



UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHÉMATIQUES ET
DE L'INFORMATIQUE



DEPARTEMENT D'INFORMATIQUE

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Technologie de l'Information et de la Communication

Par : Behlouli Selma

SUJET

**Une approche sémantique pour la classification
des traditions prophétiques**

Soutenu publiquement le : / /2016 devant le jury composé de :

.....	Université de M'sila	Président
Mm. HELASSA Madiha	Université de M'sila	Rapporteur
.....	Université de M'sila	Examineur
.....	Université de M'sila	Examineur

Promotion : 2015 /2016



UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHÉMATIQUES ET
DE L'INFORMATIQUE



DEPARTEMENT D'INFORMATIQUE

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Technologie de l'Information et de la Communication

Par : Behlouli Selma

SUJET

**Une approche sémantique pour la classification
des traditions prophétiques**

Soutenu publiquement le : / /2016 devant le jury composé de :

.....	Université de M'sila	Président
Mm. HELASSA Madiha	Université de M'sila	Rapporteur
.....	Université de M'sila	Examineur
.....	Université de M'sila	Examineur

Promotion : 2015 /2016

Remerciements

Je remercie d'abord et Avant toute chose, ALLAH
Le tout puissant qui m' a donné la volonté et la patience
nécessaires pour terminer ce travail.

J'adresse mon profonde gratitude à Madame
HELLASSA MADIHA.

Je la remercie pour sa gentillesse, et pour tous ses
conseils et remarques constructives qui m'ont permis
d'améliorer la qualité de ce travail. Je tiens à lui exprimer
mon grand respect et du plaisir que j'ai travaillé avec elle.

Comme je remercie le président de jury et l'examineur
pour avoir bien voulu accepter de juger ce travail.

Enfin, je souhaite exprimer toute ma gratitude à tous
ceux qui ont contribué par leurs conseils ou leurs
encouragements à l'aboutissement de ce modeste travail.

B. Selma

Table des matières

INTRODUCTION GENERALE	1
CHAPITRE 1 : Les caractéristiques de la langue arabe	
1 Introduction	3
2 Particularité de la langue arabe	4
2.1 a structure d'un mot arabe	5
2.2 Les catégories des mots	5
2.2.1 Verbe	5
2.2.2 Particule	5
2.2.3 Nom	6
2.3 Morphologie arabe	7
2.3.1 Principe	7
2.3.2 Objectifs	7
• Les schèmes	7
• Les racines	8
• Les stems	8
• Les signes diacritiques	9
• Mots dérivés	9
• Mot isolés	9
• Les affixes	9
3 Quelques Problèmes du traitement automatique de la langue arabe	10
2.4 Le problème de la voyellation	10
2.5 Le problème de l'agglutination	10
2.6 L'extraction de la racine	11
4 Conclusion	11

CHAPITRE 2 : Fouille de données et Fouille de textes

1	Introduction	12
2	Fouille de données	12
2.1	Définition d'ECD (KDD)	12
2.2	processus d'extraction de connaissances à partir des données..	13
2.3	Définitions de la fouille de données	14
2.4	Pourquoi la fouille de données ?	14
2.5	Les domaines d'application de la fouille de données	14
2.6	Difficultés de la fouille de données	15
2.7	Tâches de fouille de données	16
3	La fouille de textes	17
3.1	Définitions de la fouille de textes	17
3.2	Les objectifs de fouille de textes	18
3.3	Tâches de la fouille de textes	18
3.4	Processus de fouille de textes	20
3.5	Domaine d'application de la fouille de textes	21
4	Conclusion	22

CHAPITRE 3 : Corpus des hadiths

1	Introduction	23
2	Le Hadith	23
2.1	La Sunna	24
2.2	Composants du hadith	24
2.3	Classification des Hadiths.....	26
2.4	Collections importants de hadith	28
2.5	Imâm Al-Bukhârî	29
2.6	Sahih Al-Bukhârî, le livre le plus juste après le Coran	30

3 Les corpus	30
3.1 Définition	30
3.2 Le corpus en littérature	31
3.3 Le corpus en linguistique	31
3.4 Le corpus dans la science	31
3.5 Domaine d'Utilisation du corpus	31
3.6 Conditions pour construire un corpus textuel	32
3.7 Méthodologie d'utilisation d'un corpus	32
3.8 Les avantages d'une analyse de corpus	32
4 Conclusion	33

CHAPITRE 4 : Classification des documents textuels : Etat de l'art

1 Introduction	34
2 Définition de la Classification	34
2.1 Pourquoi automatiser la classification ?	34
2.2 Domaines d'application de la classification	35
2.3 Classification bi-classe et multi-classes	35
2.4 Classification de textes et Text Mining	35
2.5 Hiérarchie des méthodes de classification	36
2.6 Les types de classification	37
2.6.1 Classification Non supervisé (clustering)	37
2.6.2 Classification Supervisé (catégorisation)	37
3 Définition de la Catégorisation des textes	37
3.1 Comment catégoriser un texte ?	38
3.2 Domaine d'application de la Catégorisation de textes	39
3.3 Problèmes de la catégorisation de textes	40

4	Choix des termes	42
4.1.	Représentation en « sac de mots » « bag of words »	42
4.2.	Représentation des textes par des phrases	43
4.3.	Représentation des textes avec des racines lexicales (stemming)	43
4.4.	Représentation des textes avec des lemmes (lemmatisation)	43
4.5.	Représentation des textes avec la méthode des n-grammes	44
5	Calcul de poids	44
5.1	Codage TF-IDF	45
5.2	TFC	45
6	Réduction de la dimension	46
6.1	Sélection des termes	46
6.2	Extraction d'attributs	46
7	Algorithmes d'Apprentissage	47
7.1	Algorithme des k-voisins les plus proches	47
7.2	Arbre de décision	47
7.3	Classificateur bayésien naïf	47
7.4	Réseaux de neurones	48
7.5	Machines à support vectoriel	48
8	Critères d'évaluation d'algorithmes d'Apprentissage	48
9	Etat de l'art	50
10	Conclusion	53
CHAPITRE 5 : Réalisation et expérimentation		
1	Introduction	54
2	Le langage et l'environnement de programmation	54
2.1	L'environnement de programmation (Visual Studio 2013)	54

2.2Le langage de programmation (C#)	54
3 Présentation de corpus utilisé	55
4 Processus de catégorisation suivi par notre application	56
5 Présentation de l'application réalisée	57
5.1Interface principale	57
5.2Prétraitement sur le « Matn »	57
• Tokenization	59
• Normalisation	59
• Mot vide	60
• Lemmatisation et stemming	61
5.3L'approche utilisée pour la représentation de corpus	62
5.4Représentation de texte par tf*Idf et apprentissage	62
5.5Choix de l'algorithme d'apprentissage	62
5.6Evaluation des résultats du classifieur	63
6 Conclusion	63
CONCLUSION GENERALE	64

Liste des figures :

Figure 2.1	transformation de données brutes en connaissances utiles.	13
Figure 2.2	différentes étapes du processus ECD.	13
Figure 2.3	schéma général d'une tâche de fouille de textes.	19
Figure 2.4	Processus de la fouille de données textuelles.	20
Figure 3.1	Composants du hadith.	24
Figure 3.2	Exemple de composants hadith.	25
Figure 3.3	Classification des hadiths.	26
Figure 4.1	Hiérarchie des méthodes de classification.	36
Figure 4.2	Processus de la catégorisation de textes.	39
Figure 4.3	Exemple de la Représentation d'un texte extrait de la collection CLEF en sac de mots	43
Figure 5.1	Processus de catégorisation proposée.	56
Figure 5.2	L'interface principale de l'application.	57
Figure 5.3	parcourir des fichiers.	58
Figure 5.4	sélection d'un fichier.	58
Figure 5.5	Etape de Tokenization.	59
Figure 5.6	Etape d'éliminations des ponctuations, chiffres et caractères spéciaux.	60
Figure 5.7	Etape d'élimination des mots vide.	61
Figure 5.8	Etape de stemming	61
Figure 5.9	TF-IDF des termes du fichier.	62
Figure 5.10	Classification du hadith.	63

Liste des tableaux :

Tableau 1.1 :	Etat de transcription des lettres arabes.	3
Tableau 1.2 :	Ambiguïté causée par l'absence de voyelles pour les unités lexicales مدرسة et كتب	4
Tableau 1.3 :	classement des sous catégories de noms.	6
Tableau 1.4 :	quelques dérivations de la racine (كَتَبَ, kataba, écrire).	8
Tableau 1.5 :	les différentes voyellations du mot « شهد »	11
Tableau 4.1 :	Table de contingence à la base de l'évaluation des classificateurs.	49
Tableau 5.1 :	Le corpus utilisés dans nos expérimentations.	55

INTRODUCTION GENERALE

La recherche en informatique accorde ces dernières années, beaucoup d'importance au traitement des données textuelles. Ceci pour plusieurs raisons : un nombre croissant de collections mises en réseau et distribuées sur le plan international, le développement des infrastructures de communication et d'Internet. Les traitements manuels de ces données s'avèrent très coûteux en temps et en personnel, ils sont peu flexibles et leur généralisation à d'autres domaines est presque impossible ; Pour cela, plusieurs méthodes et techniques de gestion et de traitement d'information ont été développées.

Comme nous le savons, tout le monde a besoin d'informations dans sa vie quotidienne. Nous avons besoin de l'information pour prendre les meilleures décisions possibles. Dans chacune de nos activités personnelles, les décisions sont requises et l'information est nécessaire pour soutenir ces décisions. L'information donc est nécessaire presque dans tous les domaines de la pensée et de l'action humaine.

Le domaine de la fouille de textes (text mining) s'est développé pour répondre à volonté à la gestion par contenu des sources volumineuses de textes. A l'heure actuelle, de nombreux logiciels de classification de textes sont disponibles, ils ont fait l'objet de publications et leurs champs d'application s'élargissent de jour en jour. En général, ces systèmes sont basés sur des algorithmes d'apprentissage automatique.

Avec le développement de l'informatique et l'apparition de la fouille de textes, plusieurs applications autour des traditions du Prophète Muhammad (paix et salut sur lui) ont été développées ; malgré cela il y a absence d'une référence standard utilisable.

Notre travail se situe dans le contexte de l'extraction des connaissances intéressantes et non-triviales à partir des corpus de textes prophétiques en utilisant des techniques issues du domaine de la fouille de texte. Nous nous intéressons principalement au problème de la classification thématique des traditions prophétiques.

Organisation du mémoire

Pour l'organisation de notre mémoire, nous proposons le plan suivant :

- Une introduction générale donne un aperçu sur le sujet de notre mémoire.
- Le premier chapitre est consacré à la présentation de la langue arabe et ses principales caractéristiques influant sur son traitement automatique.

- Le deuxième chapitre permettra de présenter quelques notions sur la fouille de données et la fouille de textes. Ce chapitre est très didactique et agréable à lire car Il offre un aperçu très large du domaine et les références sont nombreuses et utiles pour le lecteur.
- Dans le troisième chapitre, nous présentons une étude sur les traditions prophétiques et quelques aspects issus des sciences de hadith, aussi nous donnons une généralité sur les corpus textuelle
- le quatrième chapitre donne une présentation générale sur l'apprentissage automatique et la catégorisation des textes, ainsi que les différentes méthodes utilisées dans la littérature. pour pouvoir passer au dernier chapitre
- Le cinquième chapitre décrit le fonctionnement de l'application suivi par une phase d'évaluation des résultats obtenus à l'aide du classifieur implémenté afin de mesurer sa performance.
- En fin, nous terminerons avec une conclusion qui dégage quelques perspectives importantes qui font suite à ce travail.

CHAPITRE 1

LES CARACTERISTIQUES DE LA LANGUE ARABE

1 Introduction

L'arabe existe et se développe à partir du 7^{ème} siècle grâce à diffusion du Coran qui est considéré comme la base de cette langue. Avec la diffusion de la langue arabe sur le Web et la disponibilité des moyens de manipulation des textes arabes, les travaux de recherche ont abordé des problématiques variées comme la morphologie, la traduction automatique, l'indexation des documents, etc. Malgré ces nombreuses recherches, la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique de la langue à cause de sa richesse morphologique.

2 Particularité de la langue arabe

La langue arabe s'écrit et se lit de droite à gauche, son alphabet compte 28 consonnes adoptant différentes graphies selon leur position (au début, au milieu ou à la fin d'une unité lexicale).

Les formes		Des lettres	
Fin	milieu	début	isolée
ب	ب	ب	ب
ت	ت	ت	ت
ث	ث	ث	ث
ن	ن	ن	ن

Tableau 1.1 Etat de transcription des lettres arabes.

Une unité lexicale arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres. Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte et elles permettent de différencier des unités lexicales ayant la même représentation.

Pour mieux comprendre prenons l'exemple de كُتِبَ du (Tableau 1.2): [20]

- كُتِبَ ، Il a écrit
- كُتِبَ ، Il a été écrit
- كُتُبَ ، Livres

Unité lexical	1 ère interprétation		2 éme interprétation		3 éme interprétation	
كتب	كُتِبَ	Il a écrit	كُتِبَ	Il a été écrit	كُتِبَ	Livres
مدرسة	مَدْرَسَة	Ecole	مُدْرَسَة	Enseignante	مُدْرَسَة	Enseignée

Tableau 1.2 Ambiguïté cusée par l'absence de voyelles pour les unités lexicales كتب et مدرسة

2.1 Structure d'un mot

En arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire, la représentation suivante schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche.

Post fixe	Suffixe	Corps schématique	préfixe	antéfixe
-----------	---------	-------------------	---------	----------

- Antéfixes sont des prépositions ou des conjonctions.
- Préfixes et suffixes expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,...)
- Post fixes sont des pronoms personnels [9]

Exemple :

- أَتَذْكُرُونَنَا

Ce mot exprime la phrase en français : "Est ce que vous vous souvenez de nous? "

La segmentation de ce mot donne les constituants suivants :

أ | ت | ذَكَّرُ | وَدْ | نَا

Antéfixe : أ conjonction d'interrogation

Préfixe : تْ préfixe verbal du temps de l'inaccompli.

Corps schématique: ذَكَّرُ dérivé de la racine: ذَكَرَ

Suffixe : وَدْ suffixe verbal exprimant le pluriel

Post fixe : نَا pronom suffixe complément du nom. [9]

2.2 Les catégories grammaticales

Il existe trois catégories pour un mot arabe : (nom, verbe et particule).

2.2.1 Le verbe

Nous pouvons classer les verbes arabes selon plusieurs critères : Selon le nombre et la nature des consonnes de leurs racines, et selon leurs modèles.

En classant les verbes selon le nombre des consonnes de la racine, nous aurons soit des verbes trilitères qui ont trois consonnes, soit des verbes quadrilatères, peu nombreux, qui ont quatre consonnes. Selon le modèle et le nombre de consonnes qui constituent la structure verbale, nous avons soit des verbes nus (مُجَرَّدٌ) qui sont composés seulement par les consonnes de leurs racines et des voyelles brèves, soit des verbes augmentés ou dérivés (مَزِيدٌ) qui sont 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.

La conjugaison des verbes dépend de plusieurs facteurs :

- Le temps (accompli, inaccompli).
- Le nombre du sujet (singulier, duel, pluriel).
- Le genre du sujet (masculin, féminin).
- La personne (première, deuxième et troisième)
- Le mode (actif, passif). [25]

2.2.2 Le particule

Les particules sont des lemmes invariables et en nombre limité. Ils indiquent l'articulation de la phrase. Elles sont classées selon leur champ sémantique et leur fonction dans la phrase; on en distingue plusieurs types :

- **Préposition** : exemple (عَنْ, بَ, لَ, حَتَّى)
- **Particules de coordination** : exemple (وَأَوْ, ثُمَّ, فَ, وَ)
- **Particules interrogatives** : exemple (أَمْ, هَلْ, مَا)
- **Particules d'affirmation** : exemple (بَلَى, أَجَلْ, نَعَمْ)
- **Particules de négation** : exemple (لَمْ, لَنْ, لَا)
- **Particules distinctive** : exemple (أَيَّ)
- **Particules relatives** : exemple (مَا)
- **Particules de future** : exemple (فَ, سَوْفَ, لَنْ)
- **Particules conditionnelles** : exemple (لَوْ, إِنْ) [3]

2.2.3 Le Nom

Les noms arabes regroupent les substantifs, les adjectifs et les pronoms, ainsi que d'autres noms invariables. Les substantifs et les adjectifs sont créés en prenant pour origine tantôt un type verbal, tantôt un type nominal. Nous pouvons distinguer dans le (tableau 1.3) deux classes de noms : la première regroupe les noms conjugables ou semi -conjugables qui peut avoir la forme duelle, plurielle, etc. la deuxième classe regroupe les noms non- conjugables qui gardent la même forme quel que soit le contexte. Les noms conjugables sont soit des noms primitifs, qui échappent à toute dérivation comme « غَزَال » (gazelle), soit des noms dérivationnels qui sont formés à partir d'une racine comme « مَكْتَبَة » (bibliothèque) de la racine « كَتَبَ ». [25]

Catégorie	Dérivation	Conjugaison	Sous-catégorie	Exemple
Nom	Dérivation nel irrégulier	Non conjugable	Adverbe	أَيْنَ، حَيْثُ، قَبْلَ
			Nom de voix	كَخْ، نَخْ
			Nom de verbe	هَيْهَاتَ، أَهْ، أَفْ
			Pronom personnel (affixé ou isolé)	هُوَ، أَنَا، تُو، تَنْ
			Pronom interrogatif	كَيْفَ، مَتَى، مَا
			Pronom conditionnel	مَنْ، إِذَا
			Pronom allusif	كَمْ، كَأَيَّ
		Conjugable	Pronom relatif	الَّذِي، الَّتِي
			Nom de nombre	ثَلَاثَةٌ، وَاحِدٌ، خُمْسَةٌ
			Pronom démonstratif	هَذَا، هَذِهِ
			Nom propre	مُحَمَّدٌ، هُنْدٌ، صَحْرَاءُ
			Nom commun	قَلَمٌ، أَرْنَبٌ، رَجُلٌ
	Dérivation nel régulier	Conjugable	masdar	كِتَابَةٌ، الْقَتْلُ
			Participe actif	قَاتِلٌ، شَارِبٌ
			Participe passif	مَكْتُوبٌ، مَضْرُوبٌ
			Nom d'une fois	جَلْسَةٌ، ضَرْبَةٌ
			Nom de manière	نَظَرَةً، جَلْسَةً
			Nom de temps	مَغْرَبٌ
			Nom de lieu	مَكْتَبٌ، مَقْبَرَةٌ
			Nom d'instrument	مَطْرَقَةٌ، مَسْمَارٌ
			Adjectif	حَسَنٌ، جَمِيلٌ، بَاطِلٌ
			Elatif	أَحْسَنُ، أَفْضَلُ
			Nom diminutif	كُتَيْبٌ، شَوَيْعِرٌ
			Nom de relation	تُونِسِيٌّ، مِصْرِيٌّ
			intensif	قَتَالٌ، غَوَاصٌ

Tableau 1.3 Classement des sous catégories de noms.

2.3 Morphologie arabe

2.3.1 Principe

La morphologie est un domaine de la langue qui permet la description des règles régissant la structure interne des mots (unités lexicales), chez les grammairiens la morphologie est l'étude des formes des mots (flexion et dérivation), en d'autres termes, la morphologie est l'étude des mots considéré isolément (hors contexte) sous le double aspect de la nature et des variations qu'ils peuvent subir . En langue arabe, l'analyse morphologique est d'autant plus importante que les mots sont fortement agglutinés, c'est-à-dire qu'ils sont formés dans leur majorité par assemblage d'unités lexicales et grammaticales élémentaires.

