Chapitre 1

Classification automatique de textes

Table des matières

1.1- Introduction

La classification automatique de textes consiste à regrouper de manière automatisée des documents qui se ressemblent suivant certains critères à savoir les critères observables tels que le type du document, l'année, la discipline, l'édition, etc... ou le critère du contenu.

Elle connaît ces derniers temps un fort regain d'intérêt. Cela est dû essentiellement à la forte croissance des documents numériques disponibles et à la nécessité de les organiser de façon rapide.

La classification de textes est une tâche générique qui consiste à assigner une ou plusieurs catégories, parmi une liste prédéfinie, ou non à un document.

Actuellement, la classification de textes est un domaine de recherche très actif et l'automatisation de cette opération est devenue un enjeu pour la communauté scientifique, les travaux évoluent considérablement depuis une vingtaine d'années et plusieurs modèles ont vu le jour comme le filtrage (classification supervisée bi-classe), le routage (classification supervisée multi-classe) ou le classement ordonnée (classement des textes par ordre de pertinence pour chaque catégorie).

Avec ces modèles, des méthodologies de tests et des outils d'évaluation ont été mises en place. Les méthodes de représentation ainsi que les prétraitements correspondants sont maintenant bien connus. Les algorithmes de classification fonctionnent correctement mais déterminer les avantages des uns par rapport aux autres reste souvent délicat ou même améliorer les performances de la même méthode en intégrant d'autres paradigmes comme nous le faisons nous ici dans le présent mémoire reste toujours un domaine de recherche très prometteur.

Le domaine du traitement intelligent de données textuelles regroupe tout les outils et méthodes capables d'extraire des informations de textes écrits dans une langue naturelle.

Il existe essentiellement deux domaines de recherche qui traitent cette problématique avec chacune ses propres méthodes :

1- Les approches issues de l'analyse de données et de la statistique étudient et chercher surtout à proposer des dispositifs aux statisticiens et aux linguistes pour leur permettre d'analyser les grandes bases de données textuelles en fournissant des informations synthétiques sur les corpus. Les logiciels d'analyse de corpus qui fournissent des listes de fréquence de mots et les représentations graphiques issues de l'analyse factorielle des correspondances font partie de cette catégorie.

2- Les approches qui proposent des systèmes de type « boite noire », ces méthodes traitent les documents de façon automatique sans intervention humaine. Elles réalisent souvent des fonctions de bas niveau : analyse lexicale, analyse syntaxique de surface, recherche d'information par mots-clés. Les moteurs de recherche popularisés avec le réseau Internet présentent un exemple typique des applications qui s'appuient sur cette approche.

Dans ce chapitre préliminaire, nous allons entreprendre notre sujet en répondant à la question pourquoi automatiser la classification ? Puis par rappeler de l'ancienneté de cette discipline, ensuite définir la classification et les différents jeux de mots utilisés dans la discipline, ensuite éclaircir la notion de classe et la notion de catégorisation déterministe et floue. Les différents objectifs et intérêts et attendus de la discipline ainsi que les confluents avec d'autres disciplines comme le Text Mining ou Recherche d'Informations seront exposés par la suite. Nous finirons par développer la démarche classique d'un système de classification automatique de textes de la représentation des documents jusqu'aux évaluations des résultats

ainsi que les différentes contraintes qui s'opposent au processus qui sont soit liées à la nature des données traitées (textuelles) soit au corpus lui-même soit aux techniques de représentation ou même le type de classifieur.

1.2- Pourquoi automatiser la classification ?

