UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTE DES SCIENCES ET TECHNIQUES - FES
Centre d'Etudes Doctorales : Sciences et Techniques de l'Ingénieur

N° d'ordre 09/2014

THESE DE DOCTORAT

Présentée par

Mme : Hanane FROUD

Spécialité : Informatique

Sujet de la thèse :

Contributions au Text Mining sur la langue Arabe: Application au Clustering des Documents Texte Arabe

Thèse présentée et soutenue le samedi 22 février 2014 à 09h au Centre de conférence devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Mohammed MEKNASSI	PES	Faculté des Sciences Dhar El Mehraz de Fès	Président
Rachid OULAD HAJ THAMI	PES	ENSIAS de Rabat	Rapporteur
Brahim OUHBI	PES	ENSAM de Meknès	Rapporteur
Noureddine CHENFOUR	PH	Faculté des Sciences Dhar El Mehraz de Fès	Rapporteur
Hicham EL BAHJA	PH	ENSEM de Casablanca	Examinateur
Noureddine RAISS	PES	Faculté des Sciences Dhar El Mehraz de Fès	Examinateur
Said El ALAOUI OUTIK	PH	Faculté des Sciences Dhar El Mehraz de Fès	Examinateur
Abdelmonaime LACHKAR	PH	Ecole Nationale des Sciences Appliquées de Fès	Directeur de thèse

Laboratoire d'accueil : **Sciences de l'Information et Systèmes**

Etablissement : **Ecole Nationale des Sciences Appliquées -Fès**

Table des matières

Liste des tableaux X

Liste des figures XII

Introduction générale 1
1 Contributions 2
2 Organisation de la Thèse 3

CHAPITRE 1: Applications du text mining : Etat de l'art 5
1 Introduction 6
2 Applications du text mining 6
2.1 Définition du text mining (c'est quoi ?) 6
2.2 Système d'indexation et recherche d'information 7
2.2.1 Définitions d'un système d'indexation et recherche d'information 7
2.2.2 Modéles de recherche de documents texte 8
2.2.3 Recherche d'information en langue Arabe 17
2.3 Système question/réponse 20
2.3.1 Systèmes de questions/réponses pour la langue Arabe 21
2.3.2 Architecture d'un système de questions-réponses 21
2.3.3 Fonctionnement d'un système de Questions/Réponses 22
2.4 Résumé automatique des documents texte 23
2.5 Traduction automatique 24
2.5.1 Définitions de la traduction automatique 25
2.5.2 Le processus de traduction 26
2.5.3 Difficultés de la traduction automatique 27
2.5.4 Méthodes des systèmes de la traduction automatique 27
2.6 Catégorisation des documents 28
2.7 Clustering des documents 29
2.7.1 Quelques définitions 29
2.7.2 Clustering partitionnel 30
2.7.3 Clustering hiérarchique 30
2.7.4 Clustering des documents pour la recherche d'information 30
2.8 Clustering des documents : cas de la langue Arabe 34
2.9 Recherche et extraction d'information 36
2.9.1 Recherche d'information 36
2.9.2 Extraction d'information 37
2.10 Extraction des phrases pertinentes à partir des documents texte 37
3 Prétraitement et représentation des documents texte 38
3.1 Prétraitement des documents texte 39
3.2 Représentation des documents texte 40
3.2.1 Représentation vectorielle 40
3.2.2 Représentation conceptuelle 42
3.2.3 Représentation mixte 42
3.2.4 Représentation utilisant les groupes nominaux 43
3.2.5 Représentation simple vs représentation complexe 43

3.2.6 Types des documents 45
4 Modèle d'Analyse Sémantique Latente 46
4.1 Principe 47
4.2 Utilisation de l'Analyse Sémantique Latente pour la langue Arabe 47
5 Réduction de la dimension 48
5.1 Transformation des termes 49
5.2 Sélection des termes 49
5.2.1 Définition 49
5.2.2 Méthodes de sélection des termes 50
6 Corpus Arabe 52
6.1 Corpus AFP 52
6.2 Journal Al-Hayat 53
6.3 Arabic Gigaword 53
6.4 Treebanks 53
6.5 D'autres efforts 53
7 Conclusion 55