Ainsi Le traitement morphologique est considéré comme une introduction principale à compréhension globale d'une langue naturelle ; il joue un rôle très important aussi bien du côté linguistique que du côté technique [34]

2.3.2 Objective

La plupart des études faites sur la morphologie arabe dans le passé ou bien aujourd'hui visent généralement à satisfaire les points suivants:

- La formation de nouveaux mots à partir des éléments lexicaux disponibles.
- L'analyse des mots réellement existant.
- Fournir les données nécessaires aux travaux des différents niveaux de traitement (syntaxe, sémantique et pragmatique). [34]

Les éléments essentiels de la morphologie de la langue arabe sont :

- **Les schèmes**

Le schème est un mot composé de trois consonnes ف [f], ع[a], et ل [l], qui sont vocalisées et qui peuvent être augmentées par d'autres lettres (préfixe, suffixe et infixes). Le schème joue un rôle très important dans le processus de génération des formes dérivées à partir d'une racine. Ce processus de génération consiste à remplacer racine du schème par les consonnes de la racine en question, tout en gardant les mêmes voyelles et les mêmes lettres augmentées tout en respectant le même ordre des consonnes, autrement dit le schème peut être considéré comme une moule sur laquelle coule la racine [3]

• Les racines

Les racines sont à l'origine de la plupart de mots arabes. Elles sont des verbes formés de trois à cinq lettres consonnes. Elles sont aux alentours de 10000 racines dont la grande majorité (85%) sont trilatérales. Les restes sont des racines quadrilatérales ou quintilatérales. Une racine définit la signification fondamentale des mots dérivés en utilisant différents diacritiques et affixes avec les lettres de la racine pour créer l'inflexion de la signification, Exemple la racine <كتب, kataba, il a écrit> plusieurs mot sans dérivé de cette racine : [1]

Racine : <كتب, Kataba, écrire >						
Conjugué	كتب	ktb	il a écrit	يكتب	yiktb	il écrit
	كتبنا	ktbna	nous avons écrit	يكتبون	yiktbwun	ils écrivent
	كتبت	ktbt	elle a écrit	تكتب	tktb	tu écris
	تكتبون	tktbwun	vous écrivez	نكتب	nktb	nous écrivons
Noms	كاتب	katb	auteur	كتابة	ktabah	écriture
	كتاب	ktab	livre	مكتوب	mktwub	lettre
	مكتب	mktb	bureau	إكتتاب	ikttab	enregistrement

Tableau 1.4 quelque dérivations de la racine (كتب, kataba, écrire).

• Les stems

Un Stem est la dérivation obtenue à partir d'une racine donnée selon un modèle. L'arabe classique à un grand nombre des Stems qui ne sont pas tous utilisables, 2% seulement sont utilisables selon Rashwan, Le Stem correspond à un schème si et seulement s'il possède le même nombre de lettres et les mêmes lettres dans les mêmes positions. Une exception est accordée aux consonnes <ف, f>, <ع, â>, <ل, l> qui sont les lettres de la racine de base <فعل, fâʿl, faire>. Par exemple, on y trouve : <مَكَاتِب, mkatb, bureaux>, il est obtenu à partir de la racine <كتب, ktb, il a écrit> selon le schème <مفاعل, mfaʿāl>. Les Stems produits ne sont pas tous utilisables [1]

• Les signes diacritiques

Les signes diacritiques sont des signes ajoutés au-dessus ou en dessous des lettres arabes afin de signifier la prononciation du mot, ce rôle phonologique influe aussi sur le sens de ce mot.

Au nombre de trois, ces symboles sont transcrire de la manière suivant :

- La fetha [a] est symbolisée par un petit trait sur la consonne (ب\ba)
- La damma [u] est symbolisée par un crochet au-dessus de la consonne (ب\bu)
- La kasra [i] est symbolisée par un petit trait sur la consonne (ب\bi)

Un petit rond symbolisant la soukoun (سكون) et apposé sur une consonne lorsque celle-ci n'est liée à aucune voyelle (بَعْدَ/baâda) [23]

- **Mots dérivés**

Selon la grammaire traditionnelle, le lexique arabe comprend Trois catégories de mots : verbes, noms, et particules. Hormis les mots propres, les mots des deux premières catégories sont dérivés à partir d'une racine. Ils sont nommés mots réguliers ou mot dérivable.

- **Mots isolés**

Les mots isolés sont les mots qui n'ont pas des racines. Les mots sont en général, les noms propres, les noms communs et les particules «Un nom propre désigne toute substance distincte de l'espèce à laquelle elle appartient. Il ne possède en conséquence aucune signification, ni aucune définition. Exemple : Paris, Jules César, Louis XIV, etc. Par contre, un nom commun est toute substance non distincte de l'espèce à laquelle elle appartient. Il est pourvu d'une signification et d'une définition. Comme <بلد, bld, pays>, <انسان, insane, une personne>, <حيوان, hayiwuan, un animal>, etc. [1]

- **Les affixes**

Les affixes sont des lettres qui s'ajoutent au début (les préfixes) ou à la fin des mots arabes (les suffixes). En général, Ils sont utilisés pour accorder aux mots des éléments syntaxiques. Ils marquent l'aspect verbal, le mode, les propriétés transitives, etc. Ils sont aux alentours de 150. [20]

Les préfixes dépendent des mots auxquels ils s'attachent. En effet, la plupart des mots arabes commencent par le préfixe <ال التعريف, al altaâryif, l'article de définition> qui est utilisé en tant que terme déclaratif. Pour cela, il y a trois types des préfixes. Premièrement, les préfixes nominaux qui sont réservés pour les noms et les adjectifs. Deuxièmement, les préfixes verbaux qui sont réservés aux verbes. Et troisièmement, les préfixes généraux qui sont utilisés indépendamment de type des mots.

Il y a deux types de suffixes, les suffixes verbaux et les suffixes nominaux. Les premiers dépendent de la transitivité et de la personne conjuguée. Les suffixes nominaux indiquent la flexion casuelle du nom (nominatif, accusatif, et génitif), le genre (masculin et féminin), le nombre (singulier, duel et pluriel), etc. [7]

شَهْدٌ	Miel (cire d'abeille)
شَهِدَ	Informé, affirmer, a été présent, a vu
شَهِدَ	A fait une déposition
شَهْدٌ	Comme رُكْعٌ
شَهِدَ	Pluriel de : شَاهِدٌ des témoins
شَهْدٌ	Nom propre féminin, plante

Tableau 1.5 Les différentes voyellations du mot « شَهِدَ »

3.3 L'extraction de la racine

Pour détecter la racine d'un mot, il faut connaître le schème par lequel il a été dérivé et supprimer les éléments flexionnels (antéfixes, préfixes, suffixes, post fixes) qui ont été ajoutés. En général des tables de préfixes et de suffixes sont utilisées, La nature agglutinative de l'Arabe rend cette tâche, assez difficile. Cette difficulté est encore plus accrue, lorsqu'il s'agit de textes non voyellés. L'analyse morphologique devra donc découper le mot et identifier des préfixes comme les conjonctions (وَ = et) et (فَ = puis), des prépositions comme (بِ = avec) et (لِ = pour), l'article défini (ال = le, la, les) et des suffixes de pronom possessif (لَهُ = à lui, لَهَا = à elle, لَهُمْ = à eux, لَهُنَّ = à elles) etc. [26]

La phase d'analyse morphologique détermine un schème possible. Les préfixes et suffixes sont trouvés en enlevant progressivement des préfixes et des suffixes et en essayant de faire correspondre toutes les racines produites par un schème afin de retrouver la racine. [9]

4 Conclusion

Contrairement aux autres langues, la langue arabe possède un système dérivationnel très riche, qui pose la difficulté de son traitement. La langue arabe se caractérise par sa directionnalité droite à gauche, sa nature cursive (agglutination des mots), ses signes de vocalisation qui s'ajoutent au-dessous et au-dessus des caractères, et par son ambiguïté due à l'absence de voyelles (cas de la majorité des textes arabes). Ces caractéristiques constituent en fait les problèmes majeurs face aux travaux effectués sur la langue arabe dans le domaine de la recherche d'information.

CHAPITRE 2

FOUILLE DE DONNEES ET FOUILLE DE TEXTES

1 Introduction

Une confusion subsiste encore entre data mining, que nous appelons en français « fouille de données », et Knowledge discovery in data bases (KDD), traduit en français par ECD (Extraction de Connaissances de Données). Le data mining est l'un des maillons de la chaîne de traitement pour la découverte des connaissances à partir des données. Sous forme imagée, nous pourrions dire que l'ECD est un véhicule dont le data mining est le moteur. [20]

Le data mining est l'art d'extraire des connaissances à partir des données. Les données peuvent être stockées dans des entrepôts (data warehouse), dans des bases de données distribuées ou sur Internet : web mining. Le data mining ne se limite pas au traitement des données structurées sous forme de tables numériques ; il offre des moyens pour aborder les corpus en langage naturel (text mining), les images (image mining), le son (sound mining) ou la vidéo et dans ce cas, on parle alors plus généralement de multimédia mining [19]

2 Fouille de données (Data mining)

2.1 Définition d'ECD (KDD)

Il est ici important de différencier les trois termes suivants :

- Donnée : valeur d'une variable pour un objet (comme le montant d'un retrait d'argent par exemple) ;
- Information : résultat d'analyse sur les données (comme la répartition géographique de tous les retraits d'argent par exemple).
- Connaissance : information utile pour l'entreprise (comme la découverte du mauvais emplacement de certains distributeurs). [36]

KDD (Knowledge Discovery in Databases) is the non-trivial process of identifying valid, potentially useful and ultimately understandable patterns in data. Fayyad (1996)

ECD (L'Extraction de Connaissances à partir des Données) est un processus itératif et interactif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par un utilisateur analyste qui y joue un rôle central. [14]

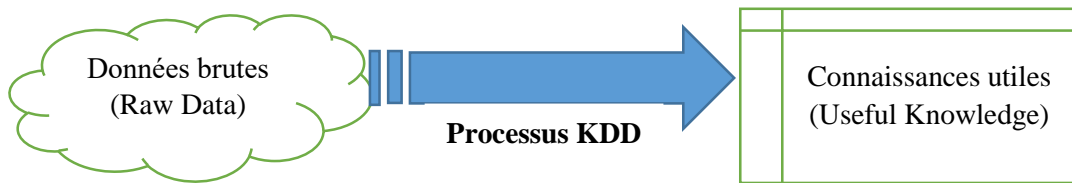


Figure 2.1 Transformation de données brutes en connaissances utiles [6]

2.2 Processus d'extraction de connaissances

Le KDD est un processus semi-automatique et itératif, constitué de plusieurs étapes allant de la sélection et préparation des données jusqu'à l'interprétation des résultats, en passant par la phase de recherche des connaissances (le data mining), les différentes étapes de ce processus sont présentées dans la figure ci-dessous :

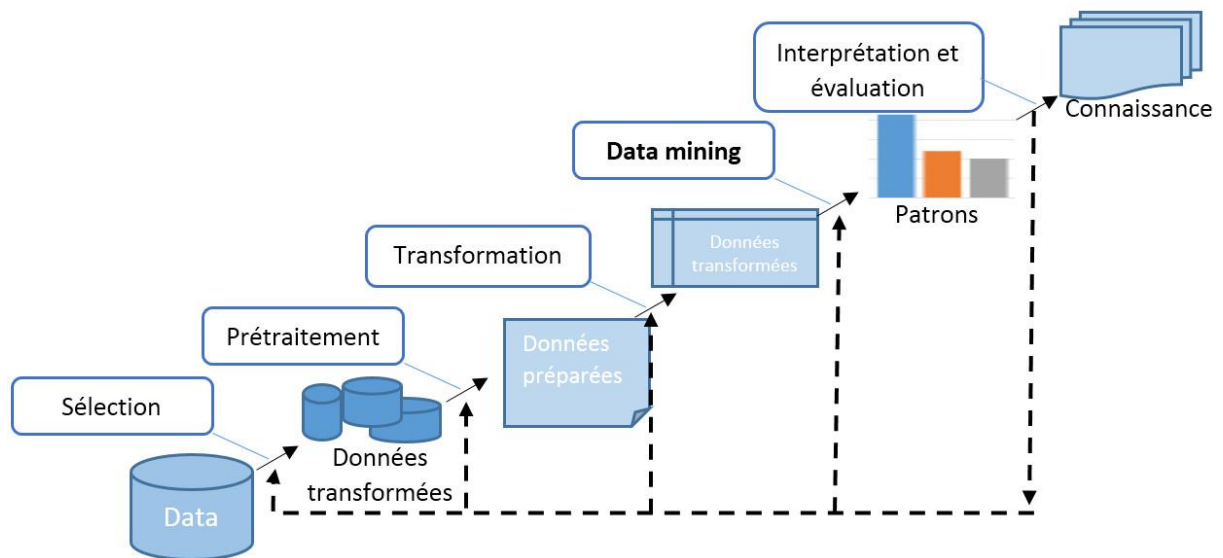


Figure 2.2 Différentes étapes du processus ECD. [19]

Ces opérations peuvent être regroupées en cinq phases majeures [Fayyad et al, 1996] :

- **Sélection de données** : le but de cette phase est l'extraction à partir d'un plus grand stock de données seulement celles qui sont appropriées à l'analyse d'exploitation de données. Cette extraction de données aide à rationaliser et accélérer le processus.
- **Préparation de données** : cette phase de KDD est concernée par les données nettoyant et les tâches de préparation qui sont nécessaires pour assurer des résultats corrects.
- **Transformation de données** : Les données sélectionnées dans l'étape précédente vont subir une transformation dont le but est de les rendre dans une forme appropriée pour les méthodes et les techniques de Data mining.

- **Data mining** : le but de la phase d'exploitation de données est d'analyser les données par un ensemble approprié d'algorithmes afin de découvrir les modèles et les règles significatifs et produire les modèles prédictifs. C'est l'élément de noyau du cycle de KDD.
- **Interprétation et Evaluation** : tandis que les algorithmes de data mining ont le potentiel de produire un nombre illimité de modèles cachés dans les données, beaucoup de ces derniers peuvent ne pas être significatifs ou utiles. Cette phase finale est visée choisissant ces modèles qui sont valides et utiles pour prendre de futures décisions économiques. [19]

2.3 Définitions de la fouille de données

Le Data mining est un processus non trivial qui consiste à identifier, dans des données, des schémas nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables. [14]

Le Data mining est L'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de grandes bases de données informatiques, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données permettant d'étayer des prises de décisions. [19]

2.4 Pourquoi la fouille de données ?

La fouille de données est une discipline en vogue. Ce n'est cependant pas un mode ni une coquetterie, l'appel croissant et varié à la fouille de données tient aux facteurs suivants :

- Augmentation des capacités de stockage des données (disques durs de giga octets).
- Augmentation des capacités de traitements des données (facilité d'accès aux données : il n'y a plus de bandes magnétiques ; accélération des traitements).
- Maturation des principes des bases de données (maturation des bases de données relationnelles).
- Croissance explosive de la collecte des données (scanners de supermarché, internet, etc.)
- Plus grande disponibilité des données grâce aux réseaux (intranet et internet).
- Développement de logiciels de data mining

2.5 Les domaines d'application de la fouille de données

Le data mining utilise le savoir-faire de plusieurs domaines : l'intelligence artificielle, l'apprentissage, la reconnaissance du modèle, l'acquisition de connaissances pour les systèmes experts, la visualisation des données, etc. la plupart des algorithmes du data mining sont dérivés

de ces domaines. Les point commun entre tous ces algorithmes est l'extraction de modèles dans un contexte de grandes quantités des données. [10]

Voici une liste non exhaustive des applications possibles du datamining par secteur d'activités :

- **Grande distribution et VPC** : Analyse des comportements des consommateurs, recherche des similarités des consommateurs en fonction de critères géographiques ou sociodémographiques, prédiction des taux de réponse en marketing direct, vente croisée et activation sélective dans le domaine des cartes de fidélité, optimisation des réapprovisionnements.
- **Laboratoires pharmaceutiques** : Modélisation comportementale et prédiction de médications ou de visites, optimisation des plans d'action des visiteurs médicaux pour le lancement de nouvelles molécules, identification des meilleures thérapies pour différentes maladies.
- **Banques** : Modélisation prédictive des clients partants, détermination de pré-autorisations de crédit.
- **Assurance** : Modèles de sélection et de tarification, analyse des sinistres, recherche des critères explicatifs du risque ou de la fraude, prévision d'appel sur les plates-formes d'assurance directe.
- **Aéronautique, automobile et industries** : Contrôle qualité et anticipation des défauts, prévision des ventes, dépouillement d'enquêtes de satisfaction.
- **Télécommunications, eau, énergie** : Simulation de tarifs, détection de formes de consommation frauduleuses, classification des clients selon la forme de l'utilisation des services, prévision de ventes.

Comme on peut le voir, le datamining peut s'appliquer à tous les domaines.

2.6 Difficultés de data mining

La mise en œuvre de data mining rencontre trois difficultés principales :

- **Qualité des données** : Les statistiques ont montré que 60% à 70% du temps de travail dans un projet de data mining est consacré au prétraitement des données (sélection, correction, transcodage, chargement...), ce qui montre que le temps de préparation est un inconvénient majeur qui influe sur le temps global du projet.
- **Choix des algorithmes** : Pour pouvoir répondre aux questions qui se posent, les algorithmes doivent être choisis en fonction du problème traité. Il faut que l'expert en data mining soit aussi

un animateur et possède des qualités que l'on trouve rarement ensemble chez la même personne: rigueur dans la méthode, ouverture et chaleur humaine dans la communication.

- **Evaluation des résultats** : Avant de procéder au déploiement final du modèle, il est important de l'évaluer plus complètement et de passer en revue toutes les différentes étapes exécutées pour construire ce modèle. Ceci permettra d'être certain qu'il permet d'atteindre les objectifs fixés. Lors de la définition du problème, un objectif principal est de déterminer s'il y a un aspect important du problème à résoudre qui n'a pas été suffisamment considéré. A la fin de cette phase, une décision sur l'utilisation des résultats fournis par les outils de data mining devrait être prise. [31]

2.7 Tâches de fouille de données

Des nombreuses tâches peuvent être associée au data mining parmi ces tâches nous pouvant citer :

- **La classification** : Etant donné un ensemble prédéfini de classes d'objets, affecter un objet à une classe, selon une certaine mesure de proximité est le rôle de la classification. Les techniques de classification commencent par définir un plan d'expérience ou un ensemble de données d'apprentissage sur lequel on applique les méthodes de classification. Puis, pour mesurer leur pouvoir de classement correct, on applique les mêmes méthodes sur un jeu d'essai (testing set).

- ✓ Des exemples de tâche de classification sont :

- Attribuer ou non un prêt à un client.
- Établir un diagnostic,
- Accepter ou refuser un retrait dans un distributeur, [18]

- **L'estimation** : Elle consiste à estimer la valeur d'un champ à valeurs continues à partir des caractéristiques d'un objet. L'estimation peut être utilisée dans un but de classification. Il suffit d'attribuer une classe particulière pour un intervalle de valeurs du champ estimé.

- ✓ Des exemples de tâche d'estimation sont :

- Noter un candidat à un prêt ; cette estimation pourra trouvera une application pour attribuer un prêt (classification),
- par exemple, en fixant un seuil d'attribution,
- Estimer les revenus d'un client.

- **La prédiction** : Cela consiste à estimer une valeur future. En général, les valeurs connues sont classées chronologiquement. On cherche à prédire la valeur future d'un champ. Cette tâche

est proche des précédentes. Les méthodes de classification et d'estimation peuvent être utilisées en prédiction.

✓ Des exemples de tâche de prédiction sont :

- Prédire les valeurs futures d'actions,
- Prédire au vu de leurs actions passées les départs de clients. [6]

- **La segmentation** : Il s'agit de créer des groupes homogènes dans la population (l'ensemble des enregistrements). Il appartient ensuite à un expert du domaine de déterminer l'intérêt et la signification des groupes ainsi constitués. Cette tâche est souvent effectuée avant les précédentes pour construire des groupes sur lesquels on applique des tâches de classification ou d'estimation.

- **L'association** : Elle consiste à induire des corrélations entre les données. Très répandue dans le secteur de la distribution car leur principale application est « l'analyse du panier de la ménagère » qui consiste en la recherche d'associations entre produits sur les tickets de caisse. Trouver tous les articles achetés ensemble et ceux qui ne sont jamais achetés ensemble dans un supermarché est un exemple de la fonction d'association. [19]

3 Fouille de textes

La fouille de données textuelles est la branche de la fouille de données qui offre des moyens capables de sélectionner, d'analyser, et d'extraire les formations textuelles non structurées en langage naturel.