On assiste aujourd'hui à un accroissement de la quantité d'information textuelle disponible et accessible d'une manière exponentielle. D'après les derniers chiffres, on parle de plus de 200 millions de serveurs hôtes sur Internet et plus de 3 milliards de pages, la taille des corpus tests utilisés est passée de quelques mégaoctets à plusieurs Gigaoctets.
Dans les années 1996-1997, Reuters a produit un peu plus de 800 000 nouvelles en anglais par année. Si l'on ajoute aux articles écrits par les journalistes de l'agence ceux provenant d'autres sources, on arrive à un total de 5.5 millions de textes anglais par année à catégoriser. À un moment, l'organisation employait 90 personnes dédiées à l'étiquetage de ces documents. Il serait à coup sûr très intéressant de pouvoir déterminer avec précision le coût de classification. De combien de temps a besoin un humain pour associer un texte à une catégorie ? En pratique, il s'agit d'une question difficile à répondre. Assurément, plusieurs variables influencent le phénomène et les lignes qui suivent porteront sur certaines d'entre elles.
Certainement une grande partie du temps consommé pour classer un document est employé dans sa lecture, puis éventuellement à sa relecture. On peut aussi imaginer que la longueur des textes à classer est assez déterminante du temps qui va être requis pour cette opération, et sans doute, d'une personne à une autre, la vitesse de lecture varie. Une fois cette étape achevée, il faut trancher à quelle(s) catégorie(s) ce texte appartient. Au temps de réflexion exigé s'ajoute, certainement, le temps de se référer à la description des classes et éventuellement de consulter d'autres textes préalablement associés à certaines classes, pour valider la décision. D'autres facteurs interviennent également comme, comme par exemple le nombre de classes qui peut faire la différence : plus il y a de classes différentes, autrement dit plus il y a d'étiquettes possibles pour un texte donné, plus il est difficile de faire un choix parmi celles-ci. Aussi, plus la sémantique des catégories est précise, fine, détaillée, plus il faut faire attention avant d'y associer un document. À cet égard, classer des documents appartenant soit à la catégorie *«informatique»* soit à la catégorie *«mathématiques»* est vraisemblablement plus aisée que celle de classer des documents appartenant à l'une ou l'autre des catégories *«Intelligence artificielle»* , *«Génie logiciel»* et *«Système d'information»* .

En conséquence, nous pouvons résumer les contraintes majeures qui s'opposent au traitement manuel de classification des documents textuels dans les trois points suivants :

- La réalisation manuelle de cette tâche par un expert est extrêmement coûteuse en terme de temps et personnel car il s'agit de lire attentivement chaque texte, au vu de la quantité phénoménale de textes aujourd'hui accessibles (par le biais du réseau Internet en particulier) (Moulinier, 1996), (Sebastiani, 2002)
- Les traitements manuels sont peu flexibles et leur généralisation à d'autres domaines est quasi impossible; c'est pourquoi on cherche à mettre au point des méthodes automatiques (Moulinier 1996), (Sebastiani 2002)
- Cette opération peut être perçue comme subjective puisque basée sur l'interprétation du document, deux experts peuvent classer différemment un même document, ou encore un même expert peut classer différemment un même document soumis à deux instants différents (Clech & Zighed, 2004), (Clech, 2004)

Ainsi l'intérêt de la recherche d'automatisation de la classification de textes n'est plus à démontrer, et c'est dans cette perspective que plusieurs travaux de recherche se concentrent ces dernières années.

1.3- Historique de la Catégorisation de textes

C'est une discipline assez ancienne, en 1627, Gabriel Naudé propose un classement selon cinq grands thèmes : théologie, jurisprudence, histoire, sciences et arts, belles lettres. Le désir de maîtriser l'Univers se fait sentir dans la multiplication des encyclopédies. L'encyclopédie de Diderot (parue entre 1751 et 1772) est organisée selon l'ordre alphabétique avec des renvois associatifs alors que celle de Panckoucke (parue de 1776 à 1780) suit une organisation méthodique selon un ordre arborescent (Fayet & Scribe, 1997).
Le système de classification par thème, apparu dès les débuts de l'écriture et institutionnalisé à Alexandrie conduisit à la création par Dewey, en 1876, d'un système de classification « universel ». Il s'agit d'une classification documentaire de type encyclopédique.
Toutefois l'idée d'effectuer la classification de textes par des machines remonte au début des années 60 et qui a connu des progrès considérables à partir des années 90 avec l'apparition d'algorithmes beaucoup plus performants qu'auparavant.
Jusqu'au début des années 80, pour construire un classifieur, il fallait consacrer d'importantes ressources humaines à cette tâche. Plusieurs experts éditaient des règles manuellement puis les affinaient au fur et à mesure des tests. L'avènement des de l'AA s'est donc traduit par un gain de temps conséquent. Il n'est plus nécessaire par exemple de reconfigurer tout le système en cas de changement d'arborescence.
Ces évolutions technologiques et algorithmes avancées font aujourd'hui de la catégorisation un outil fiable.
Au début des années 90, les travaux proviennent essentiellement de la communauté de Recherche d'Information (RI). En effet, les méthodes de numérisation, les algorithmes de classification et les méthodologies de test ont été adaptés à la CT en particulier au cours des conférences TREC (**T**ext **RE**trieval **C**onference. http://trec.nist.gov).
La communauté d'Apprentissage Automatique (AA) s'est intéressée elle aussi à ce problème il y a une dizaine d'années en le considérant comme domaine d'application à ces algorithmes de reconnaissance des formes. Actuellement, les méthodes de numérisation de texte restent largement inspirées de la RI alors que les classifieurs les plus performants sont issus de l'AA.
Une autre communauté composée essentiellement de statisticiens et de linguistes, traite également le problème de la CT en s'appuyant sur les méthodes d'analyse de données. Le but ici n'est pas de créer un système qui classe automatiquement des documents sans intervention humaine mais d'extraire des informations synthétiques du corpus. Les problématiques traitées ici sont par exemple l'étude des genres littéraires ou la détermination de l'auteur d'un texte.

