présentent un exemple des termes avec leurs termes similaires en utilisant dans les différents catégorie :

Terme	Terme similaires							Catégorie
هجوم	بغداد	انفجار	شخصا	نفجير	عراق	استهداف	مقتل	0
	قراصن	تعرض	الكثيرون	كثيرون	مستخدم	بريد	قرصن	1
	انتحار	مقتل	مسلح	باكستان	هجم	امريك	قتل	2
	تفجير	حرس	ايران	ثور	سيست	قائد	بلوشست	3
	ثان	مصر	حارس	دقيق	شوط	مبارا	شباك	4
	صين	حزب	دول	رئيس	مركز	اول	انطلاق	5
	حرب	تقرير	قطاع	صواريخ	اسرائيل	عام	فلسطين	6
جوجل Google	اشور	اتجار	اثار	بثمن	استخدما	اتصور	اريك	0
	شركة	بحث	انترنت	محرك	تكنولوجيا	علوم	الكثرون	1
	/	/	/	/	/	/	/	2
	انترنت	معارض	الكثيرون	ميل	معلوم	بورما	خدم	3
	/	/	/	/	/	/	/	4
	انترنت	موقع	شركة	محرك	بحث	رقاب	صين	5
	/	/	/	/	/	/	/	6
روسيا	خلال	اوسط	اير	يرطانيا	عراق	عالم	ضخم	0
	نهر	موسكو	انجلترا	صين	عواصف	دول	ارتياكا	1
	محطت	روس	قوقاز	قطار	ميترو	موسكو	شيش	2
	نوو	اير	يورانوم	فرنسا	روس	دول	موسكو	3
	تجار	شرق	موسكو	برتغال	نيجيريا	اوسط	تاهلت	4
	روس	يوت	غاز	دول	امداد	صين	موسكو	5
	بان	رئيس	نتنياهو	متحد	دول	اسرائيل	روس	6

Tableau IV.8 *: Exemple d'extraction des termes similaire par catégorie*

Les différents termes similaires parfois n'ont aucun sens, car ils sont dans la forme prétraitée (après le stemming). Par exemple «ايرا» qui va dire « ايران»

On remarque par la suite qu'un terme à plusieurs termes associe par rapport à la catégorie, si on prend le terme « هجم » dans la catégorie 0 qui signifie une action terroriste, et dans la catégorie 4 il signifie une action reliée au sport (football).

Après le calcul des valeurs de précision et du rappel à chaque restitution d'un document par le SRI, nous déduisons les résultats ci-après.

Le tableau suivant montre les valeurs des précisions au rang x pour les quatre requêtes ou x ∈{5, 10, 15, 20}, et la précision moyenne (Average Precision).

On effectue une comparaison entre 6 types d'index :

1- Index Classique : qui est également l'index le plus simple dans le domaine de la RI, il se base sur le fichier inverse pour retourner les différents résultats de la recherche (FI) ;

2- Index Classique avec l'utilisation des termes similaires : différent du premier index par l'utilisation de la notion des termes similaires (FI_S) ;

3- Index par catégorie sans l'utilisation des termes similaires et par la mesure IDF Inversée pour la sélection de la catégorie pour la requête (C_NS_IDF) ;

4- Index par catégorie sans l'utilisation des termes similaires et par la mesure des Centroïdes pour la sélection de la Catégorie pour la requête (C_NS_CENT) ;

5- Index par catégorie avec l'utilisation des termes similaires et par la mesure IDF Inversée pour la sélection de la Catégorie pour la requête (C_S_IDF) ;

6- Index par catégorie sans l'utilisation des termes similaires et par la mesure des Centroïdes pour la sélection de la Catégorie pour la requête (C_S_CENT).

	Précision au Rang X	Index Classique	Index Classique avec l'utilisation des termes similaire	Index par catégorie sans l'utilisation des termes similaires et par la mesure IDF Inversée pour la sélection de la Catégorie pour la requête	Index par catégorie sans l'utilisation des termes similaires et par la mesure des Centroïdes pour la sélection de la Catégorie pour la requête	Index par catégorie avec l'utilisation des termes similaire et par le mesure IDF Inverse pour la sélectionne de la Catégorie pour la requête	Index par catégorie sans l'utilisation des termes similaires et par la mesure des Centroïdes pour la sélection de la Catégorie pour la requête
Q1	P5	0.200	0.200	0.200	0.200	0.200	0.200
	P10	0.200	0.200	0.100	0.100	0.100	0.100
	P15	0.133	0.200	0.067	0.067	0.067	0.067
	P20	0.150	0.200	0.050	0.050	0.050	0.050
	Average Precision	**0.160**	**0.196**	**0.146**	**0.146**	**0.147**	**0.147**
Q2	P5	0.800	0.800	0.000	0.800	0.000	0.600
	P10	0.600	0.600	0.000	0.600	0.000	0.600
	P15	0.533	0.533	0.000	0.533	0.000	0.533
	P20	0.600	0.550	0.000	0.600	0.000	0.450
	Average Precision	**0.275**	**0.318**	**0.000**	**0.297**	**0.000**	**0.249**
Q3	P5	0.200	1.000	0.000	0.400	0.000	0.600
	P10	0.200	0.800	0.000	0.200	0.000	0.300
	P15	0.400	0.600	0.000	0.333	0.000	0.333
	P20	0.500	0.500	0.000	0.500	0.000	0.350

	Average Precision	**0.573**	**0.560**	**0.000**	**0.603**	**0.000**	**0.518**
Q4	**P5**	0.200	0.200	0.000	0.000	0.000	0.200
	P10	0.200	0.200	0.000	0.200	0.000	0.200
	P15	0.200	0.133	0.000	0.200	0.000	0.200
	P20	0.250	0.250	0.000	0.250	0.000	0.200
	Average Precision	**0.218**	**0.215**	**0.000**	**0.207**	**0.000**	**0.212**
Q5	**P5**	0.400	0.400	0.000	0.800	0.000	0.800
	P10	0.300	0.300	0.000	0.800	0.000	0.800
	P15	0.200	0.200	0.000	0.667	0.000	0.733
	P20	0.200	0.150	0.000	0.750	0.000	0.750
	Average Precision	**0.180**	**0.206**	**0.000**	**0.712**	**0.000**	**0.727**
Q6	**P5**	0.400	0.600	0.000	0.400	0.000	0.400
	P10	0.200	0.300	0.000	0.200	0.000	0.200
	P15	0.267	0.267	0.000	0.267	0.000	0.267
	P20	0.400	0.400	0.000	0.400	0.000	0.400
	Average Precision	**0.414**	**0.439**	**0.000**	**0.418**	**0.000**	**0.428**
Q7	**P5**	0.600	0.600	0.000	0.600	0.000	0.200
	P10	0.500	0.600	0.000	0.500	0.000	0.200
	P15	0.467	0.467	0.000	0.533	0.000	0.267
	P20	0.450	0.500	0.000	0.400	0.000	0.250
	Average Precision	**0.285**	**0.321**	**0.000**	**0.318**	**0.000**	**0.187**
Q8	**P5**	0.200	0.200	0.000	0.200	0.000	0.200
	P10	0.200	0.200	0.000	0.300	0.000	0.300
	P15	0.133	0.133	0.000	0.400	0.000	0.400
	P20	0.150	0.200	0.000	0.400	0.000	0.350
	Average Precision	**0.288**	**0.288**	**0.005**	**0.361**	**0.005**	**0.334**
Q9	**P5**	0.200	0.200	0.000	0.400	0.000	0.200
	P10	0.500	0.300	0.000	0.600	0.000	0.600
	P15	0.467	0.400	0.000	0.667	0.000	0.600
	P20	0.550	0.450	0.000	0.600	0.000	0.600
	Average Precision	**0.438**	**0.305**	**0.000**	**0.530**	**0.000**	**0.515**
Q10	**P5**	0.000	0.400	0.200	0.000	0.000	0.200
	P10	0.300	0.200	0.100	0.200	0.000	0.100
	P15	0.333	0.133	0.067	0.200	0.000	0.133

	P20	0.250	0.150	0.050	0.250	0.000	0.200
	Average Precision	0.161	0.183	1.000	0.140	0.006	0.140
Q11	P5	0.800	0.800	0.000	0.800	0.000	0.800
	P10	0.900	0.800	0.000	0.700	0.000	0.800
	P15	0.600	0.667	0.000	0.533	0.000	0.600
	P20	0.600	0.600	0.000	0.550	0.000	0.550
	Average Precision	0.660	0.653	0.000	0.722	0.000	0.740
Q12	P5	0.200	0.200	0.000	0.400	0.000	0.200
	P10	0.300	0.400	0.000	0.300	0.000	0.200
	P15	0.533	0.600	0.000	0.533	0.000	0.133
	P20	0.650	0.650	0.000	0.650	0.000	0.150
	Average Precision	0.321	0.339	0.000	0.338	0.000	0.179
Q13	P5	0.800	0.800	0.000	0.800	0.000	0.800
	P10	0.600	0.600	0.000	0.600	0.000	0.600
	P15	0.600	0.667	0.000	0.600	0.000	0.600
	P20	0.600	0.700	0.000	0.600	0.000	0.600
	Average Precision	0.598	0.610	0.000	0.604	0.000	0.571
Q14	P5	1.000	1.000	0.000	1.000	0.000	1.000
	P10	0.700	0.700	0.000	0.700	0.000	0.800
	P15	0.733	0.733	0.000	0.733	0.000	0.800
	P20	0.650	0.650	0.000	0.650	0.000	0.700
	Average Precision	0.741	0.741	0.000	0.751	0.000	0.757
Q15	P5	0.400	0.400	0.400	0.400	0.400	0.400
	P10	0.400	0.400	0.400	0.400	0.400	0.400
	P15	0.533	0.533	0.533	0.533	0.533	0.533
	P20	0.600	0.600	0.600	0.600	0.600	0.600
	Average Precision	0.545	0.545	0.545	0.545	0.545	0.545
Q16	P5	0.800	0.800	0.000	0.800	0.000	1.000
	P10	0.600	0.600	0.000	0.700	0.000	0.800
	P15	0.600	0.600	0.000	0.667	0.000	0.667
	P20	0.500	0.500	0.000	0.550	0.000	0.700
	Average Precision	0.382	0.382	0.000	0.483	0.000	0.569
Q17	P5	0.800	1.000	0.000	1.000	0.000	0.200
	P10	0.900	0.900	0.000	0.900	0.000	0.100