3.1 Définitions de fouille de textes

La fouille de textes est un ensemble de processus permettant, à partir d'un ensemble de ressources textuelles, de construire des connaissances pouvant être représentées dans un langage formel de représentation de connaissances et exploitées pour raisonner sur le contenu des textes [27]

On appelle fouille de texte un ensemble de techniques d'analyse linguistique, essentiellement statistique, visant à faire émerger des relations, a priori inconnues, entre éléments de connaissance (matérialisés par des séquences textuelles). La fouille de texte est particulièrement utilisée pour la veille, où la découverte de nouvelles connaissances à partir de texte est primordiale. [8]

Schématiquement, on peut énoncer :

Fouille de texte =linguistique +fouille de données

3.2 Les objectifs de la fouille de textes

La fouille de textes est utilisée pour :

- mieux comprendre le positionnement d'un discours, d'une thèse, d'un communiqué,
- appréhender les thèmes récurrents qui sont associés à une activité, une entreprise ou des concurrents,
- mesurer les points faibles et les points forts dans une revue de presse,
- comparer des textes sur un même thème afin d'en déterminer les points communs ou au contraire de distinguer les différences stylistiques,
- créer automatiquement des répertoires de sites Web ou emails associés à des thématiques....
- Quantifier un texte ou les parties d'un texte pour en extraire les structures significantes les plus fortes (Résumé automatique, Segmentation thématique),
- Etablir des règles de classification automatique de documents (Classification, Clustering).
- Etablir des liens entre les termes et les documents (Indexation), [8]

3.3 Tâches de la fouille de texte

Une tâche, au sens informatique, est la spécification d'un programme qui mime une compétence précise d'un être humain. Dans celui-ci, on commencera donc par analyser les composantes qui rentrent dans la définition d'une tâche de fouille de textes quelconque à l'aide d'un schéma général. On passera ensuite en revue les principales tâches "élémentaires" qui seront détaillées par la suite, en montrant comment elles rentrent dans ce schéma général. On montrera aussi que, via recodage de leurs données ou reformulation de leur objectif, ces tâches élémentaires sont en fait très liées les unes aux autres, par exemple que certaines d'entre elles permettent d'en simuler certaines autres. On expliquera aussi comment, en combinant plusieurs, on peut parvenir à en réaliser d'autres plus complexes. Enfin, on s'attardera sur un prétraitement des textes qui s'avère indispensable pour plusieurs des tâches élémentaires abordées ici, qui consiste à les transformer en un tableau de nombres. Cette étape préliminaire permet d'appliquer sur les textes les techniques directement issues de la fouille de données, qui s'est spécialisée dans la manipulation de tels tableaux. [37]

3.3.1 La notion de tâche et ses composantes

- **Schéma général d'une tâche**

Dans la suite de ce document, nous nous efforcerons de garder le même mode de description pour chacune des tâches abordées. Ce mode est synthétisé dans le schéma très simple de la figure suivant :

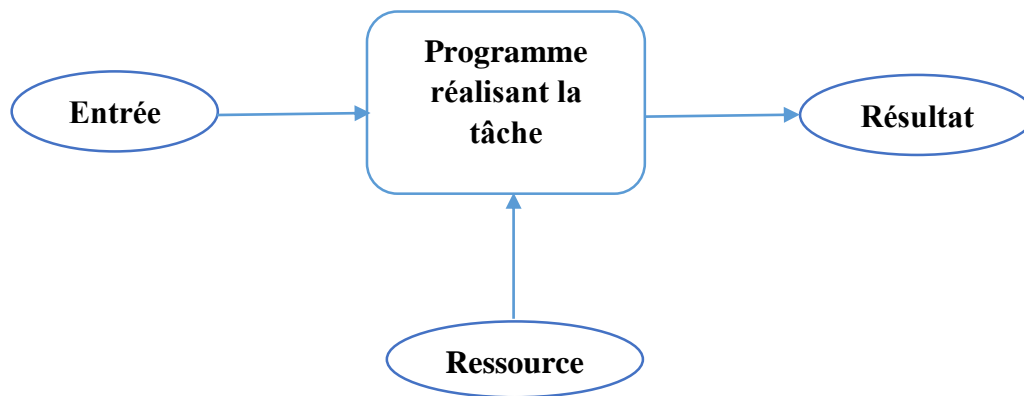


Figure 2.3 Schéma général d'une tâche de fouille de textes.[37]

Dans ce schéma, les données figurent dans des ovales tandis que le programme réalisant la tâche est matérialisé par un rectangle. C'est bien sûr dans les différentes données que la spécificité de la fouille de textes se manifestera : tout ou parties d'entre elles seront de nature textuelle, ou en découleront après un prétraitement.

Ce schéma est très simple, mais nous verrons qu'il oblige tout de même à se poser quelques bonnes questions. Par exemple, il n'est pas toujours facile de distinguer ce qui joue le rôle de données d'entrée ou de ressources dans la définition d'une tâche.

Un bon critère serait le suivant : une ressource est une donnée stable, qui n'est pas modifiée d'une exécution du programme à une autre, alors que la donnée d'entrée, elle, change chaque fois. Certaines ressources sont obligatoires dans la définition de certaines tâches, d'autres facultatives. C'est souvent par ce biais que des connaissances externes et générales peuvent être intégrées au processus de réalisation de la tâche. Les ressources sont donc un des principaux leviers pour faire rentrer un peu de linguistique dans le domaine de la fouille de textes. C'est le cheval de Troie des linguistes. [37]

3.4 Processus de fouille de textes

Un système de fouille de données textuelles reprend les étapes du processus du Data Mining et il en ajoute d'autres pour les adapter à son objectif. Quelques systèmes de fouille de données textuelles ont pour objectif de structurer le contenu des textes en découvrant des modèles pour les écrire. Ils se basent sur l'hypothèse d'une catégorisation a priori où il s'agit d'un prétraitement manuel des textes afin d'en extraire un certain nombre d'attributs comme les mots-clés ou les URLs. Une fois les attributs extraits, les méthodes classiques de Mining, telles que l'analyse statistique, association, etc.

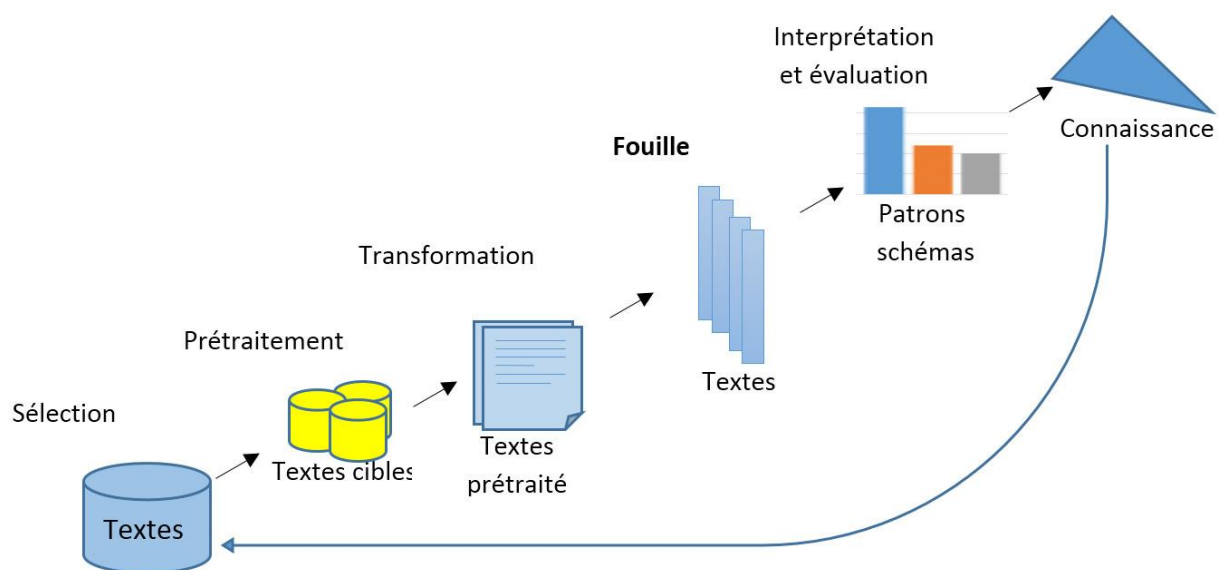


Figure 2.4 Processus de la fouille de données textuelles. [11]

Dans la suite, les étapes de sélection, de prétraitement et d'indexation des textes orientent, globalement le processus de fouille. Ces étapes ont une grande influence sur la qualité des connaissances extraites à partir des textes. De plus, les textes étudiés ont été indexés automatiquement puis nettoyés et nous montrons comment le résultat du processus de fouille peut améliorer in fine cette indexation.

- **Sélection :** Cette première phase concernant la recherche ou la collecte du corpus homogène sur un thème donné ; il est important de noter que cette étape constitue la brique de base de l'ensemble du processus de fouille de textes. Le succès du processus est fonction de la qualité et de l'homogénéité du corpus collecté. La constitution du corpus est donc confiée explicitement à un expert.

Dans la suite de nos travaux, le corpus étudié présente des critères de qualité et d'homogénéité nous permettant d'appliquer nos outils de fouille de textes.

- **Le prétraitement des textes :** Il existe de nombreux types de données à partir desquels des informations pertinentes pour une tâche donnée peuvent être extraites. Cependant, quel que soit le type de données, il est nécessaire de prétraiter les données brutes afin de pouvoir ensuite les traiter. Ce type de traitements prend actuellement le nom de normalisation. Les prétraitements des données textuelles consistent à normaliser les diverses manières d'écrire un même mot, à corriger les fautes d'orthographe évidentes ou les incohérences typographiques et à expliciter certaines informations lexicales¹ exprimées implicitement dans les textes.
- **Transformation :** Dans la mesure où l'analyste n'avait pas formulé d'objectifs de fouille particuliers, nous avons adopté une modélisation du contenu des résumés de textes dont la sémantique est simple : chaque résumé est modélisé par l'ensemble des termes qu'il possède. C'est une indexation contrôlée à partir d'une liste de termes attestée. Cette indexation terminologique contrôlée permet d'associer à un groupe de mots, un concept i.e. Une notion participant à une base de connaissance et permet ainsi de passer d'un élément de nature linguistique à un élément de nature connaissance. Elle présuppose simplement que la cooccurrence de termes dans un même texte reflète une certaine proximité sémantique entre ces termes. On y retrouve donc une forte densité de termes, un maximum de contenu informationnel et un minimum d'information inutile.
- **Fouille de données :** La fouille de données est le cœur du processus car elle permet d'extraire des connaissances à partir des données. C'est souvent une étape difficile à mettre en œuvre, coûteuse et dont les résultats doivent être interprétés et relativisés. Dans cette phase, des techniques du data mining sont utilisées : association ; classification...etc.
- **Interprétation et évaluation :** Enfin, cette étape identifie les modèles intéressants représentant les connaissances, en se basant non seulement sur des mesures d'intérêt mais aussi sur l'avis de l'expert. [12]

3.5 Domaine d'application de la fouille de texte

Les applications de la fouille de données textuelles sont multiples : d'une simple indexation pour les moteurs de recherche à l'extraction de connaissances dans des documents non structurés. Les applications principales sont :

- **La recherche d'information (RI) :** s'intéresse aux documents dans leur globalité et aux thèmes qu'ils abordent pour comparer les documents et détecter les typologies. Comme Moteurs de recherche de type Google, Détection de plagiat...

- **Extraction d'Information (EI) :** L'extraction d'information «RI» consiste en l'alimentation d'une base de données structurée à partir de données exprimées en langage naturel. Il s'agit de détecter dans le texte en langage naturel les mots correspondant à chaque champ de la base de données. L'analyse est donc locale et l'extraction de l'information est plus complexe, car elle nécessite d'effectuer une analyse lexicale et une analyse morpho-syntaxique pour reconnaître les constituants du texte (phrases, mots, verbes, adjectifs) qui permettent de détecter les phrases pertinentes pour l'extraction. Comme Veille d'information basée par exemple sur les "who did what, where and when, and why", Bibliométrie (réseaux de citations)
- **La classification de documents :** Sert à classer des documents similaires. Comme Tri d'email (détection de SPAM), Fouille d'opinions (évaluations positives ou négatives d'un produit, service, etc.), Surveillance et détection de menaces (espionnage, etc.)
- **Le traitement automatique de la langue « TAL » :** Depuis une vingtaine d'années, et avec la généralisation de l'outil informatique et d'internet, les applications du TAL au sens large du terme se multiplient dans les disciplines philologiques. Le TAL est une discipline à la frontière de la linguistique et de l'informatique. Le TAL concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain. On distingue donc usuellement cinq niveaux linguistiques de profondeurs successifs d'analyse automatique des langues : segmentation de phrase «analyse lexicale», décodage acoustico-phonétique ou «traitement de parole», analyse morphologique, analyse sémantique et analyse pragmatique. [11]

4 Conclusion

La fouille de données est l'exploration et l'analyse de grandes quantités de données afin d'y découvrir de l'information implicite. Cette information peut être de différente nature.

Avec le développement du web et la grande quantité des documents disponibles de nos jours ; les documents textuels non structurés sont devenus prédominants. L'information utile étant enfouie dans le texte, il devient alors indispensable de proposer de nouveaux systèmes permettant l'analyse, l'organisation et la représentation des différents contenus textuels.

CHAPITRE 3

CORPUS DES HADITHS

1 Introduction

Le monde aujourd'hui s'interroge sur la nécessité d'une éthique pour encadrer les développements techniques et d'une loi pour organiser les nouveaux réseaux familiaux et le nouveau portrait des familles. D'une part, l'Islam organise les rapports de l'homme avec Dieu et, d'autre part, les rapports de l'homme avec ses semblables. Il n'est pas seulement une religion, mais également une conduite ou une loi. La Charia constitue un système de droit autonome et s'impose à tout musulman puisqu'elle émane de la volonté divine. A l'origine, la Charia signifie chemin vers l'eau. En Islam, ce sont le Coran et la Sunna qui constituent la référence suprême et globale de la loi islamique.

Le Coran est le recueil des paroles que le Prophète (que la prière et la paix d'Allah soient sur lui) a reçues en état de Révélation (6235 versets). La Sunna (encore appelée Hadith ou Tradition Prophétique) est un immense corpus littéraire qui s'est cristallisé aux 3ème/9ème siècle après une longue période d'élaboration. Dans ce chapitre on va traiter des concepts liés aux corpus prophétiques.

2 Hadith

Le mot Hadith à plusieurs sens : nouvelles, histoire, communication, conversation rapport. Dans le contexte religieux islamique, cela signifie un rapport individuel d'une action, d'instruction ou de dire, du Prophète, ou son approbation, désapprobation, ou le silence (approbation tacite) en ce qui concerne une question ou action. Par sa nature même, la fiabilité des rapports du Hadith dépend de la compréhension par le rapporteur du contexte et des mots et de leur applicabilité. Indépendamment de la complexité, cependant, le Hadith est la deuxième principale source d'orientation islamique. En existe autres mots ont été également utilisés dans le même sens, tel que al-khabar et al-athar [38].

Quoi qu'il en soit, le hadîth explique le Coran, interprète ses lois et établit des prescriptions non exprimées dans le texte coranique : les deux se complètent pour la compréhension de la parole divine. [39]

L'ensemble des hadîths représente un corpus de traditions qui constitue la sunna (le chemin) du Prophète.

2.1 La Sunna (Sunnah) :

"سنة الله في الدين خلوا من قبل ولن تجد لسنة الله تبديلا"

« Telles était la loi établie par Allah envers ceux qui ont vécu auparavant et tu ne trouveras pas de changement dans la loi (sounna) d'Allah » .«*Sourate (el-Ahzabe) 33-vaste 62*»

La Sunna (Sunnah) : *tradition*, Ce terme désigne la tradition islamique tirée de l'exemple de la vie du prophète Mohammed (sur lui la prière et la paix). Ainsi elle est la seconde source, après le Noble Coran, de la connaissance des statuts religieux pour le musulman, elle vient compléter et préciser le sens du message Coranique. Donc, le Hadîth est le moyen de connaître la Sunna et il est considéré comme la seconde source de la Législation islamique, après le Noble Coran. Le mot Sounna a été employée par les Arabes du pré islam dans le sens de conduite (sîrah) ou méthode (tarîqa).

2.2 Composants du hadith

Un hadith se compose de trois parties, comme on le verra :

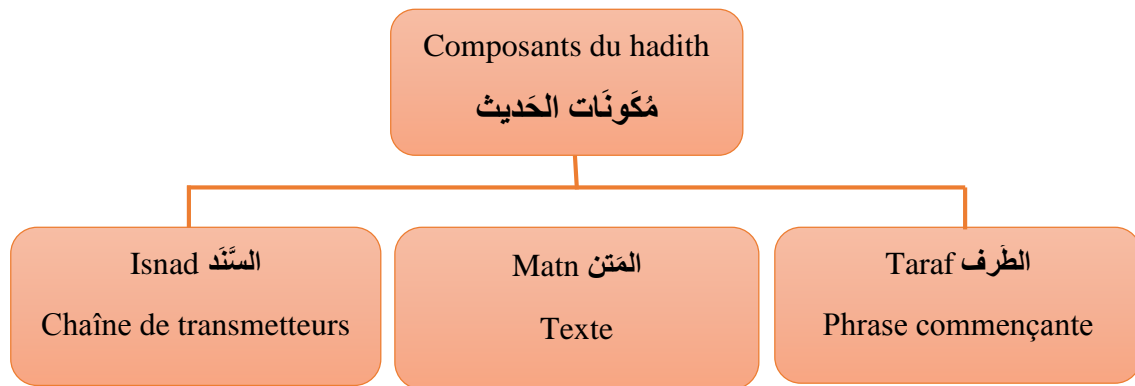


Figure 3.1 Composants du hadith. [29]

- **Matn** : est le texte ou le contenu du hadîth ; il se réfère à ce que le prophète a dit réellement ou faire.
- **Isnad** : la chaîne des transmetteurs du texte du hadîth à garantisiez l'authenticité et l'exactitude verbale de chaque énonciation.
- **Taraf** : est la partie ou la phrase commençante du texte qui fait référence à la parole, à l'action ou à la caractéristique du prophète (que la prière et la paix d'Allah soient sur lui), ou de son accord donné à d'autres actions.