1.4- Les systèmes de classification et vocabulaire utilisé

L'objectif de la CT est de classer de façon automatique les documents dans des catégories qui ont été définies soit préalablement par un expert, il s'agit alors de classification supervisée ou catégorisation, soit de façon automatique, il s'agit alors de classification non supervisée ou encore clustering.

Classification, catégorisation ou encore clustering ? C'est des termes qu'on peut rencontrer dans la littérature puisque la CT provient de plusieurs domaines scientifiques différents qui n'utilisent pas toujours le même vocabulaire pour la dénomination des différentes tâches.
Les deux termes « Classification », « Catégorisation » ont des histoires et des origines très différentes. La confusion entre ces deux termes persiste depuis le temps, dans le langage

courant voire philosophique. La première définition de la classification apparaît pour la première fois dans la cinquième édition du dictionnaire de l'académie française en 1798 « Distribution en classes et suivant un certain ordre ». Le mot « Catégorisation » n'existait pas dans le dictionnaire français, contrairement au mot « catégorie », quoiqu'il puisse être défini comme étant l'action de créer des catégories ou le résultat de cette action. Ce terme vient du grec « katêgoria : qualité attribuée à un objet ». Les catégories sont définies par Aristote comme étant « les espèces les plus générales de ce qui est signifié par un mot simple ». Il rassemble dans un même groupe des éléments proches et recense dix catégories différemment à certains pythagoriciens qui voulaient opposer toutes les espèces deux à deux : masculin et féminin, pair et impair, fini et infini, statique et dynamique, etc.…
Comme tenu de l'historique de ces deux termes et leur contexte d'utilisation actuelle, nous allons essayer de distinguer entre les différentes variantes de classification de textes et le vocabulaire utilisé dans la section suivante.

1.4.1- Catégorisation (Supervisé)

Ainsi, la *catégorisation* de textes correspond à la procédure d'affectation d'une ou de plusieurs catégories ou classes prédéfinies à un texte. Elle correspond à la *classification supervisée* pour l'apprentissage automatique et à la *discrimination* en statistiques alors que la recherche d'informations utilise des termes plus proches de l'application concernée : *filtrage* ou *routage*.
Cette problématique a par ailleurs dernièrement trouvé de nouvelles applications dans les domaines du traitement du langage tels que : l'affectation de sujets en recherche d'information, l'aide de l'utilisateur pour l'indexation de documents (Hayes & Weinstein, 1990), la veille technologique, le filtrage personnalisé des documents intéressant un internaute connaissant ses préférences de sujets (catégories) (Lang, 1995), le routage de textes (tels que le courrier) et l'amélioration de la recherche sur le web (Armstrong & all, 1995), et enfin l'organisation des sources textuelles de plus en plus nombreuses, en particulier des pages web.
Aujourd'hui, cette problématique utilise largement des méthodes issues de l'apprentissage automatique et beaucoup d'algorithmes d'apprentissage supervisé lui ont été appliqués (Naïve bayes, K-plus proches voisins, arbres de décision, machines à vecteurs support, réseaux de neurones, etc...)