	P15	0.733	0.733	0.000	0.800	0.000	0.133
	P20	0.600	0.600	0.000	0.600	0.000	0.150
	Average Precision	**0.508**	**0.530**	**0.000**	**0.686**	**0.000**	**0.178**
Q18	P5	0.200	0.400	0.600	0.200	0.600	0.200
	P10	0.400	0.300	0.500	0.300	0.300	0.300
	P15	0.400	0.267	0.333	0.400	0.333	0.400
	P20	0.350	0.300	0.300	0.350	0.350	0.400
	Average Precision	**0.314**	**0.295**	**0.415**	**0.319**	**0.521**	**0.329**
Q19	P5	0.600	0.600	0.000	0.400	0.000	0.600
	P10	0.300	0.300	0.000	0.300	0.100	0.300
	P15	0.267	0.200	0.000	0.400	0.067	0.400
	P20	0.250	0.250	0.000	0.450	0.050	0.500
	Average Precision	**0.266**	**0.268**	**0.092**	**0.417**	**0.102**	**0.437**
Q20	P5	0.800	0.800	0.000	0.800	0.000	0.800
	P10	0.700	0.800	0.000	0.700	0.000	0.700
	P15	0.600	0.733	0.000	0.533	0.000	0.600
	P20	0.650	0.600	0.000	0.600	0.000	0.600
	Average Precision	**0.667**	**0.644**	**0.040**	**0.663**	**0.039**	**0.640**
Q21	P5	0.800	0.800	0.000	0.800	0.000	1.000
	P10	0.800	0.800	0.000	0.800	0.000	0.800
	P15	0.800	0.800	0.000	0.733	0.000	0.667
	P20	0.750	0.750	0.000	0.750	0.000	0.700
	Average Precision	**0.747**	**0.747**	**0.000**	**0.744**	**0.000**	**0.742**
Q22	P5	0.800	0.800	0.000	0.800	0.000	0.800
	P10	0.900	0.900	0.000	0.900	0.000	0.900
	P15	0.933	0.933	0.000	0.933	0.000	0.933
	P20	0.900	0.900	0.000	0.900	0.000	0.900
	Average Precision	**0.876**	**0.875**	**0.000**	**0.871**	**0.000**	**0.882**
Q23	P5	1.000	1.000	0.000	0.400	0.000	0.400
	P10	1.000	1.000	0.000	0.200	0.000	0.200
	P15	0.800	0.733	0.000	0.133	0.000	0.133
	P20	0.600	0.700	0.000	0.100	0.000	0.100
	Average Precision	**0.558**	**0.593**	**0.000**	**0.583**	**0.000**	**0.583**
Q24	P5	0.200	0.200	0.000	0.200	0.000	0.400

	P10	0.100	0.100	0.000	0.100	0.000	0.200
	P15	0.200	0.200	0.000	0.133	0.000	0.267
	P20	0.200	0.200	0.000	0.100	0.000	0.300
	Average Precision	**0.303**	**0.306**	**0.000**	**0.337**	**0.000**	**0.175**
Q25	P5	0.400	0.400	0.000	0.600	0.000	0.800
	P10	0.400	0.400	0.000	0.500	0.000	0.700
	P15	0.400	0.400	0.000	0.467	0.000	0.467
	P20	0.350	0.350	0.000	0.450	0.000	0.450
	Average Precision	**0.273**	**0.278**	**0.000**	**0.391**	**0.000**	**0.461**

Tableau IV.9 *: Les Résultat du processus d'évaluation pour les requêtes de test*

Les résultats de table précédente peuvent être illustrés par les figures suivantes :

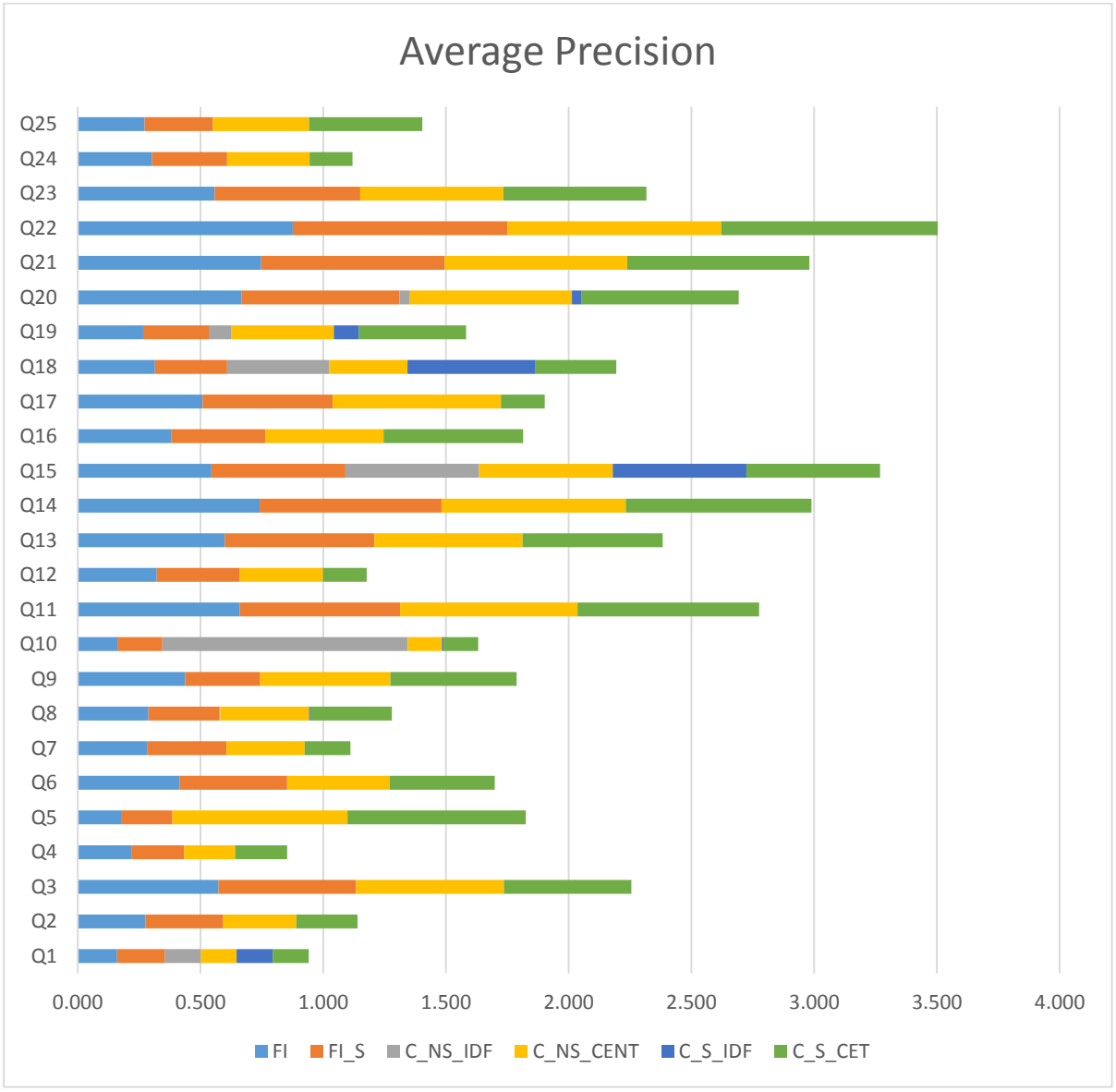


Figure IV.16 : *La Valeur de Precision Moyenne par Requête*

Pour bien comparée les index, une mesure plus générale est utilisée pour déterminer le meilleur index, on parle ici de MAP (Mean Average Precision), qui permet de mesurer les performances du SRI utilisé, par la moyenne des précisions de l'ensemble de requêtes.