L'authenticité du hadith dépend du sérieux de ses rapporteurs, et de la liaison entre eux. [29]

Exemple

حدثنا الحميدي عبد الله ابن الزبير قال: حدثنا سفيان قال: حدثنا يحيى بن سعيد الأنصاري قال أخبرني محمد ابن إبراهيم التيمي: انه سمع علقمة ابن وقاس الليثي يقول: سمعت عمر ابن الخطاب رضي الله عنه على المنبر قال، سمعت رسول الله صلى الله عليه وسلم يقول: "إنما الأعمال بالنيات وإنما لكل امرئ ما نوى، فمن كانت هجرت إلى دنيا يصيبها، أو إلى امرأة ينكحها، فهجرته إلى ما هاجر إليه"

- Source : *Sahîh Bukhârî*
- Livre : *commencement de l'inspiration*
- Chapitre : *commencement de l'inspiration*
- N° du hadith : 1

Information du hadith

حدثنا الحميدي عبد الله ابن الزبير قال: حدثنا سفيان قال: حدثنا يحيى بن سعيد الأنصاري قال: أخبرني محمد ابن إبراهيم التيمي: انه سمع علقمة ابن وقاس الليثي يقول: سمعت عمر ابن الخطاب رضي الله عنه على المنبر قال، سمعت رسول الله صلى الله عليه وسلم يقول:

Isnad

Taraf

"إنما الأعمال بالنيات، وإنما لكل امرئ ما نوى، فمن كانت هجرت إلى دنيا يصيبها، أو إلى امرأة ينكحها، فهجرته إلى ما هاجر إليه"

Matn

Figure 3.2 Exemple de composants hadith.[10]

2.3 Classification des Hadiths

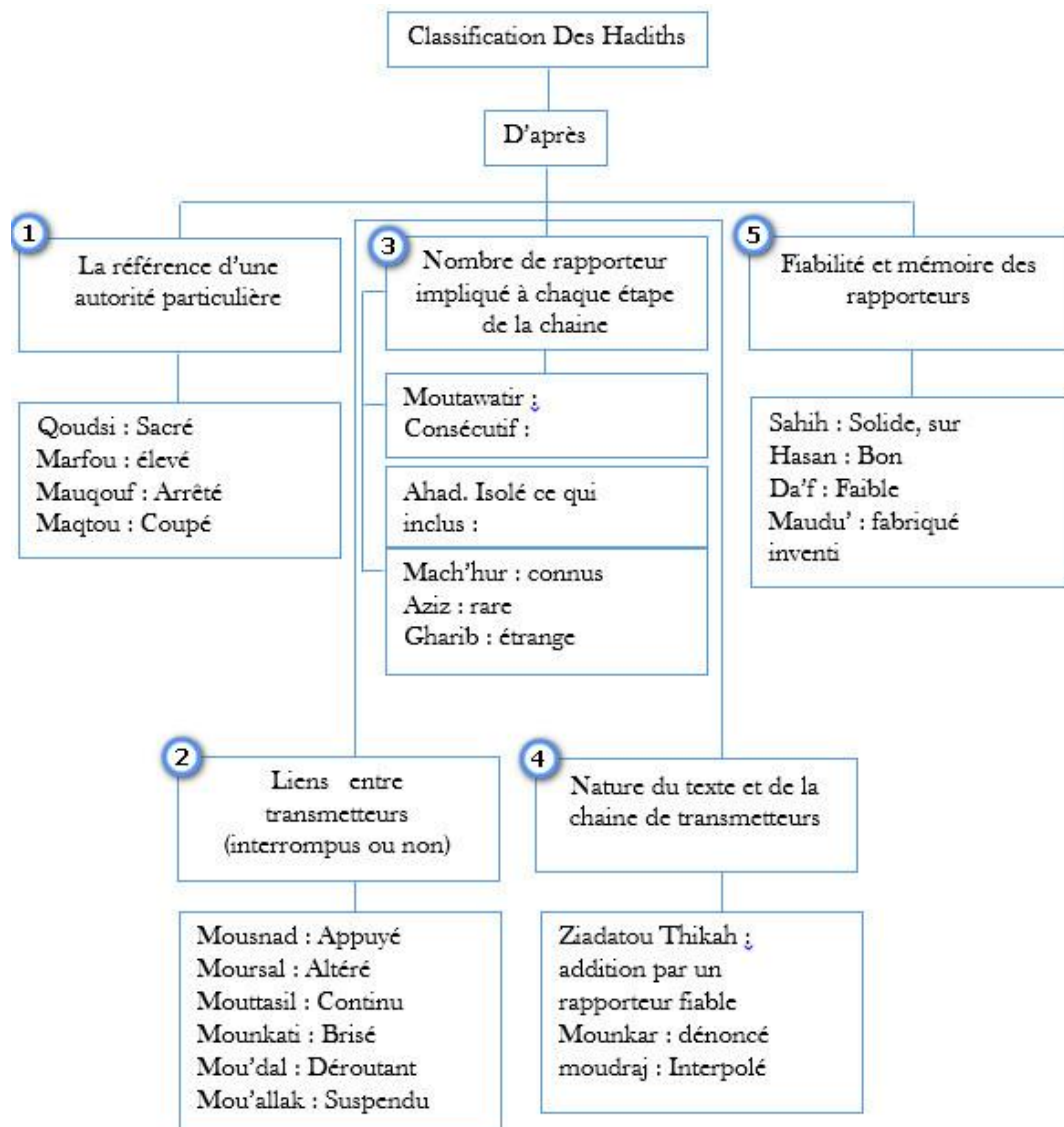


Figure 3.3 Classification des hadiths. [10]

1) Selon la référence d'une autorité particulière

- Qoudsi- Divin : une révélation de Dieu, transmis par relais des mots du prophète
- Marfou- élevé : un récit du prophète, commençant par exemple par : "J'ai entendu le prophète dire..."
- Mauqouf- arrêté : un récit rapporté par un seul compagnon, commençant par exemple par : "On nous a ordonné de..."
- Maqtou'- divisé : un récit émanant du premier successeur du compagnon.

2) Selon la chaîne de transmission (Isnad), interrompue ou non

- Mousnad - supporté : un hadith qui a été rapporté par un traditionaliste, basé sur ce qu'il a appris de son professeur à une époque appropriée à l'étude ; de même pour chaque professeur jusqu'à ce que la chaîne atteigne un compagnon bien connu, qui de son côté, rapporte des propos du prophète.
- Moutassil- continu : un hadith avec une chaîne ininterrompue qui va jusqu'à un compagnon ou un de ces successeurs.
- Moursal- altéré : si le lien entre le successeur et le prophète est manquant, par exemple quand le successeur dit " le prophète a dit...".
- Mounkati- cassé : ce dit d'un hadith dont le lien à n'importe quel endroit de la chaîne avant le successeur est manquant.
- Mou'adal- perplexe : ce dit d'un hadith dont le rapporteur omet deux (ou plus).rapporteurs de la chaîne.
- Mou'allaq- arrêté : ce dit d'un hadith dont le rapporteur omet toute le chaîne de transmission et cite directement le prophète directement.

3) Selon le nombre de rapporteurs impliqués dans chaque étape de la chaîne de transmission

- Moutawatir- Consécutif : ce dit d'un hadith qui est rapporté par un si grand nombre de personnes qu'il est impossible qu'ils se soient concertés pour convenir d'un mensonge.
- Ahad- isolé : ce dit d'un hadith qui est relaté par un nombre important de personnes mais dont le nombre n'atteint pas celui du moutawatir. Il est encore divisé en :
 - Mash'hur- célèbre : le hadith a été mémorisé par plus de deux rapporteurs.
 - Aziz- rare, fort : à n'importe quelle étape de la chaîne, seulement deux rapporteurs relate le hadith.
 - Gharib - étrange : à un certain moment de la chaîne, seulement un rapporteur relate le hadith.

4) Selon la nature du texte et de la chaîne

- Mounkar- dénoncé : ce dit d'un hadith qui est rapporté par un narrateur faillible, et dont le récit va à l'encontre d'un hadith authentique.
- Moudraj- interpolé : un ajout au texte du hadith par un rapporteur.

5) Selon le sérieux et la mémoire des rapporteurs

- Sahih- Sûr, solide. L'imam Al-shafi'i indique les obligations pour un tel hadith, qui n'est pas moutawatir, afin qu'il soit acceptable : "Chaque rapporteur doit être digne de confiance dans sa religion ; il devra être connu pour être véridique dans son récit, de comprendre ce qu'il rapporte, savoir comment une expression différente peut modifier la signification du hadith, et de rapporter les mots du hadith in extenso, et pas seulement au niveau de sa signification".
- Hasan - bon : c'est celui où sa source est connue et ses rapporteurs digne de confiance.
- Da'if- faible : ce dit d'un hadith qui n'atteint pas le statut de hasan. Habituellement, la faiblesse est :
 - une discontinuité dans la chaîne, dans ce cas le hadith pourrait être - selon la nature de la discontinuité - mounqati (cassé), mou'allaq (arrêter), mou'dal (perplexe), ou moursal (altéré),
 - un des rapporteurs ayant un caractère suspect, par exemple en raison de ses mensonges, erreurs excessives, opposition au récit des sources plus fiables, participation dans l'innovation, ou ambiguïté entourant sa personne.
- Maudou'- fabriqué ou forgé : ce dit d'un hadith dont le texte va à l'encontre des normes établies pour les paroles du prophète, ou la chaîne comprend un menteur. Un hadith fabriqué peut également être identifié par une anomalie présente à une époque particulière (rébellion, etc.)

2.4 Collections importants de hadith

Beaucoup de collections ont été étaient fragmentaires et ont été entreprises pour des buts spéciaux. La plupart ont survécu comme parties d'arguments juridiques et spirituels ou ont été incorporées dans les collections plus complètes. Des collections importantes et systématiques ont été faites vers la fin de la deuxième et le début des troisièmes siècles de Hijri :

- Le Mu'watta de Malik est le plus tôt. Il était né à Madinah : « de 93H à 179H »
- Le Musnad d'Ahmad est prochain. Il était né à Bassora : « de 164H à 241H »

Mais le « Sihah Sittah » (littéralement, le six le plus rigoureusement authentifié) est :

1. Le Sahih de Bukhari. Il était né à Boukhara : « de 194H à 256H »
2. Le Sahih des musulmans. Il était né à Nishapur : « de 204H à 261H »
3. Le Sunan d'Abu Dawud. Il était né à Sajistan : « de 202H à 275H »
4. Le Sunan de Tirmidhi. Il était né à Khurasan : « de 209H à 279H »

5. Le Sunan d'An-Nisa'i. Il était né à Khurasan : « de 214H à 303H »
6. Le Sunan d'Ibn Majah : « de 209H à 273H » [39]

2.5 Imam Al-Bukhârî

قال فيه الإمام مسلم: "والله ما يغيضك إلا حاسد اشهد ان ليس في الدنيا مثلك"

L'imam Al-Bukhârî naquit le 13ème jour du mois de Chaoual 194 de l'année hégirienne, le territoire Khouârâsân (actuellement ville de la république d'Ouzbékistan). Imâm al-Bukhari (rahimahullah) est connu comme le Commandeur des croyants dans hadîth. Sa généalogie est la suivante : Abu Abdullah Muhammad Ibn Ismâ'il Ibn Ibrahim Ibn al-Mughirah Ibn Bardizbah al-Bukhârî. Son père mourut alors qu'il n'était encore qu'un jeune enfant et c'est sa mère qui l'éleva. A l'âge de dix ans, il commença à acquérir la connaissance du hadîth. Il voyagea à Makkah (La Mecque) à l'âge de seize ans accompagné par sa mère et son frère aîné. Il visita Bagdad plusieurs fois et y rencontra beaucoup de savants y compris l'Imam Ahmad Ibn Hanbal. De par son honnêteté et sa gentillesse et le fait qu'il était digne de confiance, il était à l'écart des princes et des souverains de crainte qu'il ne soit amené à tordre la vérité pour leur faire plaisir. Les récits sur la persévérance de l'Imam al-Bukhârî à rassembler les hadîths sont nombreux. Il ne cessa de voyager vers l'un ou l'autre des territoires islamiques pour rassembler les précieux propos du noble Prophète Muhammad (que la prière et la paix d'Allah soient sur lui). On dit que l'Imam al-Bukhârî avait rassemblé plus de 300 000 hadîths mémorisant 200 000 dont quelques-uns étaient peu fiables. Il naquit au temps où on falsifiait le hadîth pour faire plaisir aux souverains et aux rois ou pour corrompre la religion de l'Islam. On dit aussi que l'Imam al-Bukhârî (avant qu'il ne rassemble les hadîths dans le Sahîh al-Bukhârî) avait vu dans un rêve, comme s'il était debout devant le Prophète Muhammad (que la prière et la paix d'Allah soient sur lui) portant un chasse-mouches à la main qu'il utilisait pour chasser les mouches autour du Prophète (que la prière et la paix d'Allah soient sur lui). L'Imam al-Bukhârî demanda l'interprétation de ce rêve aux connaisseurs, et ils répondirent qu'il chasserait les mensonges attribués au Prophète (que la prière et la paix d'Allah soient sur lui). Ainsi, il avait la grande tâche de trier les hadîths, de séparer les falsifiés des authentiques et pour cela il travaillait jour et nuit. Toutefois, malgré le grand nombre de hadiths qu'il avait mémorisés, il n'en choisit que 7275 avec répétition et 2230 sans répétition, hadîths dont l'authenticité n'était sujette à aucun doute.

À chaque fois qu'il était sur le point d'enregistrer un hadîth, il faisait les ablutions, effectuait une prière de deux Râkasa et suppliait Allah. Nombre de savants musulmans ont essayé de trouver une faille dans cette grande et remarquable collection de Hadîths réunis dans Sahîh al-Bukhârî, mais sans succès. C'est pour cette raison qu'il est établi sans aucun doute que le livre le plus authentique après le Livre d'Allah est Sahîh al-Bukhârî. L'Imam al-Bukhârî mourut le premier jour du mois de Chawwâl en l'année 256 H/870 AD, et fut enterré à Khartank, un village près de Samarkand. Qu'Allah lui accorde la miséricorde [46]

2.6 Sahih Al-Bukhârî, le livre le plus juste après le Coran

قيل في الصحيح: "هو اصح كتاب بعد كتاب الله"

Sahih al-Bukhârî couvre presque tous les aspects de la vie du musulman, en fournissant des conseils essentiels de la part du Messenger d'Allah (que la prière et la paix d'Allah soient sur lui). Toutes les écoles musulmanes (madhab) s'accordent pour dire que Sahih al Bukhari est l'ouvrage le plus authentique après le Coran. Pour ces raisons :

- Il retenait pratiquement tous ce qu'il entendait et lisait et il avait appris par cœur durant toute sa vie 100.000 hadiths authentiques et le double de ce chiffre de hadiths qui n'avait pas le degré d'authenticité (ce qui ne veut pas dire pour autant que ce sont des hadiths faux).
- Al Bukhârî aurait entendu plus de 600 000 hadîth et mémorisé plus de 100 000 hadîth. Si l'on exclut les répétitions, son recueil contient 2762 hadîth. Cela signifie que les critères de validation d'un hadîth étaient particulièrement stricts.
- Comme tous les rapporteurs de hadith, n'a pas fait que noté ce qu'on lui rapportait, mais il a aussi établi des règles strictes pour classé ses hadith. Ainsi, un hadith sans Sanad est rejeté sans autre forme de procès, les hadiths dont un des rapporteurs n'était pas intègre (l'établissement de l'intégrité des rapporteurs est toute une branche de la science du hadith qui s'appelle "al djar'h wa at ta'adil") ou encore un hadith dont le récit chronologique ne cadrerait pas (rapporteurs n'ayant jamais connu la personne de laquelle il prétend rapporter ou qui n'a pas pu être présent au moment des faits pour une quelconque raison). [12]

3 Les corpus

3.1 Définition

Un corpus est un ensemble de documents, artistiques ou non (textes, images, vidéos etc.), regroupés dans une optique précise, Ils sont collectées sous format électronique, on peut utiliser des corpus dans plusieurs domaines : études littéraires, linguistiques, scientifique, etc

3.2 Le corpus en littérature

En littérature le corpus regroupe un ensemble de textes ayant une visée commune. Un corpus peut être constitué de documents différents (tableau, extrait de texte...) et ces documents divers ont un point en commun. En général c'est le thème qui fait figure de leur ressemblance. Il faut avoir une technique particulière pour le déchiffrer.

3.3 Le corpus en linguistique

La branche de la linguistique, qui se préoccupe des corpus s'appelle la linguistique de corpus. Elle est liée au développement des systèmes informatiques, en particulier à la constitution de bases de données textuelles. On parle de corpus pour désigner l'aspect normatif de la langue : sa structure et son code en particulier. Afin de rendre les corpus plus utiles pour faire la recherche linguistique, ils sont souvent soumis à un processus connu sous le nom d'annotation.

3.4 Le corpus dans la science

Les corpus sont des outils indispensables et précieux en traitement automatique du langage naturel. Ils permettent en effet d'extraire un ensemble d'information utile pour des traitements statistiques. D'un point de vue informatif, ils permettent d'extraire des tendances et notamment de construire des ensembles de n-grammes. D'un point de vue méthodologique, ils apportent une objectivité nécessaire à la validation scientifique en traitement automatique du langage naturel. L'information n'est plus empirique, elle est vérifiée par le corpus. Il est donc possible de s'appuyer sur des corpus bien formés pour formuler et vérifier des hypothèses scientifiques.

3.5 Conditions pour construire un corpus textuel

Le corpus ne se laisse pas uniquement définir formellement, comme un ensemble de texte ou une suite de caractères alphanumériques. Il vérifie trois types de conditions : des conditions de signifiante, des conditions d'acceptabilité, et des conditions d'exploitabilité.

- Conditions de signifiante : un corpus est constitué en vue d'une étude déterminée, portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue. Les documents retenus doivent être adéquats comme source d'information pour correspondre à l'objectif qui suscite l'analyse.
- Conditions d'acceptabilité : le corpus doit apporter une représentation fidèle, sans être parasité par des contraintes externes. Il doit avoir une ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse.

- Conditions d'exploitabilité : les textes qui forment le corpus doivent être commensurables. Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme). [20]

3.6 Méthodologie d'utilisation d'un corpus

Il serait maladroît d'un point de vue méthodologique d'appliquer des traitements statistiques sur le corpus qui a permis de faire ressortir un classement ou une modélisation du langage. Lorsque l'on travaille avec des corpus, il convient donc de séparer un corpus initial en deux sous-corpus :

- Le corpus d'apprentissage : qui sert à retirer un modèle ou un classement à partir d'un nombre suffisant d'information ;
- Le corpus de test, qui sert à vérifier la qualité de l'apprentissage à partir du corpus d'apprentissage. [11]

3.7 Les domaines d'utilisation des corpus

- Lexicographie (aide à la constitution de dictionnaires)
- Apprentissage des langues
- Études sociolinguistiques
- Linguistique : (l'étude de vocabulaire, de la grammaire, évolution de la langue ou des sens des mots).
 - Linguistique informatique (TALN), entraîner ou tester les outils d'analyse textuelle
 - Terminologie, traduction, rédaction technique
 - analyser les caractéristiques des textes traduits.
 - aide à la traduction.

3.8 Les avantages d'une analyse de corpus

Les avantages d'une analyse de corpus sont nombreux :

- Les textes sont des ressources fiables et stables puisqu'ils fixent par écrit, à un instant d'un certain nombre de connaissances sur le domaine.
- Les textes sont des sources de savoirs privilégiées car ils contiennent des connaissances explicites mises à disposition sur un support physique, par opposition aux connaissances des experts souvent tacites, difficilement exprimables et difficilement accessibles. Mais leur participation au processus de construction peut être ponctuelle aux étapes de validation.

- L'analyse automatisée de corpus est un gain de temps par rapport aux entretiens avec des experts du domaine.
- Les modèles sont à-priori plus facilement compréhensibles et donc maintenables car le retour aux textes est possible. [28]

4 Conclusion

Dans ce chapitre, nous avons étudié les Corpus Prophétiques, leurs recueils et inspiration et leurs caractéristiques. A partir cette étude nous avons constaté le besoin d'un système d'extraction et de représentation de connaissance à partir des bases de données prophétiques, qui facilitera l'indexation et l'interrogation des documents prophétique.

CHAPITRE 4

CLASSIFICATION DES DOCUMENTS

TEXTUELS : ETAT DE L'ART

1 Introduction

La classification de textes est une tâche générique qui consiste à assigner une ou plusieurs catégories, parmi une liste prédéfinie, ou non à un document. Actuellement, la classification de textes est un domaine de recherche très actif et l'automatisation de cette opération est devenue un enjeu pour la communauté scientifique, les travaux évoluent considérablement depuis une vingtaine d'années et plusieurs modèles ont vu le jour comme le filtrage (classification supervisée bi-classe), le routage (classification supervisée multi-classe) ou le classement ordonné (classement des textes par ordre de pertinence pour chaque catégorie)

2 Définition de la classification

« Le seul moyen de faire une méthode instructive et naturelle, est de mettre ensemble les choses qui se ressemblent et de séparer celles qui diffèrent les unes des autres. »

Le processus de classification cherche à mettre en évidence les dépendances implicites qui existent entre les objets, les classes entre elles, les classes et les instances. La classification recouvre les processus de reconnaissance de la classe d'un objet, et l'insertion éventuelle d'une classe dans une hiérarchie. Ce mode de raisonnement permet de reconnaître un objet en identifiant ses caractéristiques, relativement à la hiérarchie étudiée. La classification fait intervenir un processus de décision d'appartenance [5]

2.1 Pourquoi automatiser la classification?

Une méthode de classification automatique de document est non seulement possible, mais une solution viable pour beaucoup de problèmes. Il libère les gens de traiter avec des piles de papier non organisés. En utilisant un scanner équipé avec les caractéristiques nécessaires, la classification automatique de document permet à l'utilisateur de trier rapidement des pages et des lettres séparées et des télécopies en cas où vous en avez besoin. Plus en plus organisés, sans l'utilisation de beaucoup de personnes vous permettra d'économiser de l'argent, le temps et l'espace.