1.4.2- Clustering (Non supervisé)

Toutefois quand l'ensemble des catégories n'est pas donné au départ, et qu'il s'agit de le créer en regroupant les textes en classes qui possèdent un certain degré de cohérence interne, on est dans un contexte de *classification non supervisée* pour l'apprentissage automatique.
La *classification non supervisée* consiste à trouver de manière automatique une organisation cohérente à un groupe de documents homogènes pour construire des regroupements cohérents (des classes ou clusters), elle correspond en statistiques au clustering, qui est également le terme utilisé en recherche d'informations.
Le clustering consiste donc, à diviser les objets (dans notre cas des textes) en groupes sans connaître à priori leurs classes d'appartenance.
Les techniques pour réaliser de tels regroupements constituent un domaine d'étude très riche, qui a donné lieu à de multiples propositions dont le recensement n'est pas l'objet de ce document.

- ***Dans ce qui suit notre travail va être concentré sur la catégorisation de textes (la classification supervisée).***

1.5- Définition de la Catégorisation de textes

Dans sa forme la plus simple, la catégorisation de documents consiste à assigner à un texte une ou plusieurs étiquettes permettant d'indexer le document dans un ensemble prédéfini de catégories, Originellement conçue pour assister le classement documentaire d'ouvrages ou d'articles dans des domaines techniques ou scientifiques.
La Catégorisation de Textes (C.T) est le processus qui consiste à assigner une ou plusieurs catégories parmi une liste prédéfinie à un document. L'objectif du processus est d'être capable d'effectuer automatiquement les classes d'un ensemble de nouveaux textes.
La catégorisation de documents consiste à apprendre, à partir d'exemples caractérisant des classes thématiques, un ensemble de descripteurs discriminants pour permettre de ranger un document donné dans la (ou les) classe(s) correspondant à son contenu (Brown & Chong, 1998).
Principalement, les algorithmes de catégorisation s'appuient sur des méthodes d'apprentissage qui, à partir d'un corpus d'apprentissage, permettent de catégoriser de nouveaux textes. Ce type de méthodes sont dites inductives car elles induisent de la connaissance à partir des données en entrée (les textes) et des sorties (leurs classes).
Les divers travaux dans le domaine cherchent à trouver un algorithme permettant d'assigner un texte à une classe avec le plus grand taux de réussite possible sans toutefois assigner un texte à trop de classes. Dans un tel contexte, une mesure de similarité textuelle permet d'identifier la ou les catégories les plus proches du document à classer. Si cette notion de similarité sémantique est un processus souvent intuitif pour l'homme, elle résulte d'un processus complexe et encore mal compris du cerveau.
Le problème de la catégorisation peut se résumer en une formalisation de la notion de similarité textuelle, soit en d'autres termes à trouver un modèle mathématique capable de représenter la fonction de décision d'appartenance des textes aux catégories.
Nous considérons un ensemble de classes $C = \{ci\}$ et un ensemble de documents $D = \{dj\}$.
Un système de classification associe automatiquement à chaque document un ensemble de classes (0,1 ou plusieurs). Le problème de la classification a été formalisé de plusieurs manières, nous vous proposons la formalisation de Sebastiani (Sebastiani, 1999) reprise par Yang (Yang, 1999)
Deux fonctions sont définies :

- Une **fonction de décision** qui associe à chaque document un ensemble de classes
- Une **fonction cible** qui nous renseigne sur l'appartenance exacte d'un document à un ensemble de classes.

La fonction de décision est une estimation de la fonction cible qu'on ignore. Plus cette estimation est correcte, plus le système de classification est performant.
La fonction de décision et la fonction cible attribuent à chaque couple $(dj, ci) \in D \times C$ une valeur booléenne pour indiquer si le document *dj* appartient ou non à la classe *ci* .

La fonction de décision sera définie de la manière suivante :
$\mathbf{D} : D \times C \rightarrow \{ vrai, faux \}$, $\mathbf{D}(d,c)$ = Vrai si *d* est associé à la classe *c* sinon $\mathbf{D}(d,c) = Faux$

La fonction cible sera définie de la manière suivante :
$\mathbf{C} : D \times C \rightarrow \{ vrai, faux \}$, $\mathbf{C}(d,c)$ = Vrai si *d* est associé à la classe *c* sinon $\mathbf{C}(d,c) = Faux$

Dans les systèmes de classification basés sur des méthodes d'apprentissage, la fonction de décision sera évaluée à l'aide d'un corpus d'entraînement. Cette fonction peut faire intervenir un grand nombre de valeurs numériques qu'un humain ne peut pas saisir. La détermination de cette fonction est appelée *phase d'apprentissage*, tandis que l'utilisation de cette fonction pour attribuer une catégorie à un document se fera pendant la *phase de test*.

