Introduction Générale

L'apparition et la popularisation des ordinateurs, des documents électroniques, de différents types de support pour stocker les documents et des réseaux de télécommunication ont profondément bouleversé les liens entre l'homme et l'information. Les mémoires magnétiques et optiques ont permis un stockage des documents électroniques de très bonne qualité, pour un coût qui ne cesse de diminuer depuis leur création. Les réseaux de télécommunication permettent leur diffusion et un échange rapide et plus simple que jamais. Un exemple de cette révolution est la récente annonce de la firme Google, qui gère le moteur de recherche actuellement le plus utilisé sur Internet, de créer la plus grande bibliothèque en ligne ayant jamais existé, et ce en numérisant 15 millions d'ouvrages. Le développement d'outils automatisés permettant un accès efficace à cette quantité gigantesque d'information numérique apparait comme une nécessité [Sauvagnat, 2005].

Ces nouveaux flux d'information créent des besoins réels dans le domaine de traitement de l'information textuelle. Plusieurs communautés informatiques sont actuellement en pleine évolution pour combler ces besoins et on assiste à l'émergence d'un domaine qui est au confluent de leurs préoccupations et que l'on appellera «l'accès à l'information textuelle ». Ce dernier est l'outil le plus communément offert aux utilisateurs pour accéder à des grandes collections de données, c'est un ensemble de systèmes qui s'intéressent à l'analyse et au traitement des informations textuelles. Dans cette perspective, le système qui s'occupe des problèmes d'indexation automatique et de recherche d'information est connue sous le nom de Système de Recherche d'Information (SRI) [Harrag, 2005].

L'objectif principal des Systèmes de Recherche d'Information (SRI) est de répondre au besoin en information des utilisateurs. Les utilisateurs interrogent, au moyen d'une requête, une base de documents numériques et les SRI leur renvoient une liste de documents susceptibles de répondre à leur besoin. Néanmoins, de nombreux problèmes se posent, notamment en ce qui concerne la formulation de la recherche, la représentation des textes ou la présentation des résultats. En effet, comment exprimer un besoin d'information qui puisse être compréhensible par le système? Comment formuler de manière précise sa recherche alors qu'il nous manque justement des informations sur le sujet? Comment caractériser le contenu des documents afin de pouvoir le rapprocher de la question posée? Comment estimer la pertinence des documents? Sous quelle forme produire les résultats de la recherche pour qu'ils soient le plus facilement interprétables par l'utilisateur?

Contexte de travail

Notre travail se situe dans le contexte de la recherche d'information et 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 présentation des résultats d'un système de recherche d'information. L'objectif est de permettre à un utilisateur de cerner rapidement les différents aspects de la requête formulée. Typiquement, la liste des réponses d'un système de recherche d'information à une requête est ordonnée selon un score de pertinence [Lamprier, 2008]. C'est dans le contexte de l'organisation et de la localisation des informations pertinentes que se situent plus particulièrement nos travaux. Nous nous plaçons plus précisément dans le cadre de la classification et la segmentation thématique des documents retournés par le système de recherche d'information pour permettre une localisation facilitée des documents pertinents [Tombros et al., 2002].

Problématique

La Tradition prophétique est l'ensemble des dires, des actions et des décisions du Prophète (que la paix soit sur lui) et est considérée par les Musulmans, comme la deuxième partie du discours religieux en Islam. C'est la Sagesse dont Dieu dota le Prophète quand Il lui a révélé le Saint-Coran. Dieu a dit : «Dieu t'a révélé le Livre et la Sagesse, Il t'a enseigné ce que jamais tu n'aurais pu savoir par toi-même! La faveur de Dieu a été immense à ton égard.» [Les Femmes : 133]. Ce que dit le Prophète, ce qu'il transmet aux hommes avec toute la sincérité d'un homme probe est la pure vérité, et Dieu le démontre bien quand il dit : «Ses propos ne procèdent pas de sa propre inspiration. C'est uniquement révélation inspirée» [L'Etoile : 3-4]. La Tradition a pour rôle d'expliciter ce qui a été dit brièvement dans le Saint-Coran. Elle explique en détails ses lois et ses préceptes. Les thèmes abordés par la Tradition sont nombreux. On y trouve de tout : des ordres et des interdictions, des directives et des conseils, des prières et des invocations, etc … Tous les dires du Prophète, ses actions et ses décisions sont des explicitations des préceptes islamiques dont la base est le Saint-Coran. Dans ce sens, on peut dire que la Tradition prophétique est le prolongement et le complément du Saint-Coran.