2.2 Domaines d'application de la classification

La classification est en pratique appliquée dans la plupart des domaines du monde réel.

Nous la trouvons à titre d'exemple dans:

- Le Web, pour la classification des documents en fonction de leurs sujets et le filtrage des spam (spam/non spam);
- Le secteur médical, pour la classification des patients en fonction de leurs maladies;
- La bio-informatique, pour la classification des gènes quand une grande quantité de gènes peuvent montrer des comportements similaires.
- Le marketing, pour la classification des entreprises en fonction de leurs productions.
-

2.3 Classification bi-classe et multi-classes

2.3.1 La classification bi-classe

La classification bi-classe correspond au filtrage. C’est une problématique pour laquelle le système de classification répond à la question : « Le texte appartient-il à la catégorie C ou non (i.e. ou à sa catégorie complémentaire $\neg C$? » (Par exemple, un document est-il autorisé aux enfants ou non). Cependant quand il s’agit d’effectuer une classification multi-classe qui permet de transmettre le document vers le ou les catégories(s) le(s) plus approprié(s), on parle alors de routage. Cette classification multi-classes, selon le cas, peut être disjointes ou non.

2.3.2 La classification multi-classes disjointes

La classification multi-classes disjointes est le contexte de classification en un nombre de classes supérieur à un et pour lequel un texte est attribué à une et une seule classe. Un système de classification multi-classes disjointes répond à la question « A quelle classe (au singulier) appartient le document ? ».

2.3.3 La classification multi-classes

Dans un système de classification multi-classes, on peut associer un texte à une ou plusieurs classes voire à aucune classe. Le système répond donc à la question : « A quelles classes (au pluriel) appartient le document ? ». C’est le cas le plus général de la classification.[15]

2.4 Classification de textes et Text Mining

Le Text Mining est une technique permettant d’automatiser le traitement de gros volumes de contenus texte pour en extraire les principales tendances et répertorier de manière

statistique les différents sujets évoqués ainsi découvrir des connaissances et des relations à partir des documents disponibles. L’outil de Text Mining va générer de l’information sur le contenu du document. Cette information n’était pas présente, ou implicite, dans le document sous sa forme initiale, elle va être rajoutée, et donc enrichir le document.

Les besoins en Text Mining peuvent être :

- Recherche d’information
- Correction orthographique/grammaticale
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- La veille technologique

Et notamment :

- La Classification automatique des documents

Toutes ces applications sont étroitement liées. [15]

2.5 Hiérarchie des méthodes de classification

Les méthodes de classification sont regroupées sous une forme hiérarchique, comme nous le montre la figure 4.1. Dans les méthodes exclusives, l’objet Doit appartenir à une seule classe. Par contre, dans les méthodes non exclusives, l’objet Peut appartenir à plusieurs classes en même temps avec un degré d’appartenance: c’est le Cas de la classification floue. Dans les méthodes exclusives, le problème de classification Est traite de deux façons: la classification non supervisée et la classification supervisée. [26]

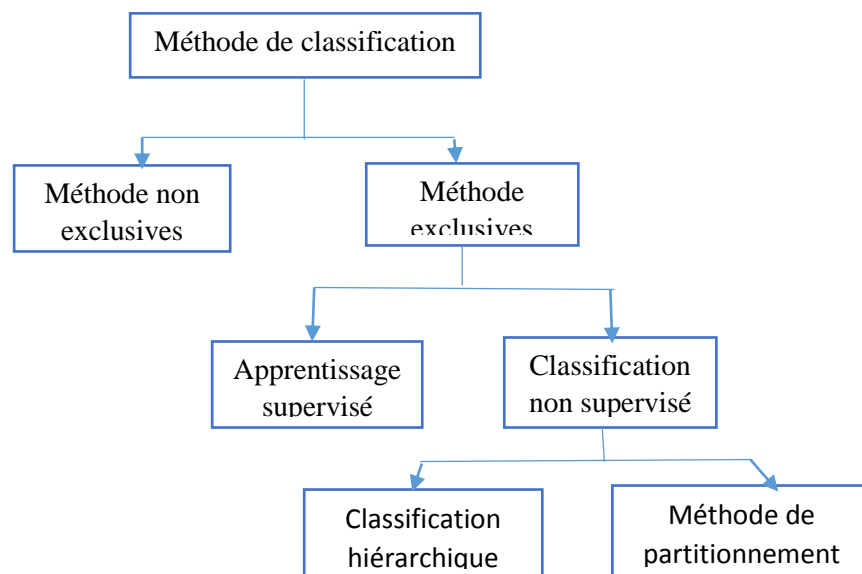


Figure 4.1 Hiérarchie des méthodes de classification. [26]

2.6 Les types de classification automatique

On distingue dans le domaine de la classification automatique deux types d’approches : la classification supervisée et la classification non supervisée. Ces deux méthodes diffèrent sur la façon dont les classes sont générées. En effet dans le cas de la classification non supervisée, les groupes de documents (classes) sont calculés automatiquement par la machine, tandis qu’ils sont, dans l’approche supervisée définis par un expert.

Cependant, Il existe d’autres types de classification qui s’appuient sur d’autres types de méthodes d’apprentissages comme « l’apprentissage semi-supervisé » et « l’apprentissage par renforcement ». En effet, l’apprentissage semi-supervisé est un bon compromis entre les deux types d’apprentissage « supervisé » et « non-supervisé », car il permet de traiter un grand nombre de données sans avoir besoin de toutes les étiqueter, et il profite des avantages des deux types mentionnés. Alors que L’apprentissage par renforcement est fort utilisé dans le cas d’apprentissage interactif

2.6.1 La classification supervisée

Dans ce type de classification, les classes sont prédéfinies avec une description des documents. Lorsqu’un nouveau document arrive, on le compare avec la description de chaque classe et on le met dans celle qui lui ressemble le plus. Plusieurs techniques sont utilisées, on peut citer K voisins Proches, Arbre de Décision, Naive Bayes, Machine à vecteur de support...

Dans ce qui suit notre travail va être concentré sur la catégorisation de textes (la classification supervisée)

2.6.2 La classification non supervisée

Lors du clustering, les objets sont groupés dans des classes homogènes disjointes. Pour faire ressortir les ensembles de documents, on doit maximiser l’homogénéité interne des classes et la dispersion entre elles. Les deux méthodes principales du clustering sont : les méthodes hiérarchiques et les méthodes non hiérarchiques.

3 Définition de la catégorisation de textes

La catégorisation de texte consiste à chercher une liaison fonctionnelle entre un ensemble de textes 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. [17]

3.1 Comment catégoriser un texte ?

Le processus de catégorisation intègre la construction d'un modèle de prédiction qui, en entrée, reçoit un texte et, en sortie, lui associe une ou plusieurs étiquettes. Pour identifier la catégorie ou la classe à laquelle un texte est associé, un ensemble d'étapes est habituellement suivies. Ces étapes concernent principalement la manière dont un texte est représenté, le choix de l'algorithme d'apprentissage à utiliser et comment évaluer les résultats obtenus pour garantir une bonne généralisation du modèle appris.

Le processus de catégorisation, intégrant la phase de classement de nouveaux textes, est résumé dans la (figure 4.2). Il comporte deux phases :

1. **l'apprentissage**, qui comprend plusieurs étapes et aboutit à un modèle de prédiction :
 - a) nous disposons d'un ensemble de textes étiquetés (pour chaque texte nous connaissons sa catégorie) ;
 - b) à partir de ce corpus, nous extrayons les k descripteurs (mots, termes) $(t_1; \dots; t_k)$ les plus pertinents au sens du problème à résoudre
 - c) nous disposons alors d'un tableau « descripteurs \times individus », et pour chaque texte nous connaissons la valeur de ses descripteurs et son étiquette ;
2. **le classement** d'un nouveau texte d_x , qui comprend deux étapes :
 - a) recherche puis pondération des occurrences $(t_1; \dots; t_k)$ des termes dans le texte d_x à classer.
 - b) application d'un algorithme d'apprentissage sur ces occurrences afin de prédire l'étiquette de ce texte d_x .

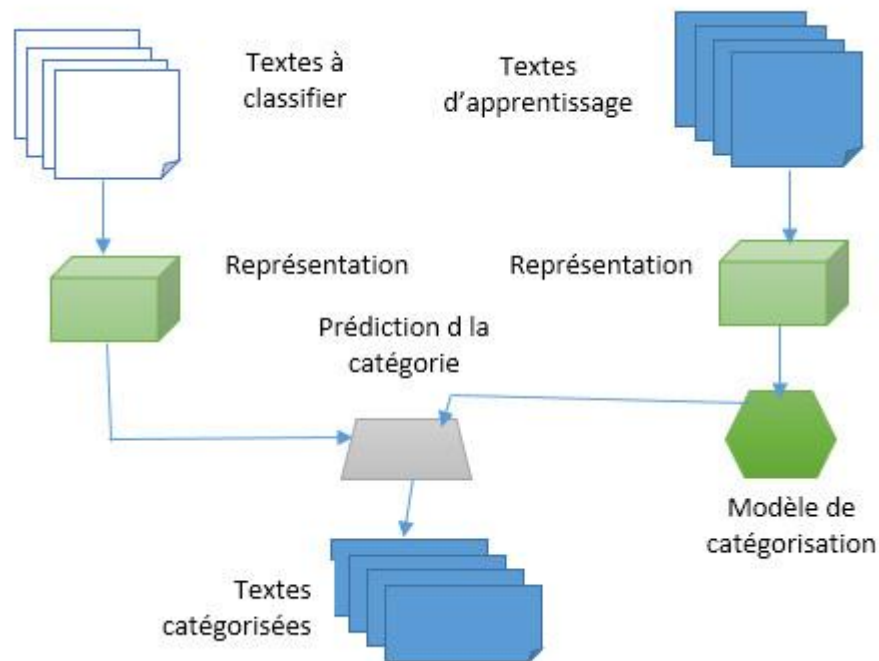


Figure 4.2 Processus de la catégorisation de textes. [47]

3.2 Domaines d'applications de la catégorisations de textes

La catégorisation de textes est utilisée dans de nombreuses applications. Parmi ces domaines figurent : l'identification de la langue, la reconnaissance d'écrivains et la catégorisation de documents multimédia. [17]

Habituellement, les catégories font référence aux sujets des textes, mais pour des applications particulières, elles peuvent prendre d'autres formes et on peut résoudre, par des techniques de catégorisation des problèmes tels que l'identification de la langue d'un document, le filtrage et la détection des spam (les courriers indésirables) afin de les supprimer, par exemple le classificateur naïve Bayes qui est utilisé pour détecter automatiquement le spam. Actuellement, les grandes entreprises ont besoin de gérer rapidement et efficacement le flux d'information pour la satisfaction des clients, plusieurs travaux de recherche existent, destinés à créer des outils informatiques et des ressources génériques pour la classification, le routage et l'acheminement des courriels vers leurs destinataires. Ces travaux cherchent aussi à développer un processus puissant de filtrage, il existe également d'autres applications comme la désambiguïsation des termes, la catégorisation des documents multimédia, l'indexation automatique des textes, et l'organisation des documents... [2]

3.3 Problèmes de la catégorisation de textes

Plusieurs difficultés peuvent s'opposer au processus de catégorisation de textes. Des problèmes connus dans la discipline liés à l'apprentissage automatique supervisé comme la subjectivité de la décision prise par les experts, le sur-apprentissage, etc.. Mais aussi des problèmes particuliers liés à la nature des données traitées à savoir des données textuelles comme la polysémie, la redondance, Les variations morphologiques ou même L'homographie, etc. Dans ce qui suit nous allons signaler les dix principales difficultés qui s'opposent à la catégorisation de textes : [15]

- **Redondance(Synonymie)**

La redondance et la synonymie permettent d'exprimer le même concept par des expressions différentes, plusieurs façons d'exprimer la même chose.

- **Polysémie (Ambiguïté)**

A la différence des données numériques, les données textuelles sont sémantiquement riches, du fait qu'elles sont conçues et raisonnées par la pensée humaine. Contrairement des langages informatiques, le langage naturel, autorise des violations des règles grammaticales engendrant plusieurs interprétations d'un même propos.

- **L'homographie**

Deux mots sont dits homographes s'ils s'écrivent de la même façon sans forcément avoir la même prononciation. L'homographie est une sorte d'ambiguïté supplémentaire.

- **La graphie**

Un terme peut comporter des fautes d'orthographe ou de frappe comme il peut s'écrire de plusieurs manières ou s'écrire avec une majuscule. Ce qui va peser sur la qualité des résultats. Parce que si un terme est orthographié de deux manières dans le même document (Ghilizane, Relizane), la simple recherche de ce terme avec une seule forme graphique néglige la présence du même terme sous d'autres graphies, ce qui va influencer les résultats puisque les différentes graphies vont être traitées séparément. Néanmoins du point de vue pratique, le fait qu'un terme inconnu est proche d'un autre terme prouve qu'il a été mal orthographié.

- **Les variations morphologiques**

Les conjugaisons, pluriels, influent négativement sur la qualité des résultats puisque les différentes variations morphologiques vont être considérées séparément et chacune va être prise comme un élément à part comme par exemple les trois termes : maître, maîtresse, maîtriser sont traités indépendamment quoique en réalité ça pivote sur la même idée.

- **Les mots composés**

La non prise en charge des mots composés comme : comme Arc-en-ciel, peut-être, sauve-qui-peut, etc. Dont le nombre est très important dans toutes les langues, et traiter le mot Arc-en-ciel par exemple en étant 3 termes séparés réduit considérablement la performance d'un système de classification néanmoins l'utilisation de la technique des n-grammes pour le codage des textes atténue considérablement ce problème des mots composés.

- **Présence-Absence de termes**

La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoique on sait très bien qu'il ya plusieurs façons d'exprimer les mêmes choses, dès lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document.

- **Complexité de l'algorithme d'apprentissage**

Nous verrons qu'un texte est représenté généralement sous forme de vecteur contenant les nombres d'apparitions des termes dans ce texte. Or, le nombre de textes qu'on va traiter est très important sans oublier le nombre de termes composant le même texte donc on peut bien imaginer la dimension du tableau (textes * termes) à traiter qui va compliquer considérablement la tâche de classification en diminuant la performance du système.

- **Sur-apprentissage**

Le nombre de termes très important et très varié qui ne se répètent dans tous les textes va causer énormément de creux dans le tableau de grande dimension (textes*termes) qui peut provoquer du sur-apprentissage qui s'explique par le fait que le modèle n'arrive pas à bien classer les nouveaux textes, pourtant il l'a bien fait dans la phase d'apprentissage en classant correctement les textes de la base d'apprentissage. Pour limiter le sur-apprentissage, on doit sélectionner des termes pour réduire la dimensionnalité. D'après les expériences antérieures, le nombre de termes doit être limité par rapport au nombre de textes de la base d'apprentissage.

Quelques auteurs recommandent d'utiliser au moins 50 à 100 fois plus de textes que de termes. En général le nombre de textes d'apprentissage est limité, c'est pour cela on cherche à agir sur le nombre des termes utilisés en les diminuant, pour éviter ce sur-apprentissage. Sans bien sûr pénaliser le système en supprimant des termes pertinents

- **Subjectivité de la décision**

Parmi les problèmes classiques usuels dans le domaine de l'apprentissage supervisé c'est la subjectivité de la décision prise par les experts qui décident de la classe à laquelle le texte va

être attribué. Certainement après la lecture du texte à classer, l’expert va trancher à quelle(s) catégorie(s) ce texte appartient en se basant sur le contenu sémantique et le contexte du texte et même en consultant d’autres textes préalablement associés à certaines classes, pour valider la décision prise qui ne peut être que subjective. Les experts humains ne lisent pas de la même manière ! Ne réfléchissent pas de la même manière ! Donc ne classent pas de la même manière. [15]

4 Choix des termes

Pour cette étape on transforme chaque document d_j en un vecteur $\mathbf{d_j} = (W_{1j}, W_{2j}, \dots, W_{|T|j})$, où T est l’ensemble des termes (ou descripteurs) qui apparaissent au moins une fois dans le corpus d’apprentissage. Le poids W_{kj} correspond à la contribution des termes t_k à la sémantique du texte d_j . [48]

Dans tous les cas (utilisation de mots simples ou de groupes de mots), il est préférable d’effectuer quelques prétraitements afin de filtrer les mots non informatifs et de regrouper les mots de même famille. Une première opération consiste à supprimer les mots faisant partie d’une liste prédéfinie : le stop words. Ce sont des mots génériques non porteurs de sens tels que les articles, les déterminants, les auxiliaires qui sont a priori inutiles pour discriminer les différentes catégories. Ils peuvent donc être supprimés sans perte d’information utile et leur suppression permet de diminuer significativement le nombre total d’occurrences dans le corpus. Cette liste doit être dressée préalablement pour chaque langue. [2]

Il y a plusieurs approches pour sélectionner ces descripteurs parmi lesquelles:

4.1 Représentation en « sac de mots » « bag of words »

La représentation de textes la plus simple a été introduite dans le cadre du modèle vectoriel présenté ci-dessus, et porte le nom de « sac de mots ». L’idée est de transformer les textes en vecteurs dont chaque composante représente un mot. Les mots ont l’avantage de posséder un sens explicite. Cependant, plusieurs problèmes se posent. Il faut tout d’abord définir ce qu’est « un mot » pour pouvoir le traiter automatiquement. On peut le considérer comme étant une suite de caractères appartenant à un dictionnaire, ou bien, de façon plus pratique, comme étant une séquence de caractères non-délimiteurs encadrés par des caractères délimiteurs (caractères de ponctuation); il faut alors gérer les sigles, ainsi que les mots composés ; ceci nécessite un prétraitement linguistique. On peut choisir de conserver les majuscules pour aider, par exemple, à la reconnaissance de noms propres, mais il faut alors résoudre le problème des débuts de phrases. [17]

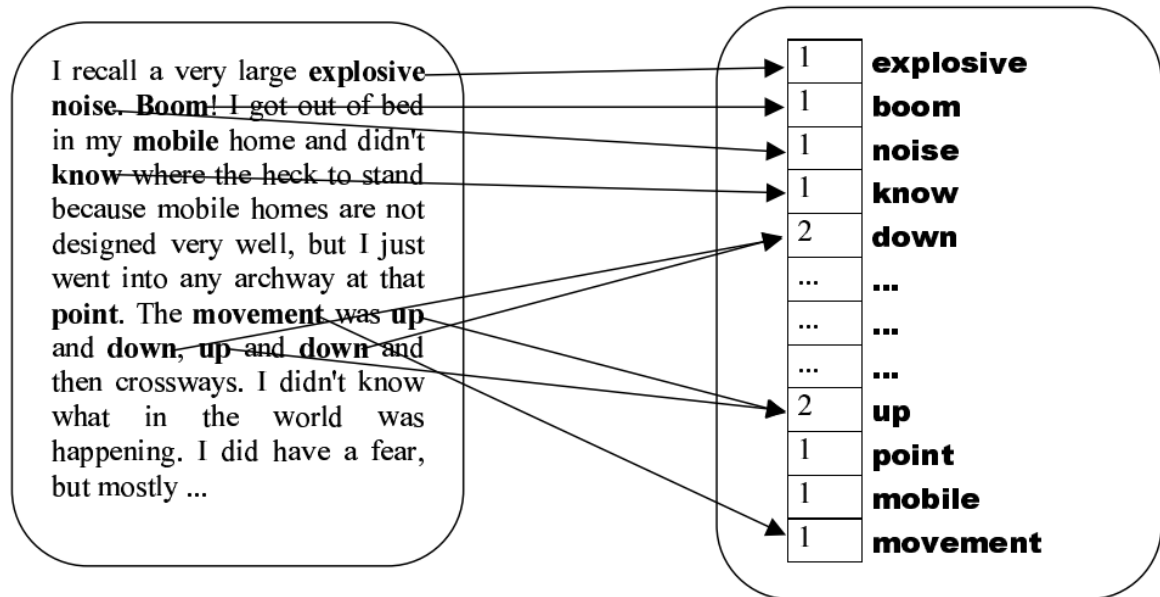


Figure 4.3 Exemple de la Représentation d'un texte extrait de la collection CLEF en sac de mots

4.2 Représentation des textes par des phrases

Un certain nombre de chercheurs proposent d'utiliser les phrases comme unité de représentation au lieu des mots comme le cas dans la représentation « sac de mot », puisque les phrases sont plus informatives que les mots seuls, par exemple « recherche d'information », « world wide web », ont un degré plus petit d'ambiguïté que les mots constitutifs, et aussi que les phrases ont l'avantage de conserver l'information relative à la position du mot dans la phrase.