1.6- La notion de classe pour les systèmes de classification

La notion de classe pour un système de classification a été habituellement synonyme de « thème ». Dans ce contexte, classer les documents revient à les organiser par différentes thématiques. (Par exemple : *Earn, Ship, Trade,* correspondent à des thèmes dans le corpus Reuters). Cependant, la problématique de classification a évolué en même temps que les besoins et elle s'intéresse aujourd'hui à différentes tâches pour lesquelles les catégories ne sont pas interprétables comme des thèmes : ainsi, par exemple, les tâches consistant à classer les documents par auteur, par genre, par style, par langue, ou encore selon que le document exprime un jugement positif ou négatif, etc.. Ainsi la classe va correspondre à un besoin d'information d'un utilisateur ou d'une société et n'est donc pas obligatoirement un thème unique. Nous considérerons dans la suite qu'une classe est simplement une étiquette à associer à des documents.
Dans la figure 1.1, un système de classification d'emails est représenté où les classes peuvent être de différentes natures (thèmes, messages provenant de certaines personnes, messages d'un certain type, etc…)

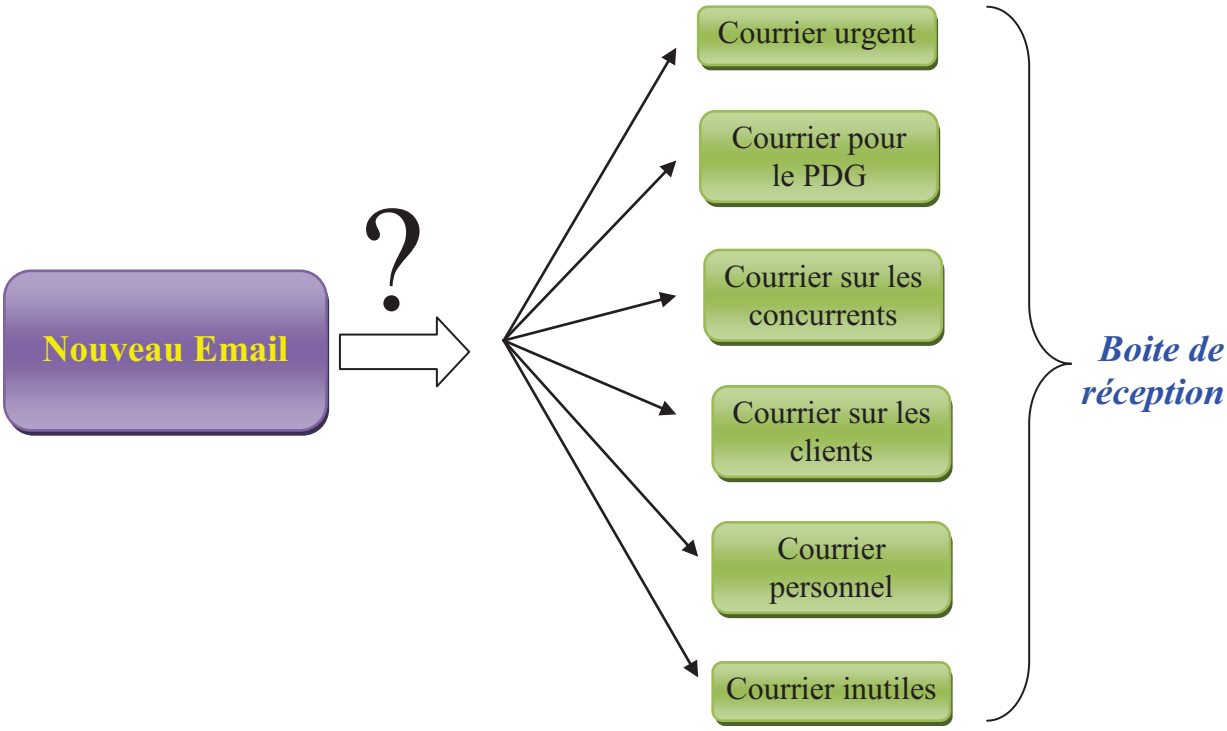

Figure 1.1 : Exemple de système de classification d'emails

Ce système organise des emails dans des boîtes aux lettres qui correspondent chacune à une classe qui sont de différentes natures (« mails urgents », « Mails du Directeur général »,etc…)

