Conclusion Générale

Les travaux présentés dans ce mémoire sont axés sur l'étude des nouvelles méthodes de classification et de segmentation des connaissances dans les bases de données textuelles telles que celles des Traditions Prophétiques (Hadith). Nous nous sommes intéressés dans ce mémoire à proposer des solutions répondant aux différentes problématiques inhérentes à ces méthodes.
Cette conclusion se divise en trois parties :

- Dans la première partie, nous avons étudié les Corpus Prophétiques et nous avons exploré les recherches faites dans le domaine de traitement d'information dans ce domaine. Les bases théoriques de la recherche d'information, la fouille de texte et les méthodes essentiellement statistiques implémentées dans ce système ont été exposées. Une place particulière a été accordée à la description des algorithmes utilisés pour l'indexation et la recherche. Les résultats expérimentaux de nos recherches dans ce domaine ont visé à mettre en œuvre un système d'extraction de connaissances prophétiques basé sur une approche de fouille de données textuelles. L'étude a porté principalement sur des corpus de traditions prophétiques «Hadith», pour des fins de découverte thématique.
- Dans la deuxième partie, nous avons présenté un état de l'art global sur la classification automatique de textes en s'intéressant plus particulièrement à la question de la représentation des documents traités par un classificateur. Les techniques de sélection et d'extraction d'attributs ont été aussi exposées. Cette partie a fait la lumière sur le processus d'évaluation des classificateurs. Nous nous sommes intéressés à l'évaluation d'un système de classification basé sur l'algorithme des arbres de décision. Nous avons également analysé les différents résultats d'évaluation obtenus pour nos corpus ainsi que les résultats de leur comparaison ce qui nous a permis de conclure qu'une série facteurs peuvent influencer le fonctionnement du système de classification en particulier la nature et la spécificité des documents de chaque corpus. Ces études ont été complétées par le développement d'un modèle de classification des textes prophétiques utilisant les réseaux de neurones et la méthode de décomposition en valeur singulière (SVD). Comme résultat tangible, la dimension réduite des vecteurs a diminué le temps de calcul de la phase d'apprentissage pour les deux réseaux de neurones MLP et RBF. Les expériences sur le corpus "Hadith" ont montré que le modèle MLP est efficace pour la représentation et la classification des documents arabes.
- Dans la troisième et dernière partie, nous avons essayé de décrire les différentes approches de la segmentation thématique à savoir les approches linéaires ou hiérarchique et les approches supervisées ou non supervisées. L'état de l'art du domaine de segmentation thématique nous a permis de distinguer trois grandes familles de méthodes. une analyse comparative des algorithmes de segmentation thématique TextTiling et C99 a été présentée pour un corpus de textes arabes. Cette comparaison a confirmé que la tâche de segmentation est dure à évaluer, cela est dû à la variation des objectifs. Nous avons conclu cette partie par l'étude et l'évaluation des techniques utilisées pour l'organisation de l'information pertinentes, afin d'améliorer l'accès à l'ensemble des résultats retournés par le système de recherche d'information. Les résultats ont démontré que la reformulation de requêtes est une phase très importante pour les systèmes de recherche d'information. Nous nous sommes intéressés à la

réorganisation des réponses fournies par un SRI tout en démontrant que le fait de considérer des passages thématiques peut s'avérer utile pour collecter de l'information pertinente, ainsi que le regroupement des segments thématiques selon leurs catégories thématiques, permet de fournir à l'utilisateur un aperçu des différentes thématiques qu'il peut trouver dans le corpus, en rapport avec une requête.

Perspectives et travaux futurs

Pour nos futurs travaux en ce qui concerne la classification, nous envisageons d'effectuer plus d'expériences afin de trouver les meilleurs facteurs qui peuvent améliorer les performances de notre système de classification. Notre but est de passer de la classification mono-classe (*hard classification*) où le document peut être classé dans une seule catégorie à la classification multi-classe (*soft classification*) où le document peut être classé dans plus d'une catégorie. D'autres méthodes de sélection et de réduction d'attributs peuvent être considérées pour nos futures études. Il serait aussi utile d'utiliser une plus grande collection de données pour améliorer la capacité d'apprentissage de langage pour les modèles utilisés. Finalement, nous prévoyons de mener plus de comparaisons avec d'autres algorithmes d'apprentissage employés dans la littérature pour la catégorisation du texte tel que SVM, Algorithme Génétiques et algorithme de Boosting.

Concernant la segmentation thématique et pour aller plus loin dans les expérimentations, nous allons essayer de nouveaux algorithmes fusionnant des méthodes supervisées avec des méthodes non supervisées ainsi que de nouvelles comparaisons entre les approches statistiques et les approches linguistiques. Une interaction avec l'utilisateur peut être mise en place en lui permettant de sélectionner les segments qui l'intéressent le plus parmi les segments du document segmenté et en relançant le processus de recherche en optimisant un critère prenant en compte les choix de l'utilisateur ce qui pourrait permettre de distinguer des thématiques qui y sont fortement liées.

Enfin, nous envisageons l'utilisation d'un modèle standard pour la représentation de connaissances dans les corpus prophétiques. Les systèmes de recherche d'informations prophétiques futurs doivent être capables de fournir les informations suffisantes pour l'accomplissent de la tâche de prise de décision par les experts du domaine. Nous visant aussi à améliorer la portabilité en fournissant un formalisme commun pour la représentation de ces documents sous forme d'une structure sémantique par 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.