4.3 Représentation avec des racines lexicales stemming

Dans la description du modèle précédent, chaque flexion d'un mot est considérée comme un descripteur différent ; en particulier, les différentes formes un verbe sont autant de mots. Par exemple, les mots franchi et franchit sont considérés comme des descripteurs différents alors qu'il s'agit de deux formes conjuguées du même verbe. Pour remédier à ce problème, il faut de considérer uniquement la racine des mots plutôt que les mots entiers (on parle de stem en anglais). Plusieurs algorithmes ont été proposés pour substituer les mots par leur racine. [2]

4.4 Représentation des textes avec des lemmes (lemmatisation)

La lemmatisation décrite auparavant, consiste à utiliser l'analyse grammaticale afin de remplacer les verbes par leur forme infinitive et les noms par leur forme au singulier. La substitution des mots par leur lemme réduit également l'espace des descripteurs comme pour les racines, et permet de représenter par un même descripteur des termes de même

signification. Par exemple, le remplacement des mots *banking*, *bank*, *banks*, par l'unique racine *bank* semble être rentable tout comme le remplacement des formes conjuguées *rebondit* et *rebondi* par le lemme *rebondir*. Les mêmes confusions d'ambiguïté peuvent être entraînées en représentant par un même descripteur des mots avec des sens différents, comme par exemple *glace* qui peut être un miroir ou « une glace aux chocolats ».

L'ambiguïté peut être causée aussi par le simple remplacement de la forme plurielle d'un mot par sa forme singulier comme *actions* qui est représenté par le descripteur *action*. Dans un contexte économique, le mot *actions* se réfère couramment à des actions d'entreprises et n'a rien à voir avec la notion *action* employée par exemple dans la phrase : « Le plan d'action de l'état ». [15]

4.5 Représentation des textes avec la méthode des n-grammes

Un n-gramme est une séquence de n caractères. Dans ce manuscrit un n-gramme désignera une chaîne de n caractères consécutifs. Dans la littérature, ce terme désigne quelque fois des séquences qui ne sont ni ordonnées ni consécutives ; par exemple un 2-gramme peut être composé de la première lettre et de la troisième lettre d'un mot, considèrent un n-grammes comme un ensemble non ordonné de n mots après avoir effectué la désuffixation (ou *stemming*) et la suppression de mots vides. Ce n'est pas l'acceptation utilisée ici.

Pour un document quelconque, l'ensemble des n-grammes est le résultat obtenu en déplaçant une fenêtre de n cases sur le texte ; ce déplacement se fait par étapes de un caractère et à chaque étape on prend une photo ; l'ensemble de ces photos donne l'ensemble de tous les n-grammes du document. Par exemple, pour générer tous les 5-grammes dans la phrase "Je suis un génie ", on obtient : *je_su*, *e_sui*, *suis_*, *_suis*, *uis_u*, etc. Le profil n-grammes d'un document est la liste des n caractères les plus fréquents, par ordre décroissant de leur fréquence d'apparition dans le document, ainsi que leurs fréquences elles-mêmes ; un document est caractérisé par son profil n-grammes. Les profils s'obtiennent en temps linéaire grâce à des tables de hachage. [17]

5 Codage des termes (Calcul du poids)

Une fois la liste des attributs déterminés, il reste à donner une pondération à chaque composant du vecteur. Il y a plusieurs façons pour coder ces composantes, nous détaillons les

deux codages les plus utilisés pour le domaine de la catégorisation automatique :

5.1 Le codage TF_IDF

Une façon largement utilisée de calculer le poids d’un terme est la fonction TF-IDF (acronyme pour « Term Frequency Inverse Document Frequency »). Issue du monde de la recherche d’information, celle-ci donne plus d’importance aux mots qui apparaissent souvent à l’intérieur d’un même texte, ce qui correspond bien à l’idée intuitive que ces mots sont plus représentatifs du document. Mais sa particularité est qu’elle donne également moins de poids aux mots qui appartiennent à plusieurs documents, pour refléter le fait que ces mots ont un faible pouvoir de discrimination entre les classes [2].

Le poids d’un terme t_k dans un document d_j est calculé ainsi :

$$Tf_IDF_{(tk,dj)} = (tk, dj). \log \frac{|Tr|}{(tk)} \quad (1)$$

- ✓ (t_k, d_j) est le nombre d’occurrences de t_k dans d_j
- ✓ $|Tr|$ est le nombre de documents d’apprentissage.
- ✓ (t_k) est le nombre de documents d’entraînements dans lesquels t_k apparaît au moins une fois

Si on désire avoir des poids entre 0 et 1, on peut appliquer une normalisation, ce qui est souvent le cas. La fonction TF-IDF a démontré une bonne efficacité dans des tâches de catégorisation de textes, et, en plus, son calcul est simple [2]

Cette pondération issue du domaine de la Recherche d’Informations (RI) tire son inspiration de la loi de Zipf introduisant le fait que les termes les plus informatifs d’un corpus ne sont pas ceux apparaissant le plus dans ce corpus. Ces mots sont la plupart du temps des mots outils. Par ailleurs, les mots les moins fréquents du corpus ne sont également par les plus porteurs d’informations [16].

5.2 Codage TFC

Le codage $TF \times IDF$ ne corrige pas la longueur des documents. Pour ce faire, le codage TFC est similaire à celui de $TF \times IDF$ mais il corrige les longueurs des textes par la normalisation en cosinus, pour ne pas favoriser les documents les plus longs.

$$TfC_{(tk,dj)} = \frac{Tf_IDF_{(tk,dj)}}{\sqrt{\sum_{s=1}^{|T|} (Tf_IDF_{(tk,dj)})^2}} \quad (2)$$

Codages sont également utilisés, comme par exemple le codage LTC qui tente de réduire les effets des différences de fréquences [17]

6 Réduction de la dimension

La taille des données peut être mesurée selon deux dimensions, le nombre de variables et le nombre d'exemples. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors de l'exploration et l'analyse de ces données. Pour cela, il est fondamental de mettre en place des outils de traitement de données permettant une meilleure compréhension de la valeur des connaissances disponibles dans ces données. La réduction des dimensions est l'une des plus vieilles approches permettant d'apporter des éléments de réponse à ce problème. Son objectif est de sélectionner ou d'extraire un sous-ensemble optimal de caractéristiques pertinentes pour un critère fixé auparavant. La sélection de ce sous-ensemble de caractéristiques permet d'éliminer les informations non-pertinentes et redondantes selon le critère utilisé. Cette sélection/extraction permet donc de réduire la dimension de l'espace des exemples et de rendre l'ensemble des données plus représentatif du problème. En effet, les principaux objectifs de la réduction de dimension sont :

- faciliter la visualisation et la compréhension des données,
- réduire l'espace de stockage nécessaire,
- réduire le temps d'apprentissage et d'utilisation,
- identifier les facteurs pertinents.

Les techniques mathématiques de réduction de dimension sont classées en deux grandes catégories :

6.1 Sélection des termes

La sélection d'attributs («feature selection») prend les attributs (ou mots) d'origine et conserve seulement ceux jugés utiles à la classification, selon une certaine fonction d'évaluation. Les autres sont rejetés. [2]

6.2 Extraction d'attributs

À partir des attributs de départ, elles créent de nouveaux attributs, en faisant soit des regroupements ou des transformations. On constate que la sélection d'attributs est meilleure

quant à l’élimination d’attributs réellement inutiles ou même d’attributs erronés («noisy») (mots mal orthographiés par exemple) tandis que l’extraction d’attributs est plutôt axée sur la réduction du nombre d’attributs redondants. [2]

7 Algorithmes d’Apprentissage supervisée

De façon simple, le but de l’algorithme est de découvrir pourquoi chaque document d’exemple a été rangé dans telle ou telle classe, afin de prédire la classe de nouveaux documents à ranger dans le futur.

Nous présentons ci-dessous quelques algorithmes d’apprentissage supervisé couramment utilisés dans le cadre de catégorisation automatique de texte :

7.1 Algorithme des K NN

KNN (K Plus Proches Voisins, ou KNN pour K Nearest Neighbours) a prouvé son efficacité face au traitement de données textuelles. La phase d’apprentissage consiste à stocker les exemples étiquetés. Le classement de nouveaux textes s’opère en calculant la distance entre la représentation vectorielle du document et celle de chaque exemple du corpus. Les K éléments les plus proches sont sélectionnés et le document est assigné à la classe majoritaire (le poids de chaque exemple dans le vote étant éventuellement pondéré par sa distance). [8]

7.2 Arbre de décision

Un arbre de décision permet de représenter les objets étudiés sous une forme arborescente, selon une hiérarchie des attributs déterminée par un calcul d’entropie. Ces méthodes sont populaires pour la présentation synthétique des données qu’elles fournissent, ainsi que pour la clarté des explications concernant la décision rendue.

7.3 Naïve Bayes (ou Simple Bayes)

L’algorithme Naïve Bayes (NB), est une autre méthode bien connue en apprentissage, elle est également utilisée dans la catégorisation de documents. Elle se base sur un modèle probabiliste, qui vise à estimer la probabilité conditionnelle d’une catégorie sachant un document et affecte au document la (ou les) catégorie(s) la (les) plus probable(s). La partie naïve de ce modèle est l’hypothèse d’indépendance des mots, c’est-à-dire que la probabilité conditionnelle d’un mot sachant une catégorie est supposée indépendante de cette probabilité pour les autres mots. Cette hypothèse fait que la catégorisation par NB est plus efficace que la

complexité exponentielle des approches bayésiennes non naïves qui utilisent des combinaisons de mots comme prédicateurs [7]

7.4 Les réseaux de neurones

Un réseau de neurones (ou Artificial Neural Network en anglais) est un modèle de calcul dont la conception est très schématiquement inspiré du fonctionnement de vrais neurones (humains ou non). Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type statistique grâce à leur capacité de classification et de généralisation, tels que la classification automatique de codes postaux ou la prise de décision concernant un achat boursier en fonction de l'évolution des cours. Ils enrichissent avec un ensemble de paradigmes permettant de générer de vastes espaces fonctionnels, souples et partiellement structurés. [12]

7.5 Machines à support de vecteurs (ou SVM)

Cette technique initiée par Vapnik tente de séparer linéairement les exemples positifs des exemples négatifs dans l'ensemble des exemples. Chaque exemple doit être représenté par un vecteur de dimension n . La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Intuitivement, cela garantit un bon niveau de généralisation car de nouveaux exemples pourront ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être tout de même situés franchement d'un côté ou l'autre de la frontière. L'efficacité des SVM est supérieure à celle de toutes les autres méthodes sur la classification de textes. Son efficacité est aussi très bonne pour la reconnaissance de formes. Un autre intérêt est la sélection de Vecteurs Supports qui représentent les vecteurs discriminant grâce auxquels est déterminé l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas. Cela en fait une méthode très rapide.

8 Critères d'évaluation d'algorithmes d'Apprentissage

Face à toutes ces différentes méthodes de classification, comment les comparer ? Sur la base de quels critères peut-on dire que tel classificateur est meilleur qu'un autre ? Ce concept, qui relève aussi de la sémantique d'un texte. Donc, pour l'instant, ce n'est pas formalisable. On se replie alors sur une évaluation empirique des classificateurs.

Presque toujours, on divise la banque de textes déjà classés et disponibles en deux ensembles : l’ensemble d’entraînement sur lequel le classificateur fait son apprentissage et l’ensemble de test sur lequel on peut évaluer sa performance. Parfois, on peut former un troisième ensemble, l’ensemble de validation, qu’on utilise pour l’optimisation de certains paramètres. Donc, l’ensemble de test contient des documents dont on connaît à l’avance les catégories auxquelles ils devraient appartenir. On pourra ainsi comparer les décisions prises par le classificateur automatique à celles des experts humains et calculer un score de performance. Ce calcul peut se faire de diverses façons, dont voici les principales.

On se considère ici en présence d’un classificateur binaire pour chaque catégorie, indiquant si oui ou non un document y appartient. Pour mieux illustrer les différentes mesures qui vont suivre, on prend pour point de départ la table de contingence illustrée par le tableau 4.1 suivant, distincte pour chaque classe.

	Document appartenant à la catégorie	Document appartenant à la catégorie
Document assignés à la catégorie par le classificateur	a	b
Document rejetés de la catégorie par le classificateur	c	d

Tableau 4.1: Table de contingence à la base de l’évaluation des classificateurs.

On définit à partir des statistiques de cette table les mesures suivantes :

- **précision** : $a / (a + b)$, soit le nombre d’assignations correctes sur le nombre total d’assignations.
- **rappel** : $a / (a + c)$, soit le nombre d’assignations correctes sur le nombre d’assignations qui auraient dû être faites
- **exactitude** : $(a + d) / (a + b + c + d)$
- **erreur** : $(b + c) / (a + b + c + d)$

Les deux dernières mesures, bien que couramment utilisées en apprentissage automatique, sont jugées moins bien adaptées à la tâche de classification de textes. Elles incluent dans leur définition le nombre total de documents. Or, comme un document appartient généralement à un petit nombre de catégories sur l’ensemble, un classificateur qui rejetterait tous les documents présenterait seulement un faible taux d’erreur et une exactitude quand même très

élevée. Entraîner un classificateur sur la base de l’optimisation d’un de ces deux critères tendrait à créer un programme qui n’accepte aucun document dans sa catégorie. C’est la raison pour laquelle la précision et le rappel sont les mesures les plus rencontrées dans la littérature. [21]

9 Etat de l’art

L’importance du Hadith chez les musulmans lui permet d’être l’un des plus importants domaines de développement des logiciels en tous genres qui existent autour de lui.

De nombreux efforts ont été consacrés au service de la tradition prophétique pour faciliter l’accès aux actes et paroles du Prophète Muhammad (paix et salut sur lui). Malgré ces efforts, nous constatons, qu’il y a un manque au niveau des systèmes de classification des traditions prophétiques.

9.1 Les travaux de Mohammed Naji Al-Kabi, Ghassan Kanaan, Riad Al-Shalabi, Saja I. Al- Sinjilawi et Ronza S. Al-Mustafa (Al-Kabi et al. 2005)

Cette recherche consiste à implémenter une méthode de classification des traditions prophétiques en adoptant la technique TF-IDF pour calculer les poids des termes dans les vecteurs des documents (Hadiths). Un terme avec un poids élevé est considéré comme un bon descripteur pour un chapitre particulier du Sahih de Al Bukhârî. L’exactitude moyenne pour cette étude était approximative de 83.2%. [41]

9.2 Les travaux de Mohammed Naji Al-Kabi et Saja I. Al-Sinjilawi (2007)

Le but de cette étude est de trouver la méthode optimale qui peut être utilisée pour classer des textes prophétiques parmi les six méthodes (produit intérieur, cosinus, Jaccard, Dice, Bayésien Naïfs, et Euclidien). Un vecteur du document a été utilisé pour calculer et comparer quatre coefficients associatifs différents du modèle de l’espace vectoriel (VSM). Les résultats expérimentaux montrent que le classifieur Bayésien Naïf dépasse les autres méthodes. L’algorithme Bayésien Naïf a été utilisé comme classifieur combiné avec la méthode Cosinus pour déterminer la ressemblance entre un Hadiths classe et son chapitre (classe) correspondant dans le livre de Sahih de Al Bukhârî. [42]

9.3 Les travaux de M.Ghazizadeh, M.H.Zahedi, M.Kahani et B.Minaei Bidgoli (2008)

Dans cette étude les auteurs ont confirmés que la “Science de Hadith” avec la “Science des Rejals”, peuvent contribuer ensemble pour prouver la validité de Hadith. L’objectif

principal de cette étude était de déterminer le taux de validité d'un Hadith à travers un système flou à l'aide d'un expert du domaine et une base de connaissance prises à partir du volume 1 du livre "Al-Kafi" pour une estimation au moyen des informations documentaires. Les résultats déduits par le système expert conçu ont été comparés avec les points de vue de l'expert et la comparaison a montré que le système a été correct dans 94% des cas. [33]

9.4 Les travaux de Harrag F, El-Qawasmeh E, Pichappan P (2009)

Dans cette étude, Harrag et EL-Qawasmeh (2009) ont utilisé un système de classification basé sur les arbres de décision (l'algorithme ID3). Ils ont utilisé un corpus de 453 documents réparti sur 14 catégories. Les résultats obtenus sont : le rappel 38%, la précision 47%, et le F1-score 40%. [33]

9.5 Les travaux de Harrag et El-Qawasmeh (2009)

Cette étude, a illustré une expérience de classification utilisant les réseaux de neurones sur un corpus de 453 Hadiths réparti sur 14 catégories de Sahih Al Bukhârî. Le rappel, la précision, et le F1-score pour la prévision de catégorie de Hadith étaient 87, 90, et 88% respectivement 88 %. [33]

9.6 Les travaux d'Alkhatib (2010)

Ici, l'auteur a comparé l'efficacité de quatre algorithmes différents pour classer le Hadith dans 8 livres sélectionnés du Sahîh Al Bukhârî mais seulement 8 catégories. Les algorithmes sont : l'algorithme Rocchio, l'algorithme du K-PPV, l'algorithme de Bayes Naïf et l'algorithme SVM. La technique TF_IDF a été utilisée pour le calcul des fréquences. Les documents prophétiques ont été divisés en deux parties, 75% des documents (1350 Hadiths) ont été utilisés comme données d'apprentissage et 25% (150 Hadiths) pour tester l'exactitude des modèles résultants. [11].

9.7 Les travaux de Harrag F, El-Qawasmeh E, Salman Al-Salman A (2011b)

Cette recherche a fait une évaluation de quelques méthodes de classification de (the stemming dictionary-lookup stemming, the root based stemming, et the light stemming). Cette expérience a utilisé sur 453 textes de Hadith de Sahih Al Bukhârî, qui a été réparti sur 14 catégories, ils ont utilisé SVM et ANN pour la phase de classification.

Pour les résultats : ANN était meilleure que SVM en termes de F1-score (ANN 42% et SVM 44%), Après le stemming, dictionary-lookup stemming était donné les meilleurs

résultats pour ANN (F1-score 50%) et light stemming était donné les meilleurs résultats pour SVM(F1- 48%) . [43]

9.8 Les travaux Aldhaln K, Zeki A, Zeki A, Alreshidi H (2012a, b) et Aldhaln K, Zeki A, Zeki A(2012)

Dans cette étude ils ont classé les Hadith dans quatre classes importantes Sahih, Hasan, Daif, et Maudo. Le corpus contient 999 Hadiths, ils sont rassemblés de trois livres : Sahih Al Bukhârî de, Jamiu Al-Termithi, et Al-Daeifah W al Al-Mawdhuah d'Al-Ahadith. Puisque l'ensemble de données s'est composé de différents livres, la tâche de classification faisait par l'intermédiaire de deux méthodes différentes : arbres de décision et Bayes naïf. Cependant, le classificateur naïf de Bayes a produit de meilleurs résultats avec exactitude 97,6% le rappel et 97,597%. [45].