1.7- Les différents contextes de classification

Plusieurs contextes de classification se distinguent, ils influents directement sur les modèles utilisés. Ludovic DENOYER a bien résumé les différents contextes de classification dans (Denoyer, 2004) que nous avons reporté dans ce qui suit, les problématiques les plus récentes comme par exemple la classification dans une hiérarchie de classes (McCallum & all, 1998),(Koller & Sahami, 1997) ne sont pas abordés ici.

1.7.1- Classification bi-classe et multi-classes

1.7.1.1- La classification bi-classe

La classification bi-classe correspond au *filtrage*. C'est une problématique pour laquelle le système de classification répond à la question : *« Le texte appartient-il à la catégorie C ou non (i.e. ou à sa catégorie complémentaire ¬C ? »* (Par exemple, un document est il autorisé aux enfants ou non).

Cependant quand il s'agit d'effectuer une classification multi-classe qui permet de transmettre le document vers le ou les catégories(s) le(s) plus approprié(s), on parle alors de *routage. C*ette classification muti-classes, selon le cas, peut être disjointes ou non.

1.7.1.2- La classification multi-classes disjointes

La classification multi-classes disjointes est le contexte de classification en un nombre de classes supérieur à un et pour lequel un texte est attribué à une et une seule classe. Un système de classification multi-classes disjointes répond à la question *« A quelle classe* (au singulier) *appartient le document ? »*.

1.7.1.3- La classification multi-classes

Dans un système de classification multi-classes, on peut associer un texte à une ou plusieurs classes voire à aucune classe. Le système répond donc à la question : *« A quelles classes* (au pluriel) *appartient le document ? »*. C'est le cas le plus général de la classification. Il correspond par exemple à la problématique de classification du corpus Reuters étudié ici dans ce mémoire.

1.7.2- Catégorisation déterministe et floue

1.7.2.1- Catégorisation déterministe

Le but des classifications précédentes est d'avoir une réponse définitive pour chaque texte (oui ou non, le texte **T** appartient à la catégorie **C**) ; qu'on peut qualifier par classification déterministe. Plusieurs fonctions de classement sont utilisées, parmi lesquelles : les règles de décisions, les arbres de décision, SVM.

1.7.2.2- Catégorisation floue ou le ranking

Contrairement aux cas précédents, on peut également souhaiter dans certains cas d'avoir simplement une évaluation des classes les plus adéquates -dans l'ordre- pour y classer le texte. Ce qu'on peut appeler par classification floue ou ranking.
Ce type de classification va permettre à l'utilisateur d'être plus indulgent si le texte est "proche" du thème que si le texte n'a absolument rien à voir avec celui-ci dans le cas ou ce dernier est incorrectement attribué à la classe.
Le ranking est une problématique de classification dans laquelle le système, au lieu d'associer un texte à une classe catégoriquement, il ordonne les classes par ordre de pertinence pour un texte donné.
Les méthodes qui évaluent une distance d'un texte à une catégorie permettent facilement ce type de classement de même pour les approches qui estiment des probabilités d'appartenance d'un texte à une classe.

Ludovic DENOYER dans (Denoyer, 2004) donne quelques exemples d'applications dans lesquelles ce système de classification est sollicité :

- Le ranking de pages Web pour une thématique définie par un internaute.
- Le filtrage avec un rajustement de seuil de tolérance, le seuil étant ajusté par rapport aux scores de ranking.
- Proposer à un utilisateur un classement d'experts compétents pour évaluer un projet.