Average Precision au Rang X	Index Classique	Index Classique avec l'utilisation des termes similaire	Index par catégorie sans l'utilisation des termes similaire et par le mesure IDF Inverse pour la sélectionne de la Catégorie pour la requête	Index par catégorie sans l'utilisation des termes similaire et par le mesure des Centroïde pour la sélectionne de la Catégorie pour la requête	Index par catégorie avec l'utilisation des termes similaire et par le mesure IDF Inverse pour la sélectionne de la Catégorie pour la requête	Index par catégorie sans l'utilisation des termes similaire et par le mesure des Centroïde pour la sélectionne de la Catégorie pour la requête
P5	0.512	0.584	0.056	0.528	0.048	0.520
P10	0.488	0.508	0.044	0.472	0.036	0.448
P15	0.475	0.475	0.040	0.480	0.040	0.435
P20	0.470	0.468	0.040	0.484	0.042	0.440
Mean Avrage Precision (MAP)	**0.430**	**0.433**	**0.090**	**0.487**	**0.055**	**0.450**

Tableau IV.10 *: Les Résultat d'Evaluations par le mesure Mean Avrage Precision*

La figure suivante traduit les résultats de table précédente :

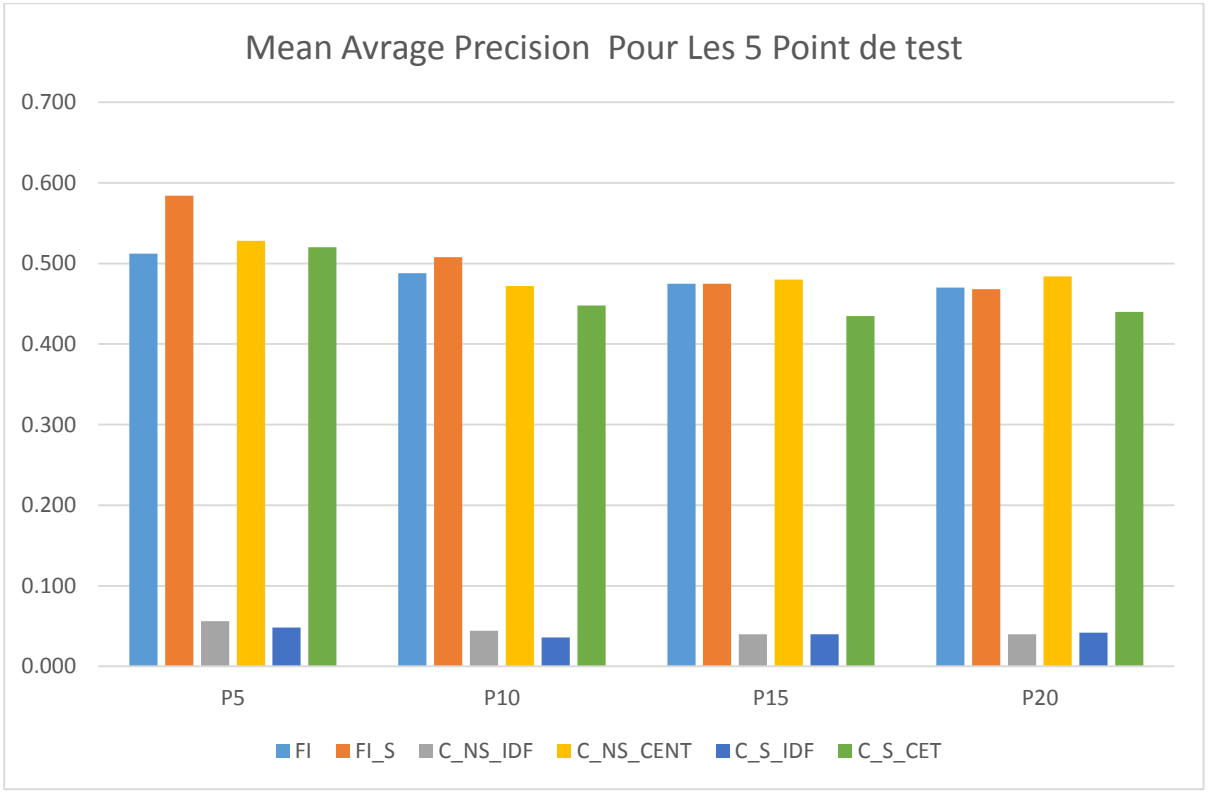


Figure IV.17 *: Les valeurs de Mean Avrage Precision pour tous les index dons les 5 point*

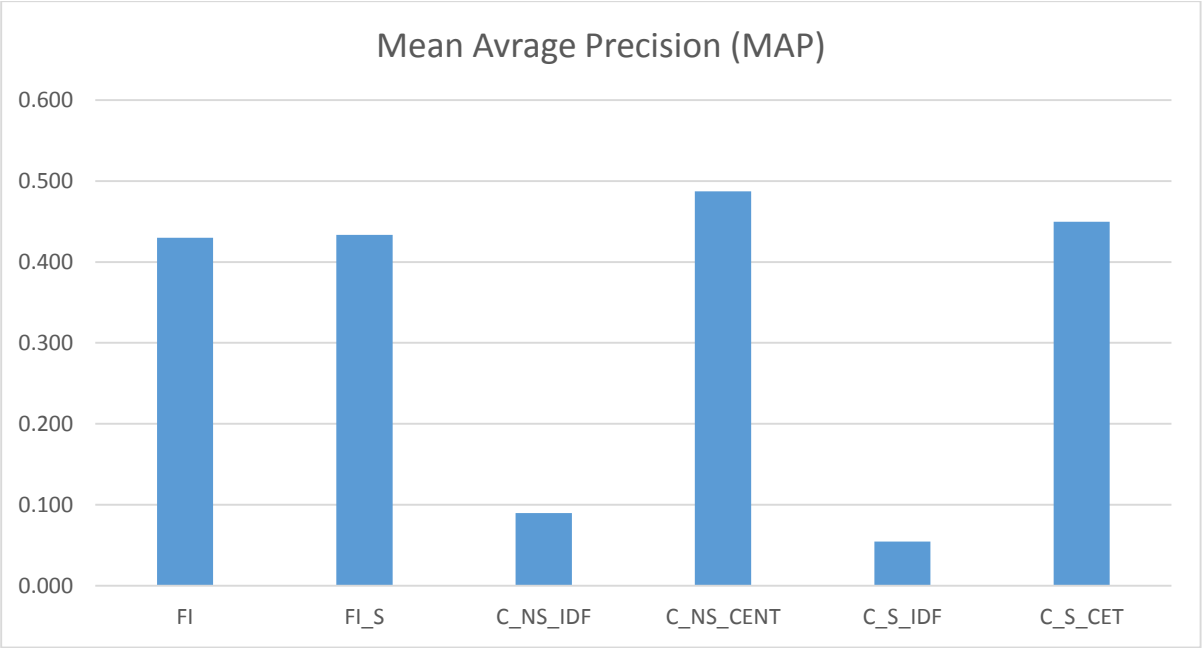


Figure IV.18 : *Les valeurs de Mean Avrage Precision pour tous les index*

Enfin en élaborée la table suivante pour illustrée l'augmentation des performances de système de recherche d'information, par l'utilisation des différents index par rapport a l'index classique :

	Index Classique avec l'utilisation des termes similaire	Index par catégorie sans l'utilisation des termes similaire et par le mesure IDF Inverse pour la sélectionne de la Catégorie pour la requête	Index par catégorie sans l'utilisation des termes similaire et par le mesure des Centroïde pour la sélectionne de la Catégorie pour la requête	Index par catégorie avec l'utilisation des termes similaire et par le mesure IDF Inverse pour la sélectionne de la Catégorie pour la requête	Index par catégorie sans l'utilisation des termes similaire et par le mesure des Centroïde pour la sélectionne de la Catégorie pour la requête
P5	14.06%	-89.06%	3.13%	-90.63%	1.56%
P10	4.10%	-90.98%	-3.28%	-92.62%	-8.20%
P15	0.00%	-91.57%	1.12%	-91.57%	-8.43%
P20	-0.43%	-91.49%	2.98%	-91.06%	-6.38%
Mean Avrage Precision (MAP)	**0.77%**	**-79.14%**	**13.33%**	**-87.30%**	**4.58%**

Tableau IV.11 : *Augmentation des performances de système de recherche d'information*

IV.4.4. Interprétation

A travers les tests que nous avons effectués, nous remarquons que

Pour les résultats obtenus à partir des deux index classiques : l'index qui est basé sur la Notion des termes similaires a introduit un résultat qui augmente la performance du système, celle-ci est due à l'utilisation des termes similaire.

la méthode d'indexation par catégorisation sans exploiter les termes similaires extraits du corpus à l'aide des méthodes d'apprentissage non supervisée, et qui utilise la mesure basée sur la sélection de la catégorie associé à la requête par l'exploitation des centroïdes, augmente globalement la performance du système de recherche d'information. Cela est remarquable à partir de la valeur de la mesure d'évaluation MAP.

De même, nous remarquons que si on applique la notion des termes similaires sur cet index, il augmente aussi la performance du SRI, en se basant toujours sur la même mesure de sélection de la catégorie de la requête, bien que la non-utilisation de ces termes similaires donne des résultats plus efficaces.

La table suivante représentée le nombre des documents dans chaque catégories :

Catégories	Nombre des documents
0	122
1	596
2	461
3	442
4	171
5	327
6	2644

Tableau IV.12 *: Nombre des documents par catégorie*

Par l'observation des requêtes de l'index C_S_CENT qui augmenté la précision moyenne par rapport au index C_NS_CENT qui sont les suivant :

«Q1, Q4, Q5, Q6, Q11, Q12, Q14, Q16, Q18, Q19 » 20% sont de la catégorie 0,10% de la catégorie 1, 20% de la catégorie 2, 10% de la catégorie 3, et 40% de la catégorie 6. En comparent cette résultat dans le table précédant en remarquent que le nombre des documents dans la catégorie joués un roule important dans la sélection des termes similaires.

Pour les index basés sur la mesure IDF Inversée, les résultats sont très mauvais, ceci est dû principalement à cette mesure qui a diminué la performance de la recherche.

Enfin on peut Déduire les résultats suivants :

- l'indexation basée sur la similarité des termes augmente la performance des systèmes de recherche d'information ;
- La sélection des termes similaire par la mesure MI dépendant de la taille de corpus ;
- l'indexation basée sur la catégorisation des documents augmente la performance des systèmes de recherche d'information.
- Le choix de la catégorie d'une requête lors de la recherche est une action très importante, et qui Contrôle le cours de ce processus de façon très significative.

- L’échec de la sélection des termes similaires lors de son utilisation dans le processus d’indexation peut causer une forte baisse dans la performance du system de recherche d’information.
- L’indexation Classique ne répond pas à toutes les questions qui concernent la requête, sa catégorie, et, ou le sens voulu par le besoin d'informations qui n’est pas exprimé par la requête.

L’échec de l’utilisation des termes similaire peut être causé par :

- Le fait de ne pas pouvoir retourner les bons termes similaires ;
- L’échec de mesure de similarité (information mutuelle) ;
- Une taille très petite du corpus.

