# Introduction

Les sciences de l'information et de la communication ont connu durant ces dernières décennies une évolution technologique remarquable après l'invention de l'internet. Cette dernière a contribué énormément à la globalisation du monde et à la réalisation du « village planétaire » par le biais du nombre croissant d'informations mises en ligne chaque jour, le développement continu des infrastructures de communication, ainsi que la progression constante du nombre de personnes connectées au réseau mondial dont la langue maternelle n'est pas l'anglais [Peters and Sheridan, 2001]. Le nombre de sites en ligne, à l'heure actuelle, est estimé à 206,026,787 selon Netcraft[1] dont la majorité du contenu textuel est écrite en plusieurs langues (anglais, français, russe, chinois, arabe, persan, hébreu, etc.). L'organisation de toute cette immense et gigantesque ressource est donc indispensable. Ainsi, les techniques de la fouille de textes par le biais de l'apprentissage artificiel, dont la classification automatique des documents fait parti, s'avèrent d'être pertinentes et très efficaces.

Depuis les années 1960, les chercheurs se sont intéressés à la question de la classification automatique des documents. La majorité des travaux a été réalisée sur des documents écrits en caractères latins (français, anglais, espagnol, etc.). En revanche, très peu des travaux se rapportent à la classification automatique des documents écrits en caractères arabes malgré la richesse morphologique de cette langue. Pour construire un modèle efficace de classification automatique des documents à catégoriser, le traitement automatique de ces derniers est un élément essentiel. Lors de ce traitement, la richesse morphologique de la langue arabe entraine des difficultés qui empêchent les chercheurs d'adopter directement les méthodes et les résultats obtenus par les travaux portant sur la classification automatique des documents écrits en caractères latins sans mettre en question leur validité. D'où la nécessité de tester la fiabilité et l'efficacité des solutions existantes sur des documents écrits en caractères arabes avant de pouvoir tirer des conclusions décisives.

[1] http://news.netcraft.com/archives/web_server_survey.html selon leur dernière enquête en Mai 2010. [Date de dernière visite : Mai 2010].

Il est indispensable de noter que, pour surmonter les difficultés liées au traitement automatique de l'arabe, nous faisons recours à un analyseur morphologique très performant qui fait partie d'un ensemble d'outils basé sur le « DIctionnaire INformatisé de l'Arabe » (DIINAR.1). L'accouplement d'un système à base de statistiques (celui des algorithmes d'apprentissage artificiel) et d'un système à base de règles (celui de DIINAR.1) a permis de réaliser tous les travaux menés dans cette thèse et d'enrichir cette dernière sur les plans théorique et pratique comme nous le verrons plus tard.

Le processus de la classification automatique des documents est composé de plusieurs éléments dont chacun joue un rôle très important dans le résultat final. Nous citons, entre autres, le choix de la nature des attributs qui mérite d'être fait avec le plus grand soin puisqu'il a un effet direct sur la fidélité des classifieurs. Dans la majorité de la littérature portant sur la classification automatique des documents écrits en caractères arabes le choix est fait, en général, entre l'approche statistique (exemple, les n-grammes) et l'approche basée sur l'analyse morphologique (exemple, les racines et les lemmes). Chacune des deux approches possède ses propres avantages incitant les chercheurs à la favoriser : ceux qui préfèrent utiliser les n-grammes argumentent que ces derniers sont indépendants de la langue et plus faciles à générer, tandis que ceux qui favorisent les racines ou les lemmes disent que l'analyse morphologique est capable de surmonter la richesse morphologique des langues ayant une morphologie aussi riche que l'arabe. Toutefois, l'analyse morphologique adoptée par la plupart de ces travaux n'est pas suffisamment profonde pour prendre en compte les spécificités majeures de la langue arabe. Cette démarche nous a empêché de tirer une conclusion décisive sur le meilleur choix à faire et nous a incité à mener dans cette thèse une étude comparative entre la performance des réseaux bayésiens naïfs et celle des machines à vecteurs de supports en utilisant plusieurs corpus conçus à partir des n-grammes, des racines, des lemmes[2], et des mots pour pouvoir tirer une conclusion sur la meilleur approche à adopter dorénavant.