CHAPITRE 2: Traitement Automatique du Langage Naturel : Cas de la Langue Arabe 56
1 Introduction 57
2 Description et caractéristiques de la langue Arabe 58
2.1 Particularités de la langue Arabe 59
2.1.1 Alphabet arabe 59
2.1.2 Structure d'un mot arabe 59
2.1.3 Morphologie arabe 61
2.1.4 Catégories des mots arabes 61
2.1.5 Grammaire et caractéristiques de la langue Arabe 61
2.2 Difficultés de l'analyse automatique de la langue Arabe 62
2.2.1 Complexité de la langue Arabe 62
2.2.2 Exemples qui montrent la complexité de la langue Arabe 62
2.2.3 Problème d'encodage 63
2.2.4 Analyse morphologique 63
2.2.5 Segmentation de texte arabe 65
2.2.6 Etiquetage grammatical 65
2.2.7 Analyse syntaxique 65
3 Techniques de prétraitement 67
3.1 Racinisation (Stemming) 67
3.2 Lemmatisation 70
3.3 Tokenisation 71
3.4 Translitération 71
3.5 Elimination des Stop-words 72
3.6 Désambiguïsation 72
3.7 Approches d'étiquetage grammatical 73
3.8 Approches et outils d'analyse morphologique 74
3.8.1 Approches d'analyse morphologique 74
3.8.2 Analyseurs morphologiques 74
3.9 Concordanciers 76
4 Conclusion 77

CHAPITRE 3: Clustering des Documents **78**
1 Introuction **79**
2 Taxonomie de méthodes de Clustering **82**
2.1 Algorithmes de Clustering 84
2.1.1 Approches de Clustering partitionnel 84
2.1.2 Approches de plongement (embedding) géométriques 86
2.1.3 Approches probabilistes 87
2.2 Autres méthodes de Clustering 87
2.2.1 Méthodes basées sur la densité 87
2.2.2 Méthodes basées sur les grilles 88
2.2.3 Clustering basé sur les mots-clés 89
2.2.4 Clustering basé sur un modèle 90
2.2.5 Approches de Clustering des pages web 92
3 Les Défis dans le Clustering des Documents **93**
4 Approches Choisies **94**
4.1 Algorithme de K-means 94
4.2 Algorithme de "Bisecting" K-means 95
4.3 Méthodes de Clustering Hiérarchique 96
4.3.1 Single Link 98
4.3.2 Complete Link 98
4.3.3 Group Average Link 99
4.3.4 Méthode de Ward 100
5 Expériences, Résultats et discussion **100**
5.1 Impact de la Racinisation (Stemming) et de l'utilisation de la Décomposition en Valeurs Singulières sur le Clustering de documents de texte Arabe 100
5.1.1 Expériences 100
5.1.2 Résultats 101
5.1.3 Discussion des résultats 103
5.2 Etude et implémentation de différents algorithmes de Clustering pour les documents texte arabe 103
5.2.1 Expériences 103
5.2.2 Résultats 105
5.2.3 Discussion des résultats 107
6 Conclusion **110**

CHAPITRE 4: Le Modèle d'Analyse Sémantique Latente **111**
1 Introduction **112**
2 Utilisation du modèle LSA en traitement automatique du langage naturel **113**
3 Description du modèle LSA **113**
3.1 Première étape : la représentation du corpus textuel sous la forme d'un tableau 114
3.2 Deuxième étape : la décomposition en valeurs singulières 114
3.3 Troisième étape : la réduction du nombre de dimensions 115
4 Réduction de la dimension avec la technique (DVS) **116**
5 Notre utilisation du modèle LSA pour la langue Arabe **117**
6 Expériences, résultats et discussion **118**
6.1 Expériences 118