Une solution proposée pour ces échecs peut être :

- Utilisée un corpus d’une grand taille ;
- Si une nouvelle ressource externe de la langue arabe a été créée, et qui ne souffre pas de manque des ressources existants, en peut l’utilisée pour aide à la discision des termes similaire lors de sont sélection ;

IV.5 Conclusion

Dans ce Chapitre, nous avons présenté l'implémentation de notre approche d'indexation automatique, en utilisant le langage Java et des outils de datamining (RapidMiner), et d'autres outils qui aident à la réalisation de cette approche. Au cours de ce chapitre nous avons effectué une série de tests sur l'ensemble des requêtes textuelles concernant notre corpus, Enfin nous nous sommes focalisés sur l'impact des diffèrent index implémenté sur le système de recherche d'information arabe.

Les résultats expérimentaux montrent que les performances du système de recherche d'information ont été améliorées de façon significative, lors de l'utilisation de l'indexation base sur la catégorisation des documents sans l'utilisation des termes similaires, et lors d'utilisation des termes similaires sans l'utilisation de la catégorisation des documents par rapport au cas d'index classique.

CONCLUSION GENERALE ET PERSPECTIVES

CONCLUSION GENERALE ET PERSPECTIVES

Le travail présenté dans ce mémoire s'inscritent dans le contexte de la recherche d'information (RI) arabe, Notre travail vise à prendre en compte les éléments de contexte dans les processus de RI. Dans cette optique, les problématiques générale auxquelles nous avons proposé des solutions ont été :

- prise en compte du sens des termes lors de son indexation ;
- prise en compte du domaine d'un terme lors de l'extraction du sens;
- pris en compte de particularité de la langue arabe lors de son indexation automatique ;
- prise en compte du manque de ressources électroniques de la langue arabe.

Suivant cette orientation, notre contribution a consisté de définir une nouvelle indexation qui supporte une RI statique / sémantique basée sur une description des termes par leur termes similaire, et qui utilise également, la catégorisation du texte afin de capturer le sens des différents termes dans leur domain d'utilisation, et qui prend en considération le manque des ressources arabes dans le domaine de la recherche d'information par l'application des methodes d'apprentissages non supervisées, et en fin adapte la particularité de langue arabe lors de l'extraction des termes et le traitment des documents a indexés.

Notre approche essaie de prendre en considération des termes similaires, et d'appliquer la catégorisation des documents afin de capturer les différents sens d'un terme dans les diffférnets domaines d'utilisation de cette dernière, face a un obstacle inattendu parfois dans le processus de la sélection des termes similaires qui fait une réduction dans les performances du système, cet échec est lié à l'un des causes suivantes :

- La petite taille du corpus de test

Notre corpus de test se compose de 4763 documents, ce corpus de taille limité ne donne pas une vraie information utile lors de l'utilisation de la notion des co-occurrences, et dans une cette partie, l'information mutuelle n'est pas considèrée une information complette sur les termes pour avoir créé un lien d'aide, qui permet de prendre une décision sur la seléction des termes similaires.

- Le choix de mesure de similairité entre les termes

La mesure utilisée pour le choix des termes similaires (mutuelle information) n'a pas donné des résultats satisfaits pour tous les termes, parfois les termes qui sont trés fréquents dans le corpus sont sélectionnés sans avoir une vrai relation au termes racine.

Les résultats obtenus par notre approche montre une amélioration significative de la précision. Nous avons conclu à travers les expérimentations qu'on l'utilise des termes similaires lors de l'indexation et la catégorisation des documents, permet de mieux répondre aux besoins de l'utilisateur, et ainsi d'améliorer en général les performances du système.

Prespectives

Ce travail ouvre une voie vers diverses prespectives qui se situent sur les plans suivants :

- Tester l'approche (index) proposé sur plusieurs collections volumineuses, Afin de valider nos propositions à grande échelle.
- Étudier plus précisément les mesures de mise en correspondance entre les termes et l'extraction des liens entre les termes par rapport à leur présence et absence dans la collection des documents.
- Améliorer le processus de catégorisation des documents par le choix des algorithmes de clustering plus spécifiquement pour ce cas et des mesures de similarité entre documents qui donne un lien plus significatif basé sur le contexte des documents.
- Appliquer un processus de prétraitement de texte arabe qui prend en compte sa particularité.

REFERENCES
BIBLIOGRAPHIE

BIBLIOGRAPHIE

Abbes.R. (2004). *La conception et la réalisation de concordancier électronique pour l'arabe. Thèse de Doctorat en Informatique.* France: Université Lyon.

Abdelali, A. (2004). *Improving Arabic Information Retrieval Using Lacal Variation in Modern Standar Arabic.* New Mexico: Institute of Mining and Technology.

Abdelli, B. (2016). *Une Approche Semantique pour les Document Numeriques.Thése de Doctorat en Informatique.* Biskara: Universite Mohamed khiderde Biskara.

Abdulhay, A. (2006). *Constitution d'une ressource s2mantique arabe a partir d'un corpus multilingue align, Thèse de Doctorat en Informatique et Sciences du Langage.* Francais: Universit´e de Grenoble.

Aggarwal.C, & Philip.S (2001). On Effective Conceptual Indexing and Similarity Search in Text Data. *In Proceeding of ICDM '01*. Washinton DC: USA.

Ali Khoudja, M. (2015). *Une Nouvelle Approche de Reformulation Conceptuelle de Requetes: Application sur la Langue arabe, Thèse de Master en Informatique.* Algérie: Universite dr.yhia fares de medea.

Apidianaki,M. (2008). *Acquisition Automatique de Sens pour la Disambiguistion et la Seletion Lexicale en Traduction.These de Doctorat en Linguistique et Descriptive Automatique.* Paris: université Diderot Paris 7.

Azzoug, W. (2013). *Contribution a la définition d'une approche d'Indexaion Sémantique de Document Textuels.* Algérie: Université m'hamed bougara boumerdes.

Badache, I. (2016). *Recherche d'Information Sociale Exploitation des Signaux Sociaux pour Améliorer la Recherche d'Information. Thèse de Doctorat en Informatique.* France: Université Touluse 3 Paul Sabatier.

Baeza-Yates.R , Ribeiro-Neto.B. (1999). *Modern Information Retrieval. Livre.* ACM/ Addison-Wesley.

Baziz, M. (2005). *Indexation Conceptuelle Guideé par Ontologie Pour la Recherche D'Information. Thése de Doctorat en Informatique.* France: Université de Paul Sabatier.

Bellaour, S. (2009). *Construction et utilisation d'un thésaurus pour la recherche d'information sur le web. Thèse de Magister en Informatique.* Alger: Université Kasdi Merbah de Ouargla.

Bellot,P. (2000). *Méthodes de classification et de segmentation locales non supervisées pour larecherche documentaire. Thèse de Doctorat en informatique.* l'université d'Avignon .

Benveniste.R. (2001). *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes, Application au calcul de similarité sémantique dans le cadre du modèle DSIR. Thèse de Doctorat en Informatique.* Suisse: Ecole polytechnique fédérale de lausanne.

Benzater, N. (2015). *Analyse Morphologique du Texte Arabe pour Son Indexation Sémantique, Thèse de Magistère en Informatique.* Algérie: Université des Sciences et de la Technologie - Mohamed Boudiaf - Oran.

Berkeley. (2003). *How Mutch Information. Livre.* http://www2.sims.berkeley.edu/research/projects/how-mutch-info-2003/.

Berkhin,P. (2002). *Survey of Clustring Data Mining Techniques.* Accurue Software.

Blachére.R, Domombynes.G. (1975). *Grammaire de l'arabe classique. Livre.* Paris: Edition Maisonneuve-Larose.

Boucham, S. (2009). *Une Approche baseé Ontologies pour L'Indexation Automatique et la Recherche D'Information Multilingue.thése de Magister en Informatique.* boumerdes: université m'hq,ed bougara de boumerdes.

Boujamaa, M., & Tawes, A. (2009). *Reformulation de requete dans un systéme de recherche d'information semi-structuré.* Alger: ESI.

Boulaknadel, S. (2008). *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, Thèse de Doctorat en Informatique.* Francais: Université de Nantes.

Bouramoul, A. (2011). *Recherche D'information Contextuelle Et Semantique Sur Le Web, Thèse de Doctorat en Informatique.* Algérie: Université Mentouri de Constantine.

Brahimi , B. (2011). *Extraction de Connaissances à Partir de Données Incomplètes et Iimprécises. Thése de Mqgister en Informatique.* M'sila: Universite De M'sila.

Brin, S. (1998). The anatomy of a large-scale hypertextualWeb search engine. 30(1–7) :107–117.

Candillier,L. (2006). *Contexualistion, Visualisation et Evaluation en Apprentissage Non Supervisé. Thése de Doctorat en Informatique.*

Catherine, R. (2010). *Une méthode d'indexation sémantique adaptée aux corpus multilingues. Thèse de Doctorat en Informatique.* Francais: l'INSA de Lyon.

Charhad, M. (Novembre 2005). Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique. 24-25.

Chibane,I. (2008). Impact des liens hypertextes sur la précision en recherche d'information.

Christopher D. Manning, P. R. (2009). *Introduction to information Retrieval, Livre.* Cambridge University Press.

Cleuziou,G. (2004). *Une Methode de Classification Non Supervisé pour L'Appretissqge de Régle et la Recherche D'Information.Thése de Doctorat.*

Cowie.J, Guthrie.J, Guthrie.L. (1992). Lexical Disambiguation Using Simulated Annealing. *International Conference on Computational Linguistics(COLING'92)*, 317-323.

Delekh, T. (2011). *Implementation d'un Outil D'Indexation etde Recherche des Textes en Arabe. thése de Magister en Informatique* . Batna: Université Hadj Lakhdar - Batna.

El ghali, B. (2016). *Modèles Contextuels pour la Recommandation et l'Expansion de Requêtes en Recherche d'Information, Thèse de Doctorat en Informatique et Télécommunications.* Maroc: Faculté des Sciences Rabat Université Mohamed V.

