UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTE DES SCIENCES ET TECHNIQUES-FES

N° d'ordre: 07/2014

Fes , le 28 Avril 2014

# THESE

*En vue d'obtenir le grade de*

**DOCTEUR DE L'UNIVERSITE SIDI MOHAMED BENABDELLAH**

Discipline : Mathématique et Informatique
Spécialité : Recherche Opérationnelle et Informatique

Préparée à l'UFR Calcul Scientifique et Informatique, Science de l'Ingénieur de la Faculté des Sciences et Techniques de Fès

Par : Mr. Abdelatif ES-SAFI

# Contribution à l'analyse de données et à la fouille documentaire: Amélioration de certains algorithmes et proposition de nouveaux modèles

Directeur de thèse : Mohamed ETTAOUIL
Soutenue le Lundi 28 Avril 2014 devant le jury :

| Nom & Prénom | Grade | Etablissement | |
|---|---|---|---|
| **Hassan QJIDAA** | PES | Faculté des Sciences de Fès | **Président** |
| **Azzedine MAZROUI** | PES | Faculté des sciences d'Oujda | **Rapporteur** |
| **Ahmed ELHILALI ALAOUI** | PES | Faculté des sciences et technique de Fès | **Rapporteur** |
| **Abdellatif ELAFIA** | P.H | ENSIAS de Rabat | **Examinateur** |
| **Abdelhak LAKHOUAJA** | PES | Faculté des sciences d'Oujda | **Examinateur** |
| **Mohamed ETTAOUIL** | PES | Faculté des sciences et technique de Fès | **Directeur** |
| **Youssef GHANOU** | PA | EST de Meknès | **Invité** |

Etablissement : Faculté des Sciences et Techniques-Fès

**Résumé de la these**

Dans ce travail, nous avons proposé quelques approches permettant de soulever certains problèmes concernant l'apprentissage automatique et la fouille textuelle. Ces approches peuvent être résumées comme suit :
La première approche concerne l'amélioration de l'algorithme de Kohonen : Dans ce contexte, afin de chercher une architecture adéquate, nous avons proposé un processus itératif qui intègre une phase de sélection permettant de corriger les erreurs commises dans la phase d'initialisation. Dans une deuxième approche nous avons proposé des techniques permettant de déterminer automatiquement le nombre de classes. Ceci permet de soulever un grand problème dont souffrent les principaux classifieurs non supervisés. En ce qui concerne la fouille textuelle, nous avons proposé une nouvelle technique concernant la représentation automatique de texte.

**Mots clés :** Analyse de données, apprentissage, optimisation d'architectures, classification, fouille textuelle, résumé automatique.

**Abstract:**

In this thesis we address some of important problems concerning the learning tools and data mining. These contributions presented in this work can be summarized as follows:

The first approach aims to improve Kohonen algorithm: In this context, in order to search an appropriate architecture of Kohonen map. We proposed an iterative process in which we integrate a selection phase allowing correcting the errors committed in the initialization phase. In the second contribution we used a technique to determine automatically the number of clusters which is a great problem in unsupervised clustering domain. About the text mining, we proposed some news techniques to represent automatically a text. We have also presented two works concerning the automatic summarization of text. Finally we presented a project on plagiarism.

**Key words:** learning, optimization of architectures, clustering, text mining, automatic
Summarization, plagiarism.

# Table de Matières


INTRODUCTION GENERALE ..... 7
partie I : Analyse de données ..... 12
CHAPITRE I. Généralités sur l'analyse des données ..... 14
1. Introduction ..... 14
2. Structures de l'analyse de données ..... 14
2.1. Notion de données ..... 15
2.2. Notion de variables ..... 20
2.3. Similarité et distance ..... 24
2.4. Outils de l'analyse de données ..... 27
2.5. Analyse factorielle ..... 28
3. Conclusion ..... 35
CHAPITRE II. Apprentissage automatique et Classification ..... 37
1. Introduction ..... 37
2. Apprentissage automatique ..... 37
2.1. Introduction ..... 37
2.2. Définition et objectif de l'apprentissage ..... 38
2.3. Différents types d'apprentissage ..... 38
2.4. Simulation de l'apprentissage humain : Réseaux de neurones artificiels ..... 42
2.5. Problèmes d'apprentissage ..... 68
3. Classification des données ..... 70
3.1. Introduction ..... 70
3.2. Principe et objectif ..... 71
3.3. Différentes approches de classification ..... 73
3.4. Principales méthodes de classification ..... 74
3.5. Evaluation de la qualité de la classification automatique ..... 102
3.6. Problèmes des classificateurs ..... 106
4. Conclusion ..... 106
CHAPITRE III. Contributions concernant la première partie ..... 107
1. Introduction ..... 107
2. Réduction de l'architecture de la carte de Kohonen ..... 107
2.1. Etat d'art ..... 108
2.2. Méthode proposée ..... 110
3. Amélioration de l'algorithme de Kohonen et recherche d'une architecture optimale ... 117
3.1. Approche proposée ..... 118

3.2.Résultats expérimentaux ..... 120
4. Amélioration de K-moyennes par la recherche du nombre optimal de classes ..... 124
4.1.Présentation de la méthode proposée ..... 126
4.2.Evaluation de la méthode ..... 127
4.3.Conclusion et perspectives ..... 134
5. Recherche du nombre optimal de classes et optimisation de la carte de Kohonen ..... 135
5.1.Méthode proposée pour optimiser la carte de Kohonen ..... 135
5.2.Evaluation de la qualité de la méthode proposée ..... 137
5.3.Interprétation des résultats expérimentaux ..... 138
6. Conclusion ..... 139
7. Résumé de la première partie ..... 139
partie II : Fouille documentaire ..... 142
CHAPITRE IV. Préparation des documents ..... 144
1. Introduction ..... 144
2. Prétraitement des documents ..... 144
2.1.Segmentation des textes ..... 144
2.2.Filtrage des documents ..... 145
2.3.Utilisation des synonymes ..... 146
2.4.Principe de stemmatisation ..... 146
2.5.Principe de lemmatisation ..... 146
3. Représentation des documents ..... 147
3.1.Principales méthodes de représentations des documents ..... 148
3.2.Exemples de représentation intégrant moyennement la sémantique ..... 151
3.3.Problèmes de la représentation ..... 153
4. Nouvelle approche pour représenter les documents ..... 153
4.1.Présentation de la méthode proposée ..... 154
4.2.Expérimentation ..... 156
4.3.Conclusion et perspectives ..... 158
5. Conclusion ..... 159
Conclusion et perspectives ..... 161
Annexe 1 : Algorithme De Kohonen ..... 164
Annexe 2 : Principaux Critères De Sélection De Variables Dans L'arbre De Décision ..... 165
Annexe 3 : Exemple Illustrant L'apprentissage Statistique Des Réseaux Bayésiens ..... 168
Annexe 4 : Détermination des classes par la loi de Bayes ..... 171
Annexe 5 : Quelques critères d'évaluation des classes ..... 173
Bibliographie ..... 175