9.9 Les travaux de Najeeb (2014)

Cette étude a proposé une nouvelle approche de classification pour l’authentification du Hadith (authentifié « Sahih » et faible « Daif »). Les règles d'association sont utilisés pour la classification .aucun taux d'exactitude explicite a été rapporté [44]

9.10 Les travaux de Mohammed N. Al-Kabi, Heider A. Wahsheh, Izzat M. Alsmadi, Abdallah Moh'd Ali Al-Akhras (2015)

Cette étude vise à évaluer l'efficacité de quatre algorithmes de classification bien connus pour classer les traditions prophétique en se basant sur la hiérarchie de Sahîh de Al Bukhârî et choisi le bon algorithme, Les algorithmes d’apprentissage automatique qui sont utilisés : (algorithme Bagging, et LogiBoost, l’algorithme de Bayes Naïf et SVM). Cette recherche a utilisé cinq classes de Sahih el bukhari. Les résultats d'évaluation ont montré que l'algorithme (NB) est plus efficace que les autres algorithmes. [30]

9.11 Les travaux de Mohammad Arshi Saloot, Norisma Idris Rohana Mahmud, Salinah Ja’afar et Dirk Thorleuchter Abdullah Gani. (2016)

Cette recherche fait une comparaison entre la plupart des travaux existante sur data mining et la classification des traditions prophétiques en utilisant le même corpus (3150 Hadiths de Sahih Al Bukhârî) les résultats d’évaluations des travaux de classification montraient que « neural networks classify the Hadith » est la meilleur avec 94% de précision. [33]

10 Conclusion

La classification automatique de textes qui est considérée parmi les composantes les plus importantes d’un système de recherche d’information, car elle permet d’organiser les documents par catégories et ainsi elle permet d’accélérer, cibler et améliorer la recherche d’information ; la représentation de textes doivent bénéficier de la même importance que les méthodes de classification cela s’explique par le fait qu’une bonne classification nécessite une bonne représentation. Nous avons également donné quelques métriques utilisées pour calculer les performances d’un système de classification utilisées,

CHAPITRE 5

REALISATION ET EXPERIMENTATION

1 Introduction

La classification automatique de documents devient nécessaire à cause du volume de documents échangés et stockés sur les supports électroniques.

Notre domaine d'application est les Traditions Prophétiques, donc il est nécessaire de présenter ce domaine et de poser sa problématique. Nous proposons d'étudier en détail la structure d'un corpus prophétique choisi qui est « Sahîh Al-Bukhârî » et nous nous intéressons au problème d'extraction d'information dans ce corpus. Une grande partie de ce chapitre est réservée à la description de notre démarche méthodologique,

2 Le langage et l'environnement de programmation

2.1 Visual Studio 2013

Visual Studio 2013 est un Environnement de développement intégré extensible, complet et gratuit pour créer des applications moderne pour Windows, Android et iOS, ainsi que des applications Web et des services Cloud. Le logiciel peut inclure des composants de Microsoft Windows, Microsoft Windows Server, Microsoft SQL Server, Microsoft Exchange, Microsoft Office et Microsoft SharePoint. Visual Basic, Visual C++, Visual C# et Visual J# utilisent tous le même environnement de développement intégré (IDE, Integrated Développement Environnement), qui leur permet de partager des outils et facilite la création de solutions faisant appel à plusieurs langages. Par ailleurs, ces langages permettent de mieux tirer parti des fonctionnalités du .NET Framework, qui fournit un accès à des technologies clés simplifiant le développement d'applications Web ASP et de Services Web (XML.) [32]

2.2 Langage de programmation

C# est un langage orienté objet de type sécurisé et élégant qui permet aux développeurs de générer diverses applications sécurisées et fiables qui s'exécutent sur le .NET Framework. Vous pouvez utiliser le langage C# pour créer entre autres des applications clientes Windows traditionnelles, des services Web XML, des composants distribués, des applications client-serveur et des applications de base de données. Visual C# fournit un éditeur de code avancé, des concepteurs d'interfaces utilisateur pratiques, un débogueur intégré et de nombreux autres

outils, pour faciliter le développement d'applications basées sur la version 4.0 du langage C# et sur la version 4 du .NET Framework. [8]

La syntaxe C# est très expressive, mais elle est également simple et facile à apprendre. La syntaxe de C# est facile à reconnaître à ses accolades si vous connaissez déjà les langages C, C++ ou Java. Si vous connaissez déjà l'un de ces langages, vous pouvez devenir très vite productif en C#. La syntaxe C# permet de répondre à de nombreuses complexités de C++ en fournissant des fonctionnalités puissantes telles que des types valeur Nullable, des énumérations, des délégués, des expressions lambda et des accès directs à la mémoire qui n'existent pas en Java. C# prend en charge des méthodes et types génériques qui améliorent la cohérence et les performances des types, ainsi que des itérateurs, qui permettent aux développeurs de classes de collection de définir des comportements d'itération personnalisés simples à utiliser par le code client. Les expressions LINQ (Language Integrated Query) transforment les requêtes fortement typées en construction de langage de premier ordre. [8]

3 Présentation du corpus utilisé

Nous avons utilisé un corpus de textes qui contient un ensemble des Aḥādīth d'après « sahih bukhari » constitué de 453 hadiths répartis en 14 groupes voire (tableau 5.1), les hadiths que on a utilisé sont sans le Sanad qui est une partie du Hadith qui fait référence à la chaîne de noms de personnes qui ont transmis Al-Hadith.

Catégorie	Nbr de document d'apprentissage	Nbr de document de test
الإيمان	23	9
العلم	22	9
اللباس-الزينة	34	8
الجنائيات	22	9
القرآن	24	10
الجهاد	24	8
الأشربة-الأطعمة	31	8
المعاملات	25	8
الأخلاق-الأداب	31	8
الأحوال-الشخصية	24	8
العبادات	33	9
الأفضية-الأحكام	24	8
الأمم السابقة	12	5
السيرة	11	6
Total	340	113

Tableau 5.1 Le corpus utilisés dans nos expérimentations.

4 Processus de catégorisation suivi par notre application

Le schéma suivant explique le processus de catégorisation selon notre application :

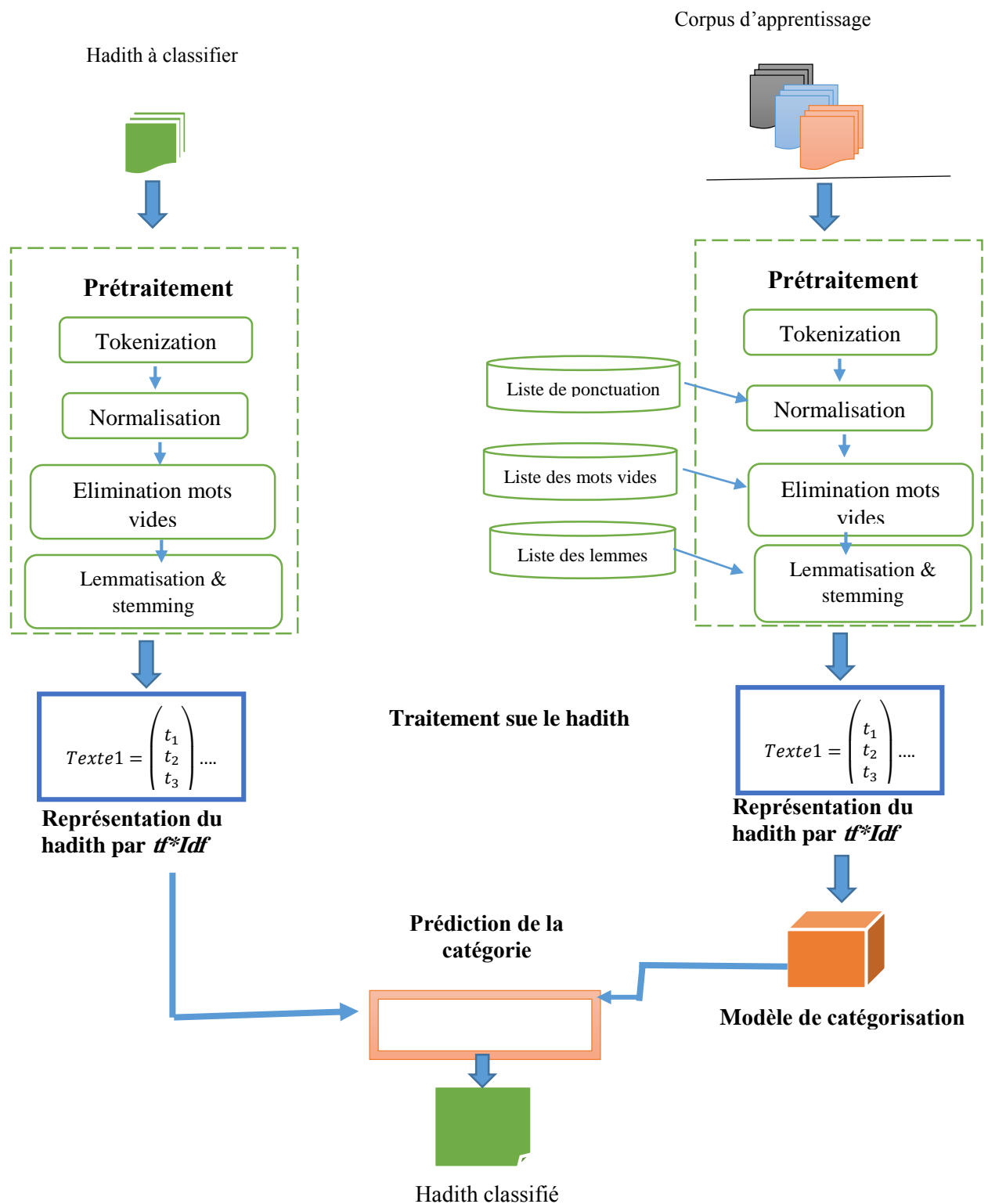


Figure 5.1 Processus de catégorisation proposée. (Adaptation [47])

5 Présentation de l'application réalisée (le classifieur)

Notre application est constituée d'un ensemble d'interfaces et de fonctions. Une interface peut être définie comme étant un interprète qui assure et facilite le dialogue entre l'utilisateur et l'application. Dans ce sens, les interfaces réalisées permettent d'une part, à l'utilisateur de formuler ses besoins et d'autre part, à la machine d'exposer des résultats compréhensibles par l'utilisateur.

5.1 Interface principale

L'interface principale de l'application, où l'utilisateur peut construire son procédé comme le montre la figure suivante :



Figure 5.2 L'interface principale de l'application

5.2 Prétraitement sur le « Matn »

Les algorithmes d'apprentissage ne sont pas capables de traiter directement les textes ni, plus généralement, les données non structurées. C'est pourquoi une étape préliminaire dite de représentation est nécessaire. Cette étape consiste généralement en la représentation de chaque document (hadith) par un vecteur, dont les composantes sont par exemple les mots contenus dans le texte, afin de le rendre exploitable par les algorithmes d'apprentissage.

Le prétraitement sera effectué sur chaque Hadith utilisés dans l'ensemble d'apprentissage et l'ensemble de test. Cette étape est nécessaire avant la phase de classification peut être appliqué pour l'extraction de connaissances à partir du Hadith et il se compose de plusieurs sous-étapes :

- Faire un parcours pour obtenir les fichiers du corpus.

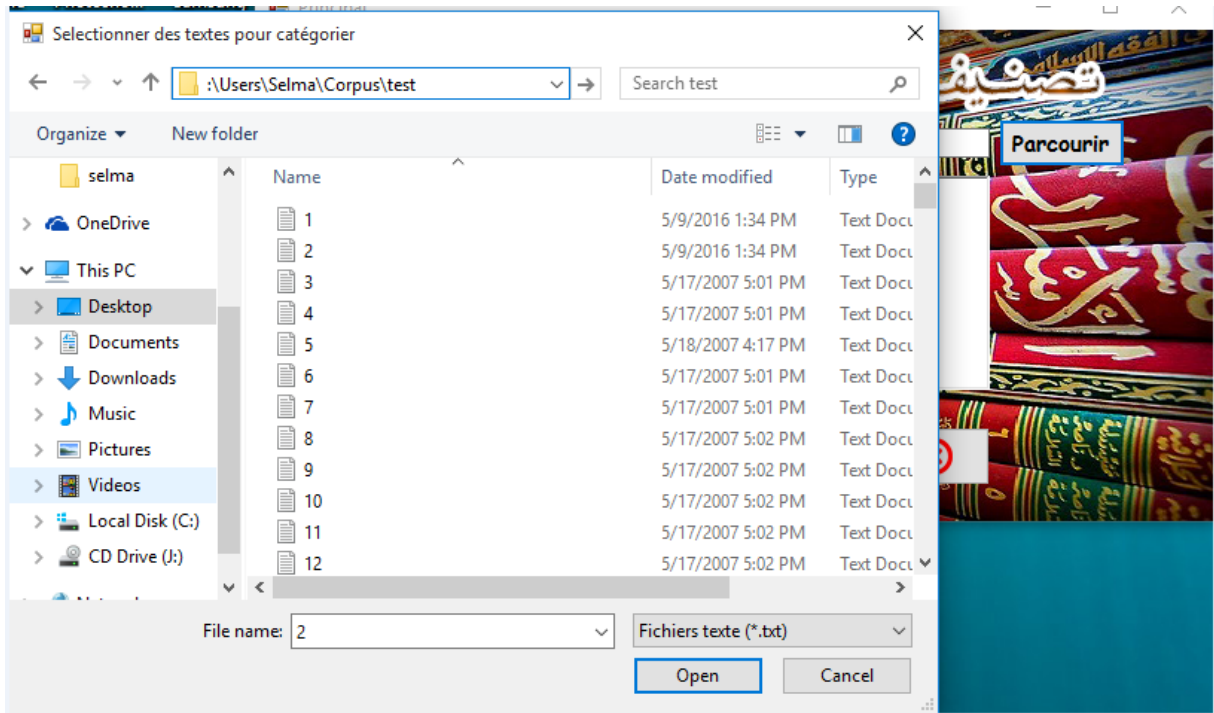


Figure 5.3 parcourir des fichiers.

- Sélectionner un fichier



Figure 5.4 sélection d'un fichier.

- **Tokenization** : Qui vise à diviser le Hadith en des mots ; le Hadith a été Facile à résoudre étant donné que chaque mot (token) peut être identifié comme une chaîne de lettres entre les espaces blancs :



Figure 5.5 Etape de Tokenization.

- **Normalisation des lettres** : La suppression des signes diacritiques et la ponctuation est importante, car ces marques sont répandues dans le hadith et n'ont aucun effet sur la détermination du classe du Hadith. Cette phase est représentée par (Elimination des ponctuations, Elimination des voyelles, Elimination des non-lettres (chiffres, caractères spéciaux,...etc.)).



Figure 5.6 Etape d'éliminations des ponctuations, chiffres et caractères spéciaux.

▪ Elimination des Mot vide

Une bonne partie des chercheurs débutent le processus d'indexation par la suppression de mots vides de sens. Cette méthode nécessite l'utilisation d'une liste de ces mots qui ne contiennent aucune information sémantique et qui ne modifient pas le sens des mots qui les accompagnent. Par exemple, en arabe, on peut compter des mots comme «من», «إلى», «عن», «في». Ou il se compose des pronoms en arabe, les prépositions, les noms de personnes (compagnons du Prophète Mohammed) et les lieux ont été mentionnés dans le corpus du Hadith. Après avoir éliminé les mots vides du Hadith, les autres mots (termes) restent.

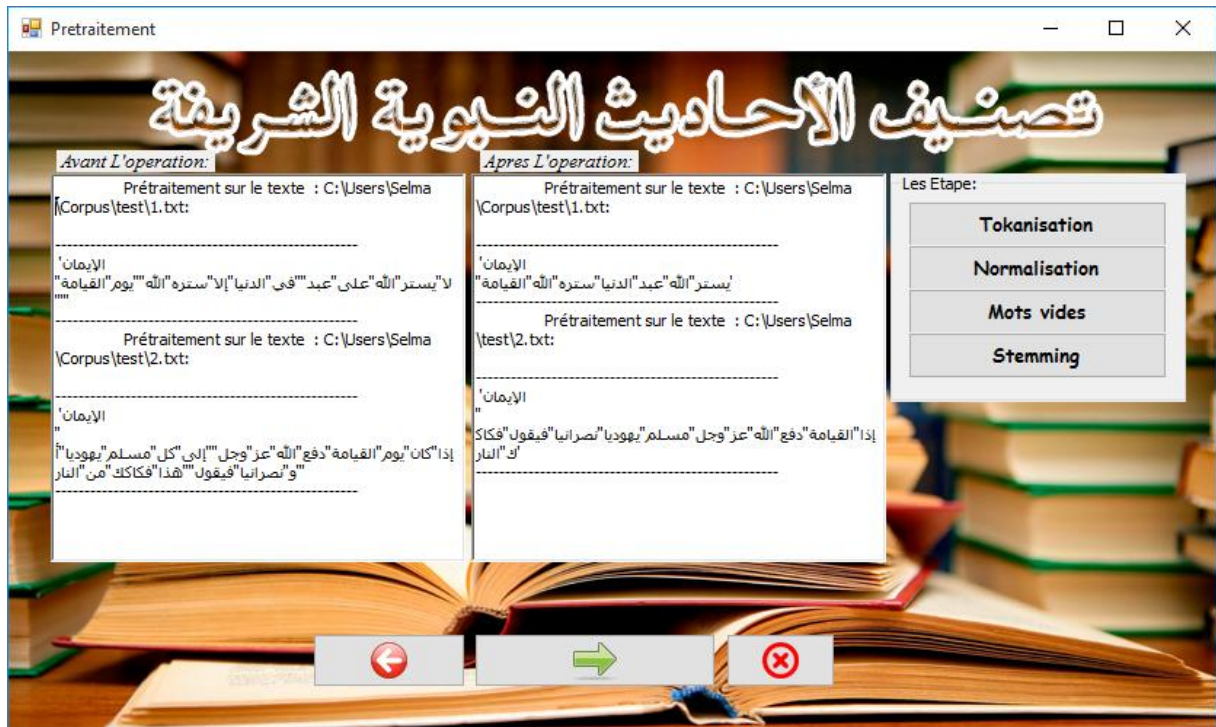


Figure 5.7 Etape d'élimination des mots vide.

- **Lemmatisation et stemming** : malheureusement c'est une opération très difficile pour le cas de la langue arabe.

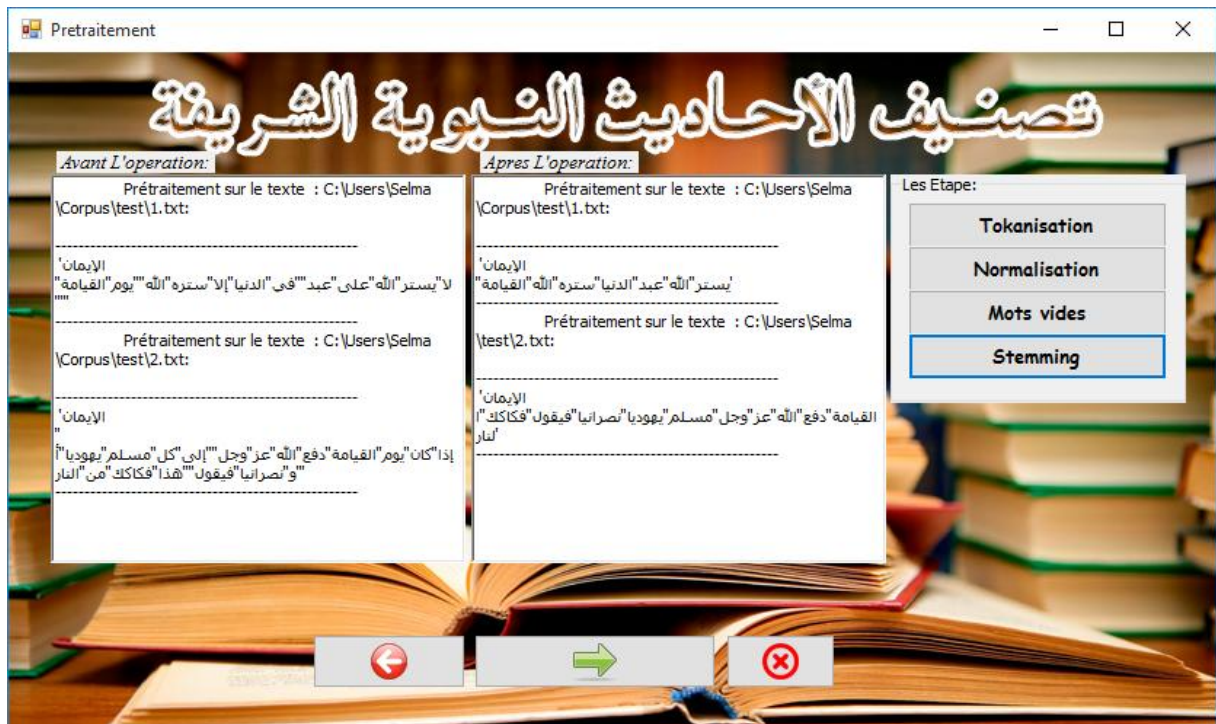


Figure 5.8 Etape de stemming

5.3 L'approche utilisée pour la représentation de corpus

Pour la représentation de corpus, nous avons utilisé l'approche « sac de mot » qui est le plus utilisé dans le domaine et vu sa simplicité et son efficacité, (voir chapitre 4) .