L'importance du Hadith aux yeux des 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) pour l'ensemble de la nation musulmane dans tous les coins de la terre. Malgré ça, nous constatons, que les systèmes de recherche d'informations prophétiques existants sont incapables de fournir les informations suffisantes pour l'accomplissent de la tâche de prise de décision par les experts du domaine. En effet, l'application des techniques de la fouille de textes sur les corpus prophétiques apporte de nombreux avantages aux applications qui concernent le traitement informatique des documents Prophétique. Elle permet en particulier d'automatiser une partie des traitements effectués sur les documents Prophétiques.

A présent, notre problématique se situe dans le cadre de 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). Ces méthodes représentent un processus de structuration de l'ensemble des informations et des réponses fournies par un système de recherche d'information. La classification thématique permet à l'utilisateur d'orienter son exploration en fonction des thématiques générales des classes et d'accéder ainsi plus rapidement au but de sa recherche. La segmentation permet de présenter à l'utilisateur les segments textuels jugés pertinents et de mieux positionner certains documents longs dans lesquels l'information recherchée ne constitue que l'une des thématiques abordée. Cela s'applique notamment à des données en quantité trop importante pour qu'une étude visuelle soit possible comme dans le cas du corpus du Hadith, par exemple.

Contribution

Basée sur une contribution antérieure [Harrag, 2005], le présent travail, dans le cadre de la fouille de texte dans les corpus des traditions prophétiques, a pour but essentiel de porter un intérêt particulier aux interactions suivantes :

- Etant donné que le domaine d'application est les Traditions Prophétiques, il est nécessaire de présenter ce domaine et de poser sa problématique. Nous proposons d'étudier en détail la structure des corpus prophétiques, leur historique, leurs méthodes et leur finalité.
- Les systèmes de recherche documentaire basés sur des méthodes essentiellement statistiques permettent le traitement de requêtes en langage naturel sur des corpus hétérogènes et volumineux. Nous nous intéressons aux systèmes de recherche basés sur les indices de ressemblance. Un tel système calcule la distance entre la requête et chacun des documents disponibles. Nous Tentons d'apporter une solution aux problèmes suivants:

- Selon les valeurs des indices de recherche, une liste ordonnée de documents est fournie à l'utilisateur. Cependant cette liste est souvent trop longue pour être pleinement exploitable par l'utilisateur. En effet, parce que mal positionnés, certains documents quoique pertinents, ne sont jamais explorés par l'utilisateur.
- Les sujets abordés par certains documents sont multiples et certains sont éloignés des thématiques recherchées par l'utilisateur, soit parce que ces dernières ne sont pas clairement exprimées dans la requête soit parce que le système n'a pas su les prendre correctement en compte.

- Parmi les méthodes qui permettent la structuration des réponses aux requêtes figure la classification thématique [Sebastiani, 2002]. Elle permet à l'utilisateur d'orienter son exploration en fonction des thématiques générales des classes et d'accéder ainsi plus rapidement aux résultats demandés. Une individualisation des thématiques abordées par les documents retournés par une recherche préliminaire pouvait conduire à l'obtention de groupes de résultats représentant mieux les différentes notions abordées que lorsque les documents sont pris en compte de manière globale. Les catégories thématiques présentées à l'utilisateur permettent ainsi une meilleure localisation des informations pertinentes. Dans ce mémoire, nous étudierons le comportement des algorithmes d'arbres de décision et de réseaux de neurones artificiels dans le domaine de classification thématique des textes.
- La segmentation thématique a pour but de diviser le texte en segments [Chuang & Chien, 2005]. Elle permet de présenter à l'utilisateur les segments textuels jugés pertinents et de mieux positionner certains documents longs dans lesquels l'information recherchée ne constitue que l'une des thématiques abordée. À partir de cette segmentation, un nouveau calcul des indices de ressemblance est réalisé entre les segments et la requête aboutissant à une nouvelle liste dont les éléments peuvent eux-mêmes être classés pour obtenir une segmentation plus raffinée. Nous proposons d'étudier le fonctionnement des deux algorithmes de segmentation thématique TextTiling et C99 sur des corpus en langue arabe et notamment les corpus des traditions prophétiques.