[2] Les lemmes et les racines sont extraits par l'analyseur morphologique de notre équipe qui performe une analyse morphologique profonde en faisant appel à la base de données DIINAR.1. Pour un exemple d'un document analysé morphologiquement voir l'annexe page 194.

Le choix de l'algorithme d'apprentissage se présente comme un autre élément essentiel pour une classification automatique efficace. La plupart des travaux menés sur des documents écrits en caractères arabes se basent sur des algorithmes d'apprentissage récents comme, par exemple, les machines à vecteurs de support[3], les réseaux bayésiens naïfs[4], et les arbres de décision[5] qui sont connus pour être parmi les plus performants classifieurs du domaine [Al Harbi, 2008], [Bawaneh, Koffash, Al Rabea, 2008], [Duwairi, 2007] et [El Kourdi, Bensaid, et Rachidi, 2004]. Néanmoins, à l'heure actuelle, aucun de ces travaux n'utilise l'algorithme du « dopage » (en anglais, « Boosting ») malgré le fait qu'il est conçu spécifiquement pour cette tâche. Cet algorithme se base sur le principe de combinaison de classifieurs dont l'idée générale est la suivante : il est légitime de croire que devant une tâche nécessitant la connaissance d'un expert, plusieurs experts travaillant ensemble et combinant adéquatement leurs jugements pourraient être plus performants que le jugement d'un seul expert pris séparément. Dans ce paradigme, « le dopage » regroupe de nombreux algorithmes (que l'on appelle des « apprentis faibles») qui s'appuient sur des ensembles de classifieurs binaires[6] faibles pour optimiser leurs performances. Les travaux utilisant cet algorithme confirment que « le dopage » produit de très bons résultats. Toutefois, ils sont menés sur des documents écrits en caractères latins ce qui nous empêche d'adopter leurs résultats pour juger son efficacité et sa performance sur des documents écrits en caractères arabes. Ainsi, nous menons dans cette thèse une étude comparative basée sur des documents écrits en caractères arabes, entre les arbres de décision dopés[7] d'une part et les arbres de décision sans dopage, les machines à vecteurs de support, et les réseaux bayésiens naïfs d'une autre part.

Une pratique d'écriture généralement utilisée dans la littérature arabe, surtout celle abordant des sujets scientifiques et techniques, consiste à introduire quelques mots écrits en caractères latins avec ceux écrits en caractères arabes pour des raisons multiples ; par exemple, il n'existe pas un mot arabe portant suffisamment de fidélité

---

[3] cf. §3.6 page 70.
[4] cf.§3.5 page 65.
[5] cf.§3.4 page 57.
[6] Une classification binaire est un système disposant uniquement de deux catégories différentes qui cherche à classer un objet dans une de ces catégories. Par contre, une classification multi-classes dispose de plus que deux catégories et peut classer un objet dans une ou plusieurs de ces catégories.
[7] i.e. on a utilisé les arbres de décision comme un apprenti faible.