5.4 Représentation de texte par tf*Idf et apprentissage

Notre travail consiste à trouver une bonne représentation des textes, en se basant sur le codage TF*IDF, afin de pouvoir appliquer les tâches de la catégorisation d'une façon la plus simple possible en s'appuyant sur les résultats de TF*IDF.

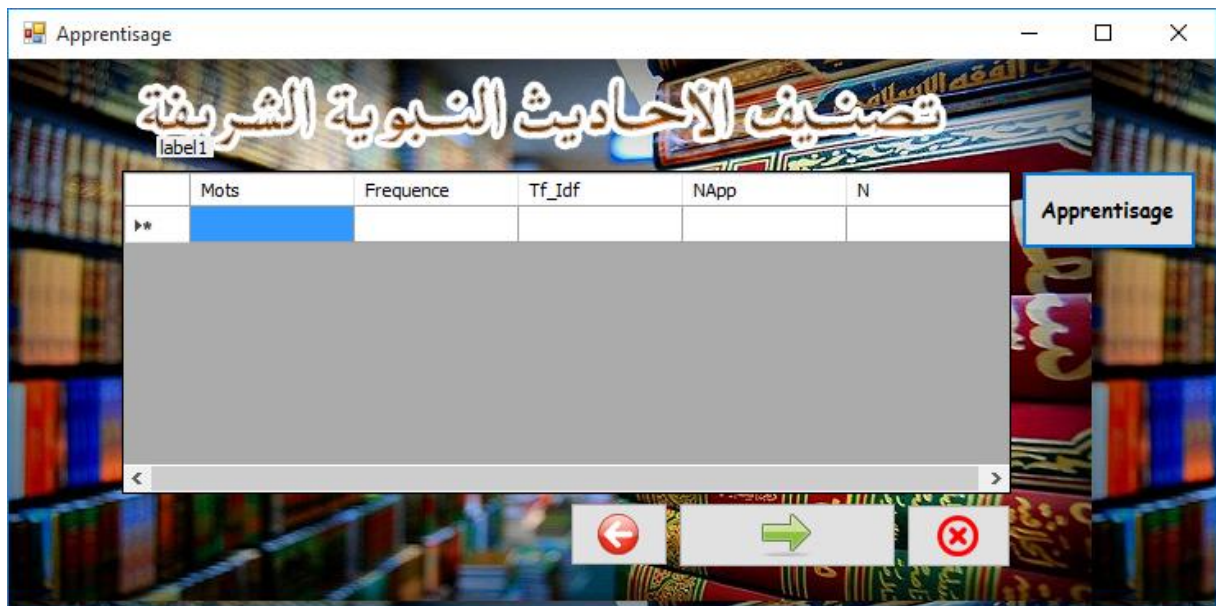


Figure 5.9 TF-IDF des termes du fichier.

5.5 Choix de l'algorithme d'apprentissage

Nous avons choisi l'algorithme NB, pour construire notre modèle de prédiction qui nous permet d'associer des catégories aux différents textes.

$$P(d_i|c_j) = P(c_j) \pi_{kj} P(t_k|c_j) \quad (3)$$

Où : d_i est le document à classer, c_j est une catégorie, t_k est un terme
 $P(c_j)$ est la probabilité associée à la catégorie C , $P(t_k|c_j)$ est la probabilité d'avoir le terme t_k sachant la catégorie C .



Figure 5.10 classification du hadith.

5.6 Evaluation des résultats du classifieur :

Pour évaluer les résultats obtenus par un classifieur, les documents de l'espace d'apprentissage sont souvent divisés en deux ensembles : le premier est utilisé pour la construction du classifieur, tandis que le deuxième est utilisé pour faire le test. Puisqu'on adopte l'approche de classification supervisée on connaît à l'avance la catégorie de chaque document. Ainsi, on compare la catégorie prédite avec celle prédéfinie et on calcule un score de performance. Ce calcul peut se faire de différentes façons

6 Conclusion :

La Langue arabe est considéré comme l'un des langues qui ne seront jamais disparus mais il y a quelques recherches ont été faites sur les corpus arabe. Toutefois, il est la langue officielle de vingt des pays de l'orient est et les pays africains, et la langue religieuse de tous les musulmans, indépendamment de leur origine.

Une méthode de classification est proposée dans cette étude, afin de découvrir les connaissances les corpus des traditions prophétiques en attribuant chaque Hadith à une des classes prédéfinies.

CONCLUSION GENERALE

La classification de textes s'est avérée au cours des dernières années comme un domaine majeur de recherche pour les entreprises comme pour les particuliers. La langue arabe est une langue riche morphologiquement ce qui ouvre les portes au plusieurs recherches dans le domaine de fouille de textes et de TALN.

Cette vue d'ensemble du domaine de la fouille de données et particulièrement la fouille de données textuelles nous permet d'affirmer que ce domaine de recherche est mûr, il est maintenant possible, en utilisant différents outils, de faire ressortir des informations de manière automatique à partir d'un ensemble de textes, sans même les lire.

Malgré toutes les difficultés qu'on les a rencontrés telle que : le manque de références et des travaux concerne la classification des textes arabes et spécialement la classification des traditions prophétiques et d'outils spécialisés, la non suffisance du temps, aussi la difficulté liées au traitement automatique de la langue Arabe. Mais, nous pensons qu'on a quand même pu relever le défi. Et par la même occasion, apprendre beaucoup de nouvelles connaissances tout au long de la réalisation de ce travail telles que :

- ✓ La programmation objet sous le langage C #.
- ✓ Des nouveaux concepts dans des domaines d'actualité tels que : le Data Mining, le Text Mining, et l'apprentissage automatique,
- ✓ La classification automatique supervisée de documents qui se situe dans l'intersection des domaines précités.
- ✓ Le traitement automatique du langage naturel TALN (en particulier la langue arabe).

Toutefois, le sujet abordé étant très vaste, il reste beaucoup à faire pour améliorer notre système. A cet effet, nous proposons comme perspectives :

- ✓ Utiliser des corpus d'apprentissage et de test de grandes tailles pour donner plus de crédibilité aux résultats obtenus.
- ✓ Utiliser d'autres formes de documents (HTML, XML, ...).
- ✓ Etudier le cas de la classification multi-classe et la classification non supervisée (clustering).
- ✓ Nous envisageons de partitionner les traditions prophétiques non plus en hadith Sahih El bukhari mais en autres Sahih ou en Charh de Hadith, cela pourrait résoudre le problème de l'homogénéité des documents.

Pour conclure, ce travail peut être amélioré au future par l'ajout des ontologies puisque ils sont devenues des outils importants pour structurer la connaissance pour des résultats plus sûr, ou bien l'élaboration d'une ontologie pour les documents prophétique ce qui va permettre de d'établir des liens entre les différentes branches de la culture islamique comme le Hadith, le Coran, la Jurisprudence et les sciences de la langue arabe.

BIBLIOGRAPHIE

- [1] Abd El Salam AL HAJJAR, Extraction et gestion de l'information à partir des documents arabes, mémoire de Doctorat d'informatique, Université paris – saint dénis, Décembre 2010.
- [2] ABIDI Karima, la catégorisation de texte multilingue, mémoire de magistère d'informatique, Ministère De L'enseignement Supérieur Et De La Recherche Scientifique Ecole Supérieure D'informatique, 2011.
- [3] Aïda KHEMAKHEM, ArabicLDB : une base lexicale normalisée pour la langue arabe mémoire de MASTER en Systèmes d'Information et Nouvelles Technologies, Université de Sfax Faculté des Sciences Economique et de Gestion, 2006.
- [4] Akila DJEBBAR-ZAIDI, Optimisation de la recherche d'un cas Bayésien, Thèse de doctorat en Informatique, université Badji Mokhtar Annaba, 2013.
- [5] Ali El Akadi, contribution à la sélection de variables pertinentes en classification supervisée : Application à la sélection des gènes pour les puces à ADN et des caractéristiques faciales, mémoire de doctorat d'informatique, Université Mohammed V – Agdal, Rabat, 2012.
- [6] Amroun Karima, découverte des dépendances fonctionnelles floues dans des modèles relationnels sous imprécision, mémoire de magister d'informatique, Université M'hamed Bougara de Boumardas, 2008.
- [7] Baali Meriem, utilisation de la technique des n-gramme dans l'extraction des racines en langue arabe, mémoire de master d'informatique, université de m'sila, 2015.
- [8] DAHOUMI Fares, Identification de la langue et catégorisation thématique des textes d'un corpus multilingue en utilisant les algorithmes : NB, SVM, mémoire de master d'informatique, Université De M'sila, 2013.
- [9] Fouad Soufiane Douzidia, Résumé automatique de texte arabe, Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique Septembre, 2004.
- [10] Harrag fouzi, Modélisation d'expertise dans les bases de données : application à l'authentification du " hadith ". Université Ferhat Abbes Sétif, 2005.

- [11] Harrag Fouzi, Une approche de fouille des textes basée sur la classification et la segmentation thématique : Application au corpus des Traditions Prophétiques "Hadith", mémoire de doctorat d'informatique, Université Ferhat ABBAS, Sétif, 2011.
- [12] HELASSA MADIHA, Extraction de connaissances à partir de données : application au hadith, mémoire de Magister Université Ferhat Abbas-Sétif, 2012.
- [13] LAMICHE Chaabane, fusion et fouille de données guidées par les Connaissances : application à l'analyse d'image, mémoire de doctorat d'informatique, Université Mohamed khider – Biskra, 2013.
- [14] Lazhar FAREK, Identification d'opinions dans les journaux arabes, mémoire de magister d'informatique, Université Badji Mokhtar–Annaba, 2009.
- [15] MATALLAH Hocine, Classification Automatique de Textes Approche Orientée Agent, mémoire de Magister en informatique, Université Aboubekr Belkaid-tlemcen, 2011.
- [16] Nicolas Béchet, Extraction et regroupement de descripteurs morphosyntaxiques pour des processus de Fouille de Textes, thèse de doctorat, Université des Sciences et Techniques du Languedoc, 2008.
- [17] Radwan JALAM, Apprentissage automatique et catégorisation de textes multilingues, thèse de doctorat, université Lumière Lyon2, Année 2003.
- [18] Rahmani Rabah, découvert d'associations sémantique dans les bases de données relationnelles par des méthodes de data mining, mémoire de magister d'informatique, Université mouloud mammeri tizi-ouzou, 2001
- [19] SASSI Amina, Une approche basée agent pour la fouille de données, mémoire de Magister en Informatique, Université HADJ LAKHDAR – BATNA, 2013.
- [20] Siham Boulaknadel, Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de Spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation, mémoire de doctorat d'informatique, Université de Nantes, octobre 2008.

- [21] SIMON RÉHEL, catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés, Mémoire présenté à la Faculté des études supérieures de Université Laval dans le cadre du programme de maîtrise en informatique pour l'obtention du grade de maître ès sciences (M.Sc.), 2005.
- [22] Slim MESFAR, Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard, mémoire de doctorat d'informatique, Université de Franche-Comté, novembre 2008.
- [23] Sofiane baloul, Développement d'un système automatique de synthèse de la parole à partir de texte arabe standard voyelle, mémoire de doctorat d'informatique, Université du Marie-France, 2003.
- [24] Soraya Zaidi–Ayad, Une plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran), mémoire de Doctorat en Informatique, Université. Badji Mokhtar, Annaba, 2013.
- [25] Tahar DILEKH, Implémentation d'un outil d'indexation et de recherche des textes en arabe, mémoire de Magister en Système d'Information et de Connaissance (SIC), Université Hadj Lakhdar – Batna, 2011.
- [26] Taoufik Guernine, classification hiérarchique floue basée sur le SVM et son application pour la catégorisation des documents, Mémoire présente au Département d'informatique En vue de l'obtention du grade de maître ès sciences (M.sc.), Université de Sherbrooke ,sherbrooke, québec,Canada,14 avril 2010.
- [27] Yannick Toussaint, Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances, Mémoire de doctorale en informatique, Université Henry Poincaré, Nancy 1, 21 novembre 2011

Article

- [28]Marti.A.Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, juillet 1992, pages 539-545.
- [29]Moath Najeeb, Abdelkarim Abdelkader, Musab Al-Zghoul, Abdelrahman Osman, A Lexicon for Hadith Science Based on a Corpus, Moath Najeeb et al, / (IJCSIT) International

Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, pages 1336-1340

[30] Mohammed N. Al-Kabi, Heider A. Wahsheh, Izzat M. Alsmadi, Abdallah Moh'd Ali Al-Akhras, Extended Topical Classification of Hadith Arabic Text, International Journal on Islamic Applications in Computer Science And Technology, Vol. 3, Issue 3, September 2015, pages 13-23.

Webographie

[31] VOLLE Michel. “Intranet et Datamining”. Article en ligne 2001. Disponible sur www.volle.com/lectures/ACM1.htm. Consulté le : 07.05.2016.

[32] <https://www.visualstudio.com/products/visual-studio-community-vs> Consulté le : 15.05.2016.

[33] Mohammad Arshi Saloot· Norisma Idris· Rohana Mahmud· Salinah Ja'afar· Dirk Thorleuchter· Abdullah Gani, Hadith data mining and classification : a comparative analysis, Article en ligne 2016.

http://www.researchgate.net/publication/290222515_Hadith_data_mining_and_classification_a_comparative_analysis Consulté le : 15.05.2016

[34] Mohamed Amine, Chérâgui Youssef Hceini et Moncef Abbas, Une approche multicritère pour lever l'ambiguïté morphologique dans le texte arabe pages 356-367, Disponible sur :

[http://dspace.univ-ouargla.dz:8080/jspui/bitstream/123456789/2820/1/Mohamed% 20Amine.pdf](http://dspace.univ-ouargla.dz:8080/jspui/bitstream/123456789/2820/1/Mohamed%20Amine.pdf) Consulté le : 15.03.2016

[35] Fathi DEBILI, Voyellation automatique de l'arabe. 42-49, Disponible sur :

<https://www.aclweb.org/anthology/W/W98/W98-1006.pdf> Consulté le : 15.03.2016

[36] Djamel Abdelkader ZIGHED & Ricco RAKOTOMALALA, Extraction des Connaissances à partir des Données (ECD), disponible sur :

<http://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/bases-de-donnees-42309210/extraction-de-connaissances-a-partir-de-donnees-eed-h3744/>

Consulté le : 18.04.2016

[37] http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf

Consulté le : 18.04.2016

- [38] Shafi Mohammad, The HADITH - How it was Collected and Compiled, Disponible sur: <http://www.daralislam.org/portals/0/Publications/TheHADITHHowitwasCollectedandCompiled.pdf> Consulté le : 22.04.2016
- [39] Julien Élie, Éléments de réflexion sur le saint Coran, Version du 24 octobre 2010. Disponible sur : <http://www.trigofacile.com/divers/spiritualite/pdf/reflexions-coran.pdf> Consulté le : 16.04.2016
- [40] Xavier Polanco, text mining et intelligence économique, Disponible sur : <http://lpvist.free.fr/IE/Text%20mining%20et%20IE.pdf> Consulté le : 15.03.2016
- [41] <http://scialert.net/fulltext/?doi=jas.2005.584.587&org=11> Consulté le : 15.03.2016
- [42] https://www.researchgate.net/profile/Mohammed_Al-Kabi/publication/241109392_A_COMPARATIVE_STUDY_OF_THE_EFFICIENCY_OF_DIFFERENT_MEASURES_TO_CLASSIFY_ARABIC_TEXT/links/00b495345aa866eeee000000.pdf Consulté le: 22.05.2016.
- [43] http://link.springer.com/chapter/10.1007%2F978-3-642-22191-0_26#page-1 Consulté le : 22.05.2016.
- [44] https://www.researchgate.net/publication/276854961_Towards_Innovative_System_for_Hadith_Isnad_Processing Consulté le: 22.05.2016
- [45] <http://www.ijitcs.com/2ndicekmt/Kawther+AAldhaln.php> Consulté le : 22.05.2016.
- [46] <http://sunnah.com/bukhari/about>. Consulté le : 20.04.2016.
- [47] http://www.memoireonline.com/12/09/2917/m_Algorithmes-dapprentissage-pour-la-classification-de-documents0.html Consulté le : 20.04.2016.

ملخص

على الرغم من العدد الهائل للمتكلمين باللغة العربية إلا أنها تحوي ظواهر جَدَّ خاصة في المورفولوجيا والنحو، هذه الخصوصية ترجع بصفة خاصة الى تركيبها الإعرابية وتعدد أشكالها وهذا أدى إلى صعوبة المعالجة الآلية لها. الهدف من هذا البحث هو دراسة إحدى الطرق لتصنيف المعارف في قواعد المعطيات النصية مثل تلك الخاصة بالأحاديث النبوية الشريفة هذه الطريقة تمثل عملية لمعالجة مدونة الأحاديث النبوية الشريفة.

في هذه الدراسة استعملنا المصنّف الآلي (Naïve Bayes) المعروف بكفاءته في تصنيف الأحاديث النبوية الشريفة، معتمدين في ذلك على تصنيف صحيح البخاري

الكلمات المفتاحية: اكتشاف المعارف، التنقيب عن البيانات، التنقيب عن النصوص، التلقين الذاتي، اللغة العربية، مدونة الأحاديث النبوية الشريفة

Abstract

Even though, The Arabic language has a very important number of speakers, it present special morpho-syntaxic phenomena. This particularity is mainly related to the inflectional and agglutinative morphology, and the multiplicity of its forms, this leads to a difficulties in the automatic processing. This research aims to study methods of classification in textual databases such as those of Prophet Sayings (Hadiths). This method provide a process of treatment for the corpora of prophetic traditions.

In this study, we use Naïve Bayes classifieur which known by its effectiveness to classify the prophetic sayings based on Sahih Al-Bukhârî classification.

Keywords: Knowledge discovery, Data mining, Text mining, Machine Learning, Texts classification, Arabic language, Prophetic sayings corpora.

Résumé

La langue arabe, bien que très importante par son nombre de locuteurs, elle présente des phénomènes morpho-syntaxiques très particulières. Cette particularité est liée principalement à sa morphologie flexionnelle et agglutinante, et à la multiplicité de ses formes ; cela induit des difficultés de traitement automatique. Cette recherche a pour but d'étudier une méthode de classification des connaissances dans les bases de données textuelles telles que celles des Traditions Prophétiques (Hadiths). Cette méthode représente un processus de traitement du corpus des traditions prophétiques.

Dans cette étude, nous avons utilisé le classifieur Naïve Bayes qui est connu par son efficacité pour classifier les traditions prophétiques en se basant sur la classification de Sahih Al-Bukhârî.

Mots clés: Découverte de connaissances, Fouille de données, Fouille de textes , Apprentissage Automatique , Classification de textes, Langue arabe, Corpus des traditions prophétiques,