El-Dahadeh. (1996). *معجم قواعد اللغة العربية في جداول و لوحات*. لبنان: مكتبة بيروت للنشر.

Fatiha, B. (2008). *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets, Thèse de Doctorat en Informatique.* Francais: Université Toulouse III - Paul Sabatier.

Forgy,E. (1965). Cluster Analysis of Multivariate Data: Efficiency VS Interpretability of Classifications. *Biometrics*, 768-769.

Gasami, M. (2009). *Utilisation des ontologies pour l'indexation automatique des sites Web en Arabe. Thèse de Magister en Informatique.* Algérie: universite kasdi merbah ouargla.

Gerard, S. (1968). Search and Retrieval Experiments in Real-Time information retrieval .

Ghenima,M. (1998). *Systéme de voyellation de textes arabes. Thèse de Doctorat en Informatique.* France: Université Lyon2.

Gilles , B. (2000). *La similarite : Une notion symbolique/numerique.*

Goker, A., & Davies, J. (2009). *Informmation Retrieval: Serching In The 21ST Century. Livre.* United Kingdom: A Jhon Wiley and Sons, Ltd, Publication.

Grolier. (1990). Grolier Multimedia Enclopedia CD-ROM. *Grolier Interactive*.

Hamada. Z., El Hady.G, Abd El-Wahed.W. (2010). *Query Recommendation for Improving Search Engine Results. Proceedings of the World Congress on Engineering et Computer Science. Livre.* San Francisco USA: WCECS.

Harman ?D (1987). *Conception d'un dictionnaire pour le traitment automatique de l'arabe dans différents contextés d'application. Thèse de doctorat en Informatique.* France: Université Lyon 1.

HarragFouzi. (2011). *Une approche de fouille des textes basée sur la classification et la segmentation thématique: Application au corpus des Traditions Prophétiques "Hadith", Thèse de Doctorat en Informatique.* Algérie: Université Ferhat ABBAS Sétif.

HDR.S. (2001). *Conception de Système de Recherche D'information. Livre.* Lyon: Université Claude Bernard.

Herbrich,R. (2001). *Learning Kernal Classifiers: Theory and Algorithms.* MIT Press.

Hlaoua, L. (2007). *Reformulationde Resuetes par Rénjection de Pertinence dans les Dpcuments Semi-Structurés. Thése de Doctorat en Informatique.* Université Pail Sabatier.

Idri, A. (2003). *Un Modèle Intelligent D'Estimation des Coûts de Développement de Logiciel.Thése de Doctoraten Informatique.* Montréal: Université du Québec à Montréal.

Jalam, R. (2003). *Apprentissage Automatique et Catégorisation de Textes multilingues. Thèse de Doctorat en Informatique.* Franc: Univérsité LUMIERE LYON2.

Khatir, N. (2012). *Clustring Dans les Bases de Données. Thése de Magister.* Oran: Université D'Oran.

Khoja,S. (1999). Stemming Arabic Text.

Kleinberg,J. (1999). Authoritative sources in a hyperlinked environment. 46(5):604–632.

Kolte.S.G, Bhirud.S.G. (2008). Word Sense Disambiguation Using WordNetDomains. *IEEE DOI 10.1109/ICETET*.

Kompaoré, N. (2008). *Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes: vers un processus de RI adaptatif, Thèse de doctorat en informatique.* Université Paul Sabatier de Toulouse.

Laporte,E. (2000). *Mot et niveau lexical. Livre.*

Largouet Anita. (2005). *La recherche d'informations sur Internet. thése de doctorat.* Bordeaux: Université Michel de Montaigne.

Leah.S, Ballesteros.L, Margaret.E, Connell.E. (2005). Light Stemming for Arabic Information Retrieval.

Lesk,ME. (1986). automatic sense disambiguation using machine readable dictionaries: how to tell a pince come from a nice cream one. *SIGDOC*.

Lewis ,D. (1992). Feature Selection and Feature Extraction for Text Categorization. *Proceeding of the Third Annual Symposium on Document Analysis and Information Retrieval.* SDAIR'94.

Luhan, H. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* , 159-165.

Maisonnasse,L. (2008). *Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale. thèse de doctorat en informatique.* France: Université Joseph Fourier- Grenoble I.

Manning, C., Raghvan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval. Livre.* England: Cambridg University Press.

Mark A, WJ, B. (1997). Conceptual and Contextual Indexing Using Wordnet-Derived Lexical Chains. *In the Proceedings of BCS IRSG*, (pp. 47-65).

Mizoguchi,R. (2004). Le Role de L'ingénierie Ontologique dans des EIAH. *Sciences et Technologies de L'information et de la Comminication pour L'Education et la Formation*, 11.

Mohamed Salim El Bazzi, T. Z. (s.d.). Indexation automatique des textes Arabes. *e-TI, Revue Electronique des Technologies de l'Information*.

Mohammed.S, Hirst.G. (2006). Determining Word Sense Dominance Using a Theasaurus. *Proceeding of the 11th Conference on European Chapter of the Assoiation for Computational Linguistics()* (pp. 121-128). Italy: EACL.

Moulinier.M, , Ganascia.J.G. (1996). Applying an Existing Machine Learning Algorithm to Text Categorization Connectionist, Sttistical, Statistical,and Symbolic Approoachees to Learning for Natural Language Processing. *Springer-Verlag*, 343-354.

Navigli,R. (2009). Word Sense Disambiguation: A survey. *ACM ComputingSurveys, Vol*.

Ourdia, B. (2007). *Prise en compte de l'Hétérogénéité Structurelle en Recherche d'Information semi-structurée, Thèse de Magister en Informatique.* Algérie: Université m'hamed bougara boumerdes.

Pantel.P, ,Lin.D. (2002). Discovering Word Sense from Text. *Proceedings of the 8th ACM SIGKDD International conference on Knowledge Discovery and Data Mining* (pp. 613-619). Edmonton, Alta: Canada.

Pederson, Weigend.A. (1995). A Neural Network Approach to Topic Spotting. *4th Annual Symposium on Document Analysis and Information Retrieval, SDAIR24*, 317-332.

Picarougne,F. (2004). *Recherche d'information sur Internet par algorithmes évolutionnaires. Thèse de doctorat en informatique.* Université François Rabelais Tours.

Ressad-Bouidghaghen, O. (2011). *Accès contextuel à l'information dans un environnement mobile.* Toulouse : l'Université Toulouse III - Paul Sabatie .

Reymonet, A. (2008). *Modélisation de connaissances à partir de textes pour une Recherche d'Information Sémantique, Thèse de Doctorat en Informatique.* Universite De Toulouse.

Riejsbergen, C. (1979). *Information Retrieval. Livre.* London: Butterworth.

Rijsbergen, K. (2004). *The Geometry of Information Retrieval. Livre.* Cambridg University Press.

Robertson, S. E. (1977). *The probability ranking principle in ir. Livre.*

Robertson.E ,Walker.S. (1994). *Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. Livre.* SIGIR.

Roget. (1995). *The New Thesaurus* . Boston: Houghton Mifflin.

ROMARY,L. (2000). *Outils d'accés à des ressources linguistiques. Livre.*

Salton ,G. (1968). *Automatic Information Organization and Retrival, Livre.* New York: Hill Book Company.

Salton, G. (1971). A comparison between manual and automatic indexing method. *Journal of American Documentation*, 61–71.

Salton.G, McGill.M. (1983). Introduction to ModrenInformationRetrieval.

Sauvagnat, K. (2005). *Modele flexible pour la Recherche d'Information dans des corpus de documents semi-structurés, Thèse de Doctorat en Informatique.* Université Paul Sabatier de Toulouse.

Schutze ,H. (1998). Automatic Word Sense Discrimination. *Computational Linguistic: Special Issue on Word Sense Disambiguaton*, pp. 97-123.

Singhal.A, Salton.G, Mitra.M, & Buckley.C (1996). Pivoted Document length normalization In Proceeding of the 19th annaul international ACM SIGIR. conference on in development in information retrival.

SOULIER, L. (2014). *Définition et évaluation de modèles de recherche d'information collaborative basés sur les compétences de domaine et les rôles des utilisateurs? Thèse de Doctorat en Informatique.* Université Toulouse 3 Paul Sabatier.

Sunssa,M. (1993). Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network. *2nd International Conference on Infromation and Knowledge management(CIKM-1993),* 67-74.

Thi Hoang, D. l. (2010). *Utilisation de ressources externes dans un modèle Bayésien de Recherche d'Information,Application à la recherche d'information médicale multilingue avec UMLS. Thése de Doctorqt en Informatiaue.* Universite joseph fourier.

Weiss,F. (1973). *Learning to Diambigute, Information Storage and Retrieval. Livre.*

William A, W. (1997). *Conceptual Indexing: A Better Way to Recognize Knowledge.* Woods.

Yarowsky,D. (1992). *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.* International Conference on Computational Linguistics(COLING-92).

Yarowsky,D. (2000). Hierarchical Decision Lits dorWord Sense Disambiguation. *Journal Computers and Huminities*, 179-186.

Zemirli,W.N. (2008). *Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif. Thése de doctorat en informatique.* Toulouse: Université Paul Sabatier de Toulouse III.

Zipf,GK . (1949). *Human Behavior and the Principale of Least Effort. Livre.* Cambridge: Addision Welsey.