6.2 Résultats 119
6.2.1 Résultats avec le Stemmer Léger de Larkey 119
6.2.2 Résultats avec le Stemmer de Khoja 119
6.3 Discussion des résultats 120
7 Conclusion 122

CHAPITRE 5: Résumé Automatique de Texte Arabe 123
1 Introduction 124
2 Intérêt et caractéristiques de résumé automatique de texte 125
2.1 Intérêt de résumé automatique du texte 125
2.2 Concision 125
2.3 Couverture 126
2.4 Fidélité 126
2.5 Cohésion et cohérence 126
3 Différents types de résumé 126
4 Approches globales de résumé automatique de texte 128
4.1 Approche numérique 128
4.2 Approche symbolique 129
4.3 Approche hybride 129
5 Méthodes de résumé automatique de texte 130
5.1 Méthodes de résumé basées sur l'extraction des phrases clefs 130
5.1.1 Méthodes à base de mots clés 130
5.1.2 Méthode à base de position 132
5.1.3 Méthode dépendant de la longueur de phrase 132
5.1.4 Méthode à base d'expressions indicatives (cue methods) 132
5.1.5 Méthode basée sur les relations (cohésion lexicale) 133
5.1.6 Méthode d'exploration contextuelle 133
5.1.7 Méthode hybride 134
5.2 Méthode de Luhn et la méthode basée sur l'Analyse Sémantique Latente (Latent Semantic Analysis (LSA)) 134
5.3 Résumé par mesure de pertinence 135
5.4 Méthodes de résumé pour la langue Arabe 135
5.4.1 Méthode symbolique proposée pour le résume automatique des documents arabes 137
5.4.2 Méthode numérique pour le résume automatique d'articles de journaux en langue Arabe 138
5.4.3 Méthode hybride pour le résume automatique des documents arabes 138
6 Méthode adaptée pour le résumé automatique de texte arabe 139
6.1 Résumé de texte avec l'analyse sémantique latente 139
6.2 Résumé du texte 142
7 Expériences, résultats et discussion 142
7.1 Expériences 142
7.2 Résultats 143
7.2.1 Résultats en utilisant la représentation textuelle complète des documents de la base de données 144
7.2.2 Résultats en utilisant les résumés des documents 145
7.3 Discussion des résultats 146

8 Conclusion **147**

CHAPITRE 6:Extraction des Phrases Pertinentes **148**
1 Introduction **149**
2 Méthodes appliquées pour l'extraction des phrases pertinentes pour les documents texte arabe **150**
3 Nouvelle approche proposée pour l'extraction des phrases pertinentes basée sur les arbres de suffixes **151**
3.1 Arbre des Suffixes 151
3.2 «Nettoyage» du Document 151
3.3 Modèle de l'Arbre des Suffixes d'un Document 151
4 Expériences, résultats et discussion **154**
4.1 Description des Expériences 154
4.2 Résultats 155
4.3 Discussion 155
5 Conclusion **158**

Conclusion et perspectives **159**

Bibliographie **162**

Annexe A: Corpus de test de documents texte arabe **175**

Annexe B: Exemple de document texte arabe **177**

Annexe C: Techniques de fusion pour l'algorithme agglomératif hiérarchique **179**

Annexe D: Evaluation de la qualité des résultats du Clustering des documents texte arabe **181**

Annexe E: Mesures de Similarité **182**
1 Métrique **183**
2 Distance euclidienne **183**
3 Similarité Cosinus **184**
4 Coefficient de Jaccard **184**
5 Coefficient de Corrélation de Pearson **184**
6 Divergence de Kullback –Leibler moyenne **185**
7 Distance de Chi-deux χ^2 ..... **186**
8 Similarité de Dice **186**
9 Distance de Manhattan dans le plan **186**