Nous proposons d'explorer ces différents points afin de mettre en place un système permettant de présenter à l'utilisateur une liste de représentants des catégories constituant un bon aperçu des différents types d'information qu'il pourra trouver en rapport avec sa requête dans le corpus de textes interrogé. L'objectif final est d'extraire les parties les plus intéressantes d'un ensemble de documents (les documents retournés par un système de recherche classique) afin de présenter à l'utilisateur une liste de passages de texte lui permettant de sélectionner les aspects, et donc les groupes de passages, qui lui semblent correspondre au mieux à ses besoins.

Organisation du mémoire

Cette thèse s'articule autour de trois parties. La première présente le contexte dans lequel se situent nos travaux, c'est à dire la fouille des données textuelles et plus précisément la recherche d'information dans les corpus prophétique:

- Le premier chapitre de cette partie traite les concepts liés aux corpus textuels proprement dits. Il est consacré à la description des corpus en langue arabe. Dans ce chapitre, nous nous intéressons à l'étude des corpus prophétiques, nous passons en revue toutes les travaux à caractère commercial et scientifique qui visent à informatiser ces corpus.
- Dans un second chapitre nous réalisons un survol des principales composantes des systèmes de fouille de textes et de recherche d'information; ce chapitre s'intéresse aussi à l'évaluation de ces systèmes.
- Le troisième chapitre est réservé à l'expérimentation des notions et des modèles rencontrés dans les deux premiers chapitres de cette partie. Dans ce chapitre, nous présentons une description textuelle du corpus « *Sahîh Al-Boukhârî* » 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, nous visons à évaluer un système de fouille de textes basé sur le modèle vectoriel pour le classement des résultats retournés lors d'une première recherche.

Cherchant à proposer des catégories de résultats plus représentatives des différentes notions abordées par les documents retournés par une recherche préliminaire, la seconde partie s'intéresse à classification thématique des textes prophétiques :

- Un premier chapitre introduit le problème de classification automatique de texte. Dans ce chapitre, nous présentons les différentes stratégies de représentation des documents traités par un classificateur, les techniques de sélection et d'extraction d'attributs sont aussi exposées. Nous détaillons le fonctionnement des classificateurs les plus connus dans le domaine et nous abordons les critères d'évaluation des classificateurs.
- Le deuxième chapitre se focalise alors sur les résultats d'évaluation de l'algorithme des arbres de décision dans la tâche de classification des textes arabes. Les expérimentations sont déroulées sur deux corpus en langue arabe. Ce chapitre s'intéresse à l'étude de l'impact des méthodes de sélection d'attribues sur les performances du processus de classification par arbres de décision.
- Le troisième chapitre concerne la présentation d'un modèle basé sur les réseaux de neurones (NN) pour la classification des textes prophétiques. Nous proposons d'utiliser la *Décomposition en Valeur Singulière* (SVD) dans la phase de prétraitement du modèle de réseaux de neurones pour réduire la dimension des termes. La technique SVD permet donc d'améliorer les performances de classification et d'accélérer la convergence du processus d'apprentissage pour le modèle des réseaux de neurones.

La dernière partie concerne la segmentation thématique des documents. Nous intéressons à l'étude de l'impact de la prise en compte des segments thématiques sur la pertinence globale des documents :

- Dans le premier chapitre, nous introduisons la problématique de la segmentation thématique des textes, sa définition, ses objectifs, ses approches et ses grandes familles de méthodes.
- La motivation principale du deuxième chapitre est d'étudier l'efficacité des algorithmes à base de cohésion lexicale comme un moyen de segmentation thématique des textes arabe. Ce chapitre présente une vue d'ensemble des approches implémentés. Une bonne partie de ce chapitre est réservée à l'évaluation des résultats obtenus par les deux algorithmes implémentés.
- Le troisième chapitre détaille les différents éléments de nos approches pour l'organisation de l'information pertinente. Le but de ce chapitre est de réorganiser les résultats retournés par un SRI. Nous évaluons l''impact de la technique de retour de pertinence sur l'amélioration des performances. Nous effectuions une évaluation de la capacité des méthodes de segmentation à mettre en valeur certains documents à partir de leur segmentation et nous explorons les différentes pistes pour considérer ces segments lors de la production de groupes thématiques de résultats.

Enfin, une conclusion générale synthétise l'ensemble de nos interprétations dans ce mémoire et donne des perspectives futures pour ce travail.