Site Web :

Java Development Kit (JDK). (2017, 05 15). Récupéré sur techopedia: https://www.techopedia.com/definition/5594/java-development-kit-jdk

Java. (2017, 05 10). Récupéré sur Edu Tech Wiki: http://edutechwiki.unige.ch/fr/Java

JavaMI. (2017, 05 15). Récupéré sur JavaMI: http://www.cs.man.ac.uk/~pococka4/JavaMI.html

NetBeans IDE 8.2 Information. (2017, 05 15). Récupéré sur NetBeans: https://netbeans.org/community/releases/82/

NetBeans. (2017, 05 11). Récupéré sur Techno-Science: http://www.techno-science.net/?onglet=glossaire&definition=5346

RapidMiner Studio. (2017, 05 10). Récupéré sur Edu Tech Wiki: http://edutechwiki.unige.ch/fr/RapidMiner_Studio

Text Processing. (2017, 05 10). Récupéré sur RapidMiner: https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_text

T*Welcome To JFreeChart*. (2017, 05 15). Récupéré sur JFreeChart: http://www.jfree.org/jfreechart/

ANNEXE

ANNEXE A

RapidMiner

1. Introduction

1.1. Présentation

RapidMiner est un logiciel open source et gratuit dédié au data Mining. Il contient de nombreux outils pour traiter des données : lecture de différents formats d'entrée, préparation et nettoyage des données, statistiques, tous les algorithmes de data Mining, évaluation des performances et visualisations diverses.

C'est un logiciel puissant, il n'est pas facile à manipuler au premier abord, mais avec un peu de pratique, il permet de mettre en place rapidement une chaîne complète de traitement de données, de la saisie des données à leur classification.

1.2. Ressources

RapidMiner est disponible sur le site http://www.rapidMiner.com, où l'on trouve aussi des tutoriaux, un blog et des forums.

2. Téléchargement de RapidMiner Studio

Suivez ces instructions pour télécharger RapidMiner Studio:

1. Pour télécharger la demande, accédez à la :

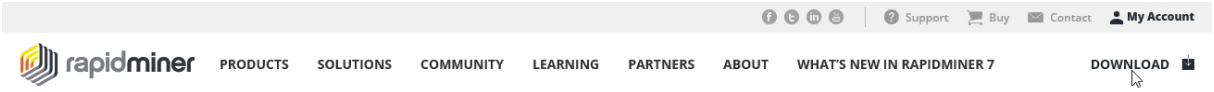


Figure A. 1 : Barre d'acaile de RapidMiner site web

2. Cliquez sur le bouton Télécharger (download en anglais) dans le coin supérieur droit.
3. Connectez-vous si vous ne l'avez pas encore. Voir ci-dessous si vous devez créer un compte. Si vous souhaitez simplement télécharger RapidMiner Studio sans vous connecter, cliquez sur Téléchargements.

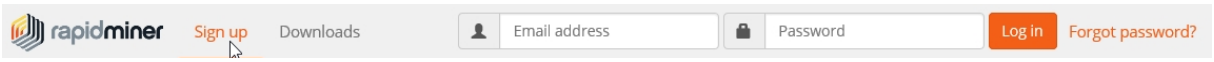


Figure A. 2 : Barre de Connexion

4. Cliquez sur votre système d'exploitation préféré pour commencer le téléchargement. Votre système d'exploitation actuel sera mis en surbrillance.

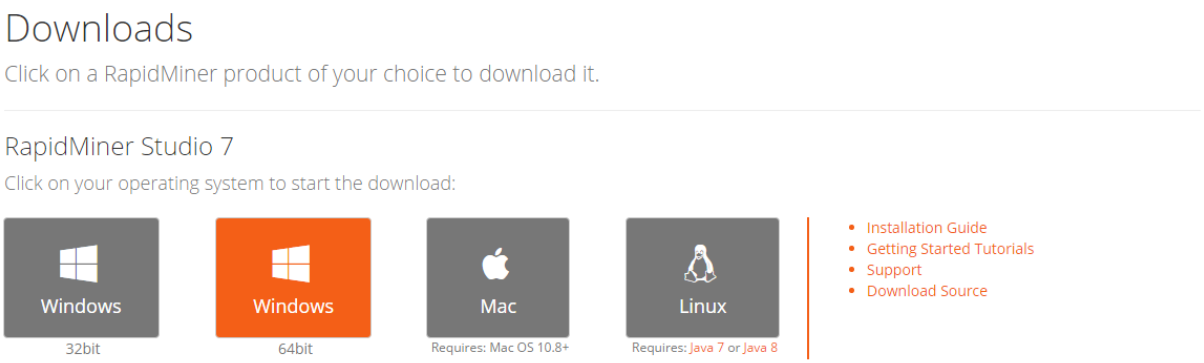


Figure A. 3 : Page de téléchargement de RapidMiner

3. Installation de RapidMiner Studio sur Windows

Suivez ces instructions simples pour exécuter le lanceur qui installe RapidMiner Studio sur Windows.

1. Double-cliquez sur le fichier téléchargé (par exemple, rapidminer-studio-<version>-win64-install.exe).
2. Si vous êtes invité, permettez au programme d'apporter des modifications à votre ordinateur. L'assistant de configuration RapidMiner Studio apparaît. Cliquez sur Suivant pour continuer.
3. Lisez les termes du contrat de licence et cliquez sur Je suis d' accord pour continuer.
4. Sélectionnez un dossier de destination (ou laissez la valeur par défaut). Assurez-vous que le chemin du dossier ne contient pas + ou % caractères. En cliquant sur Installer, l'assistant extrait et installe RapidMiner Studio. Une fois l'installation terminée, cliquez sur Suivant, puis sur Terminer pour fermer l'assistant et démarrez RapidMiner Studio.
5. Lisez les termes du contrat de licence et cliquez sur Accepter pour continuer.

4. Mise en route de RapidMiner Studio

Une fois que vous lancez RapidMiner Studio, un écran de bienvenue s'affiche, avec deux options:

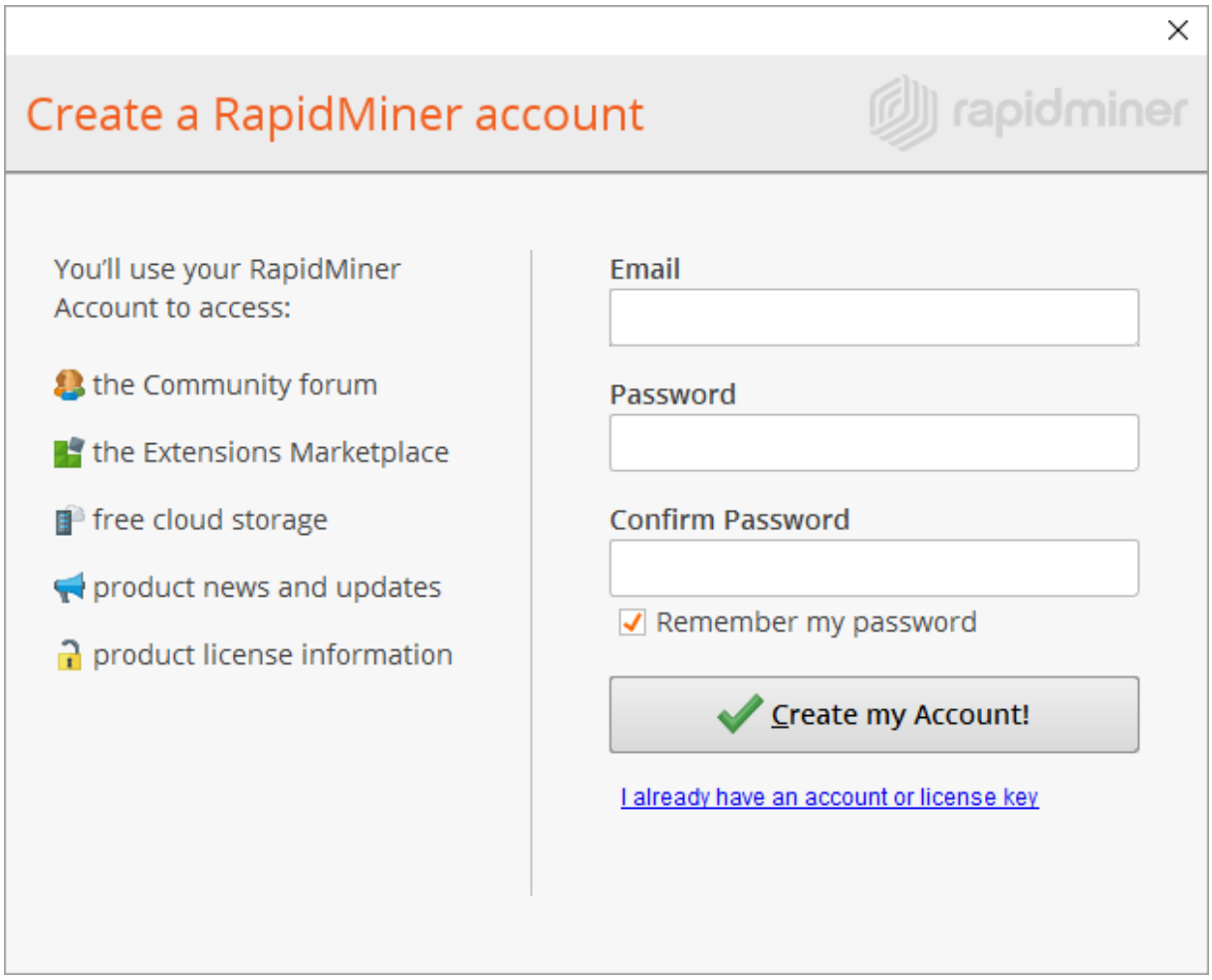


Figure A. 4 : Fenêtre de création de compte en RapidMiner Studio

Option	La description
Créer mon compte	Crée un compte avec les informations d'utilisateur fournies.
J'ai déjà un compte ou une clé de licence	Si vous vous êtes déjà inscrit, entrez vos informations d'identification de quickminer.com pour vous connecter afin que vous puissiez télécharger ou installer votre (vos) licence (s).

Tableau A. 1 : Description des option de la Fenêtre de création de compte de RapidMiner

Une fois connecté, si vous avez acheté une licence, le système l'installe en fonction de vos informations d'identification.