au sens ou à l'indication du mot écrit en caractères latins pouvant le remplacer, ou bien il est important de mentionner le mot dans sa langue d'origine puisqu'il a été traduit ou emporté de cette langue, ou bien l'auteur trouve que sa présence rend le texte plus clair. Quoique ce soit la raison pour laquelle les mots écrits en caractères latins sont mélangés avec les mots écrits en caractères arabes, il est important de souligner le fait que tous les travaux portant sur la classification automatique des documents écrits en caractères arabes traitent ces documents d'un point de vue monolingue i.e. ils exploitent uniquement les mots écrits en caractères arabes et éliminent tous les mots écrits dans d'autres langues. Cette pratique entraîne une perte d'une partie fidèle au sens des sujets traités dans les documents sachant qu'elle aurait pu contribuer à la classification. Cette perte soulève le problème du degré de la subjectivité de la décision finale prise par le modèle de prédiction. Pour cette raison, nous nous intéressons aussi dans cette thèse à la classification automatique des documents d'un point de vue multilingue. Cette prise en compte du cas multilingue entraîne des défis supplémentaires qui rendent certains aspects de la classification automatique inopérants et nous incite à leur trouver des solutions convenables. L'élément le plus affecté par cette question est celui de la réduction de la taille du vocabulaire par le biais de la sélection d'attributs. Pour cela, nous avons interrogé, dans cette thèse, six des méthodes largement connues et utilisées dans la littérature et nous avons constaté qu'elles sont incapables de préserver un nombre considérable des mots écrits en caractères latins dans le contexte d'une classification d'un point de vue multilingue. Pour résoudre ce problème, nous proposons plusieurs solutions et approches capables de conserver un nombre significatif d'attributs écrits en caractères latins, par rapport au nombre total d'attributs préservés lors de la réduction de la taille du vocabulaire. En plus, les expérimentations basées sur les corpus générés par ces méthodes produisent un taux de fidélité de classification aussi bon, voire meilleure dans la plupart du temps, que celui obtenu en se basant sur les corpus générés par les six méthodes connues.

Les points cités ci-dessus sont abordés dans six chapitres. Les cinq premiers s'intéressent à la classification automatique des documents en général, ainsi que celle des documents écrits en caractères arabes d'un point de vue monolingue i.e. où les mots écrits en caractères latins sont éliminés. Le chapitre 1 définit la classification

automatique des documents, son utilité, ainsi que les difficultés souvent rencontrées par les chercheurs du domaine. Le chapitre 2 présente les différents éléments de la classification automatique tels que la représentation du corpus, la réduction de la taille du vocabulaire et les méthodes utilisées pour cet effet, et finalement le calcul du poids des attributs. Le chapitre 3 détaille les algorithmes d'apprentissage artificiel utilisés dans tous les travaux menés dans cette thèse ainsi qu'une analyse de leurs avantages et de leurs inconvénients. La conception des jeux de données utilisées dans les travaux de cette thèse ainsi qu'une étude comparative entre la performance des réseaux bayésiens naïfs et celle des machines à vecteurs support pour trouver le meilleur choix de la nature d'attribut sont présentées dans le chapitre 4. Le chapitre 5 présente une comparaison de la fidélité des arbres de décision dopées d'une part et les arbres de décision (sans dopage), les machines à vecteurs de support, et les réseaux bayésiens naïfs d'une autre part pour trouver le meilleur choix d'algorithme d'apprentissage permettant d'atteindre le meilleur taux de fidélité de classification automatique des documents écrits en caractères arabes.

Par contre, le chapitre 6 aborde la classification automatique des documents écrits en caractères arabes d'un point de vue multilingue i.e. en exploitant simultanément les mots écrits en caractères arabes et ceux écrits en caractères latins. Nous présentons dans ce chapitre quelques motivations incitant à l'organisation des documents multilingues (dont l'arabe fait parti). La classification automatique de ces derniers entraîne des défis supplémentaires qui rendent quelques éléments de la classification automatique incompétents et nous oblige de leur trouver les solutions convenables. L'élément le plus affecté par ce point de vue est la réduction de la taille du vocabulaire par le biais de sélection d'attributs. Nous avons testé six méthodes largement connues et utilisées dans la littérature et nous avons constaté qu'elles sont incapables de préserver un nombre considérable des mots écrits en caractères latins sauf si le seuil d'attributs préservés lors de la sélection est assez large ($\geq$ 2000 attributs). Ainsi, nous proposons, dans ce chapitre, des méthodes et des approches de réduction de la taille de vocabulaire capables de conserver, par le biais d'une sélection, un nombre significatif d'attributs écrits en caractères latins même avec un seuil très bas. Les expérimentations menées sur ces méthodes confirment que les taux

de fidélité de classification obtenus sont aussi bons, voire meilleurs dans la plupart du temps, que ceux obtenus en se basant sur les six méthodes connues.