4.1. Créer un compte RapidMiner

La création d'un compte exige que RapidMiner Studio puisse accéder à Internet. Si vous ne disposez pas d'une connexion Internet pour l'application (par exemple, s'il est empêché par un pare-feu), créez un compte via votre navigateur et entrez votre clé de licence manuellement.

1. Complétez les étapes pour créer un compte en entrant votre adresse e-mail et votre mot de passe. Notez que le mot de passe doit comporter un minimum de six caractères.
2. Immédiatement, le système envoie un email d'activation à l'adresse que vous avez enregistrée. (Autorisez le courrier électronique de RapidMiner si vous ne voyez pas un courrier électronique intitulé «Vérifiez votre courrier électronique» dans votre boîte de réception.) Les contrôles périodiques vérifient si votre courrier électronique a été validé.

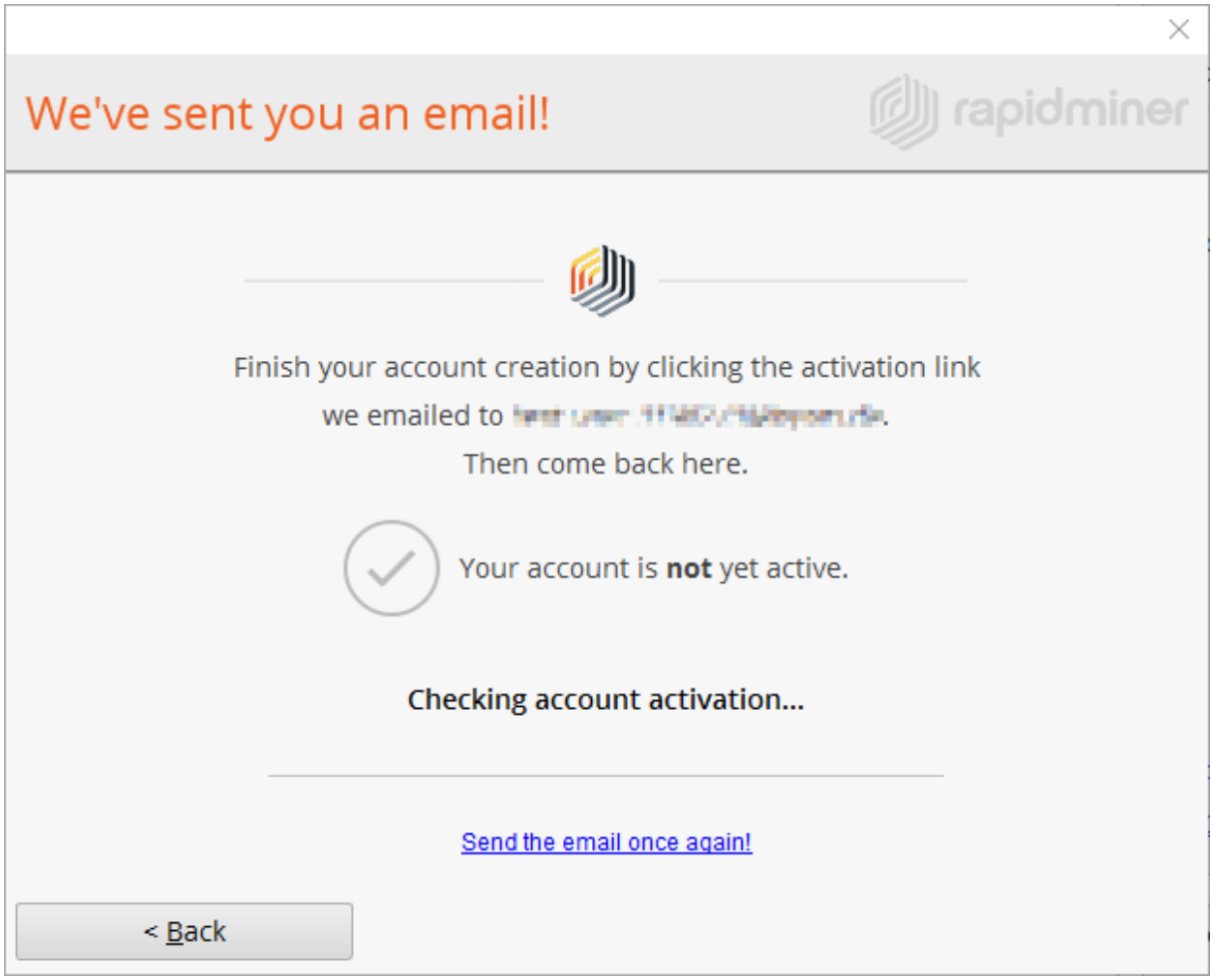


Figure A. 5 : Fenêtre de Confirmation de la création de compte RapidMiner

3. Une fois que vous validez, vous recevrez un message de réussite dans votre navigateur:

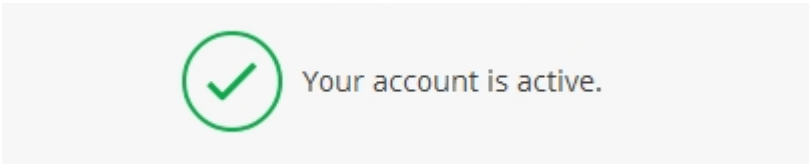


Figure A. 6 : Message de ssucsee de création de compte RapidMiner

4. Revenez à l'application RapidMiner Studio. Complétez l'installation en cliquant sur Je suis prêt.

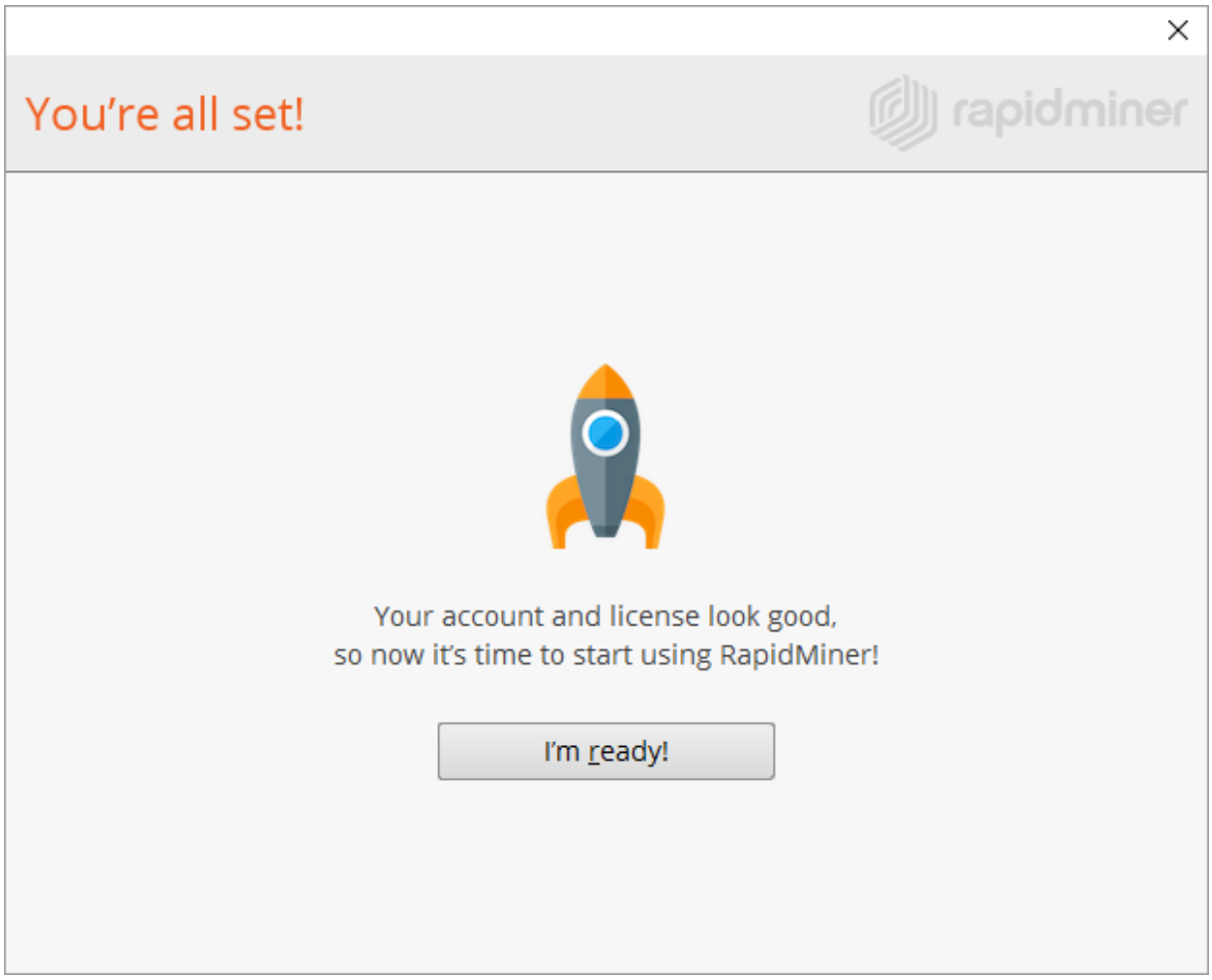


Figure A. 7 :Fenêtre de Confirmation Final de compte RapidMiner

5. Si c'est la première fois que vous installez RapidMiner Studio, nous vous recommandons de commencer avec le didacticiel affiché pour apprendre à quel point il est facile de créer un code sans code. Bonne mine!

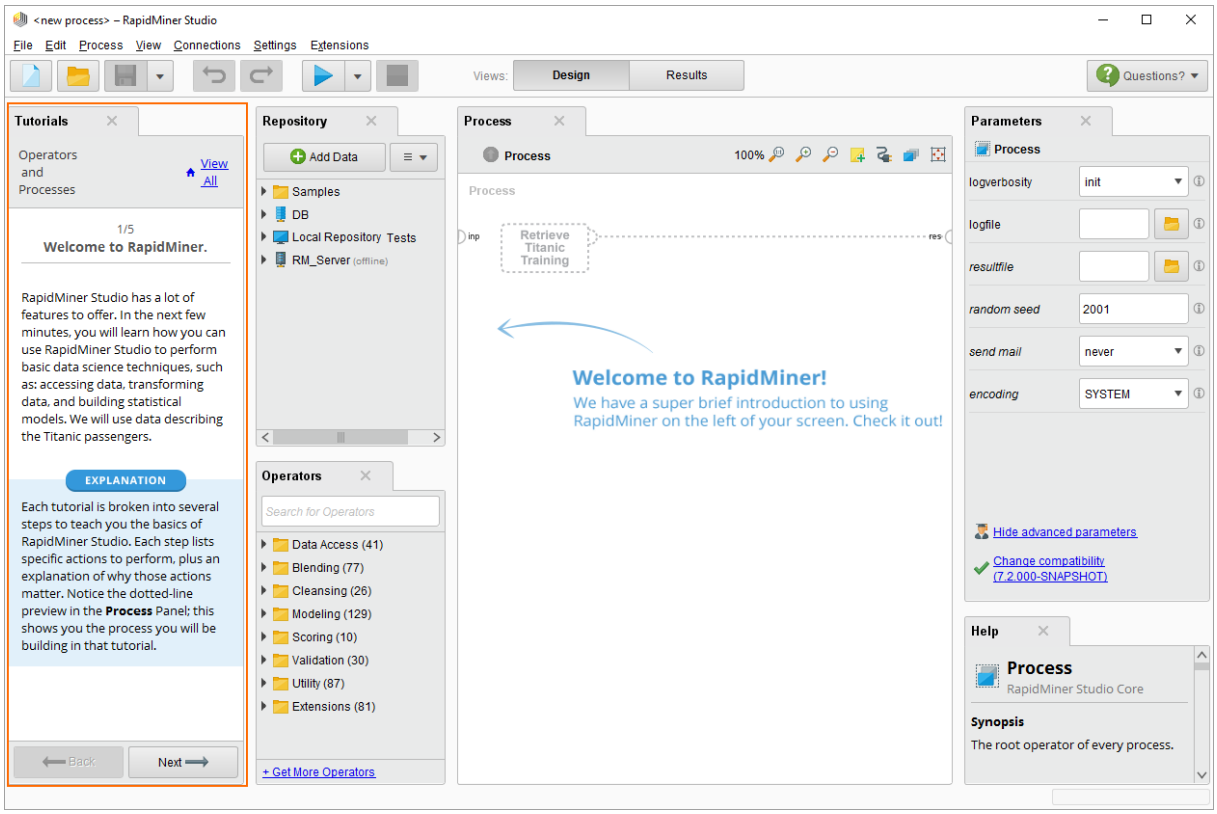


Figure A. 8 : Fenêtre principale de RapidMiner studio

5. Prise en main

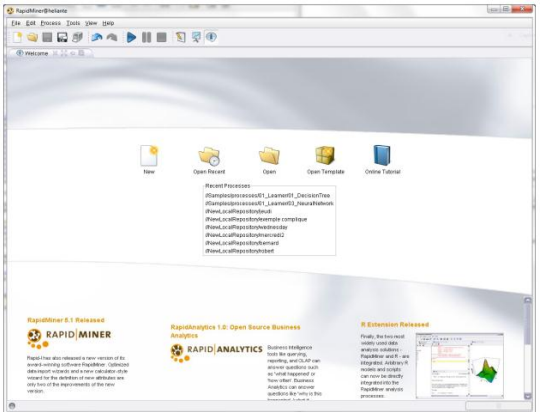


Figure A. 9 : Le premier écran

A l'ouverture du programme, RapidMiner présente l'écran de la Figure A.9, qui vous propose soit de commencer un nouveau schéma de traitement de données, soit d'en ouvrir un précédemment écrit, soit de suivre le tutoriel en ligne (celui-ci suppose une connaissance préalable des méthodes et du vocabulaire du data mining, et devient vite très compliqué à suivre). Choisissez 'new', ce qui vous amène à l'écran de la Figure A.10, où vous devez choisir l'endroit où se trouvera votre espace de travail. Choisissez 'NewLocalRepository' et entrez un nom. On arrive maintenant à l'écran de travail proprement dit, illustré sur la Figure A.11. Cet espace de travail se divise en cinq zones principales :

- La zone centrale servira à dessiner le schéma du flux de données : il partira d'une source, qui passera successivement à travers plusieurs outils (préparation des données, traitement, évaluation de performances . . .).

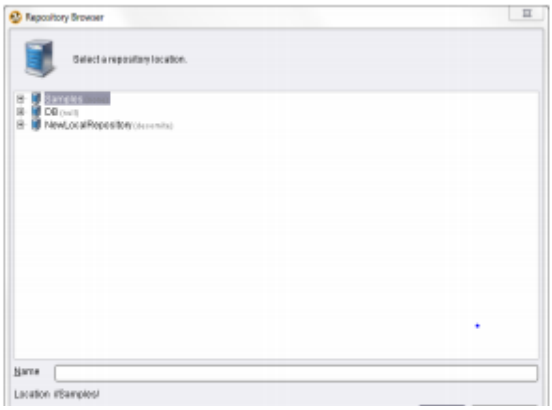


Figure A.11 : Choix du dépôt

- Le catalogues des outils, à gauche, où nous irons chercher les composants à utiliser dans la zone centrale. Ces composants peuvent être des sources de données, des modificateurs de données, des algorithmes de data mining, des évaluateurs de performances. Nous les utiliserons chacun à leur tour, un peu à la fois.
- Pour chaque outil utilisé, la zone de droite listera ses paramètres.
- La zone des informations en bas à droite renseigne sur le fonctionnement de l'outil courant.
- En bas, la fenêtre des messages d'erreurs.

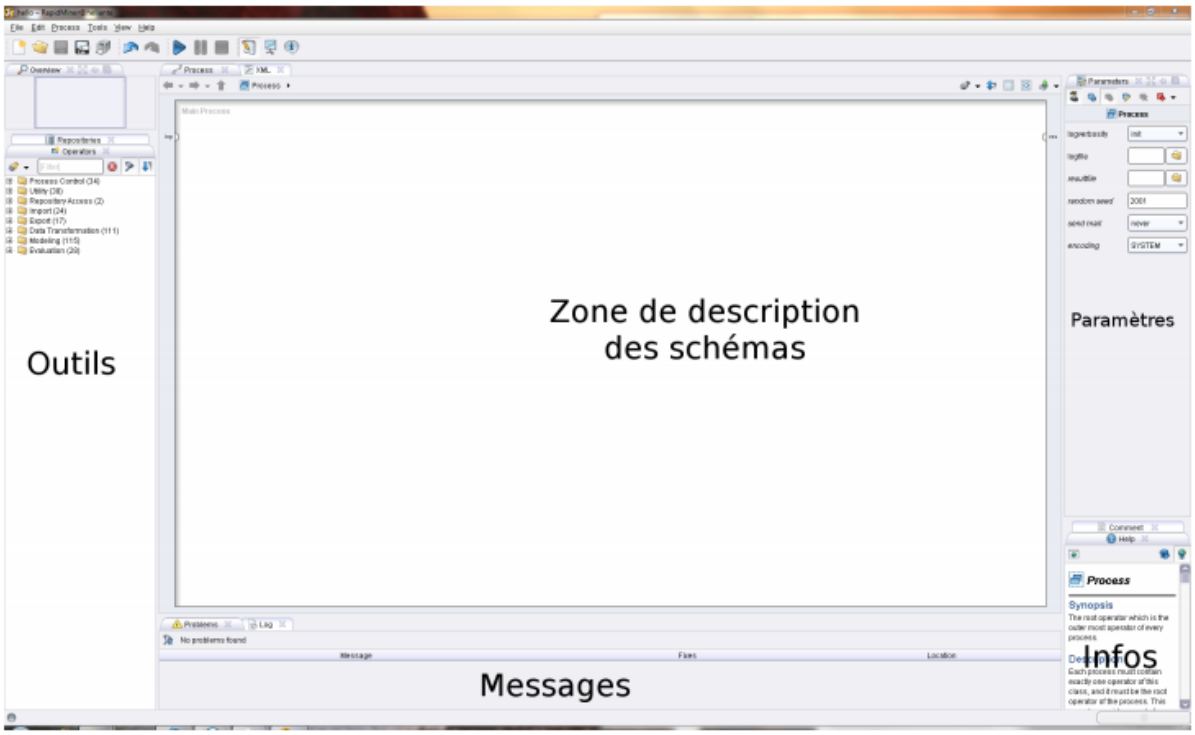


Figure A. 11 : Espace de travail

6. Plus que vous devez savoir

Juste quelques autres astuces concernant RapidMiner Studio …

Utilisation du marché RapidMiner

Le RapidMiner Marketplace est votre site de guichet unique pour télécharger et partager des extensions pour RapidMiner Studio. Suivez ces étapes pour obtenir des instructions détaillées sur l'accès et l'utilisation du marché RapidMiner.

Déplacement du script de démarrage

Par défaut, la variable RAPIDMINER_HOME est définie dans l'emplacement du dossier du script de démarrage. Si vous laissez le script dans le dossier d'installation (la méthode la plus commune), la variable d'environnement est définie correctement. Si vous déplacez le script de démarrage hors du dossier d'installation, vous devez définir la variable d'environnement RAPIDMINER_HOME sur le chemin d'installation de RapidMiner Studio. Cela permet à RapidMiner de trouver ses fichiers de ressources.