Dans ce cas spécifique, une fonction de score est définie de la manière suivante :

$$\textbf{SCORE} : \quad D \times C \longrightarrow [0,1]$$

Cette fonction nous renseigne sur le degré d'appartenance d'un texte à une classe donnée. Ainsi, plus *SCORE*(*d*,*c*) est proche de 1, plus le document *d* est proche à être attribué à la classe *c* et inversement, plus cette valeur est proche de 0, plus le document est loin d'être attribué à la classe. Le calcul de cette fonction de score nous permet alors d'organiser les classes dans l'ordre pour y classer le texte et donc de savoir par exemple quelle est la classe la plus probable à être sélectionnée par rapport aux autres.
Pratiquement, tous les algorithmes de classification calculent un score entre un texte et une classe. C'est le cas de toutes les approches probabilistes, particulièrement le classifieur Naïve Bayes. Toutefois, ces systèmes peuvent être aussi utilisés pour la classification déterministe. Dans ce cas, il est fondamental d'adopter une stratégie transformant la fonction de score en une fonction de décision. Pour cela, la stratégie habituelle consiste à utiliser un seuil L_C tel que :

$$\begin{cases} \text{si SCORE } (d, c) > L_c \text{ alors } \mathbb{D}(d,c) = \text{vrai} \\ \text{sinon } \mathbb{D}(d,c) = \text{faux} \end{cases}$$

1.8- Objectifs et intérêts

Les intérêts des méthodes de classification sont multiples, il peut s'agir d'améliorer les performances des moteurs de recherche documentaire ou aussi classer les documents en fonction de leurs références communes à d'autres documents pour faire apparaître les liens qui les unissent.

Nous pouvons citer six applications typiques qui sont :

- Le classement automatique de différents communiqués de presse, ou messages sur des forums en différentes matières (« Les actualités de la région », « la bourse », « culture », etc..), (Exemple : Une boite propose un système de tri d'informations dans des flots de dépêches d'agence de presse AFP ou Reuters etc.. ou pages web. Chaque matin les nouvelles importantes sont faxées à différentes entreprises).
- Indexation automatique sur des catégories d'index de bibliothèques : aide à la classification thématique des différentes rédactions dans une bibliothèque.
- La gestion de bases documentaires (mémoire d'entreprise). Ce système peut être utilisé pour présenter l'information à l'utilisateur selon des catégories thématiques, ce qui facilite la navigation.
- Sauvegarde automatique de fichiers dans des répertoires.
- Les filtres internet en général, et en particulier les filtres anti-spams.
- Le classement automatique des emails, et particulièrement la redirection automatique de courriers des clients et fournisseurs en fonction de leur contenu vers les personnes compétentes dans une entreprise (Service commercial, livraison, service après vente, approvisionnements, etc..) ou vers des répertoires prédéfinis dans un outil de

messagerie, ou encore le tri de courriers électroniques dans différentes boîtes aux lettres personnelles et possibilité d'envoi de réponses automatiques.

1.9- Classification de textes et Text Mining

Le Text Mining est une technique permettant d'automatiser le traitement de gros volumes de contenus texte pour en extraire les principales tendances et répertorier de manière statistique les différents sujets évoqués ainsi découvrir des connaissances et des relations à partir des documents disponibles.
L'outil de Text Mining va générer de l'information sur le contenu du document. Cette information n'était pas présente, ou implicite, dans le document sous sa forme initiale, elle va être rajoutée, et donc enrichir le document.

Les besoins en Text Mining peuvent être :

- Recherche d'information
- Correction orthographique/grammaticale
- Traduction automatique
- Résumé automatique
- Question/réponse (interfaces en langage naturel)
- La veille technologique

Et notamment

- **La Classification automatique des documents**

Toutes ces applications sont étroitement liées.

1.10- Classification de textes et Recherche d'informations

Dans la section suivante, nous allons rappeler les définitions de la recherche d'informations et la catégorisation de textes et essayer de positionner l'un par rapport à l'autre.

- La recherche d'informations (RI), aussi appelée recherche documentaire (RD), est la problématique la plus ancienne de ce domaine, elle consiste à trouver, dans une importante base de documents, les documents pertinents correspondant à des requêtes qui peuvent être de différentes natures (liste de mots clefs, langage naturel, langage spécifique comme le SQL par exemple etc.).
La RI correspond à la tâche classique d'interrogation par des requêtes, aujourd'hui démocratisée par le Web avec des moteurs tels que Google ou Altavista ou encore la recherche informatisée de documents dans de sources bibliothécaires. Beaucoup de modèles ont été développés et continuent aujourd'hui de l'être, ces modèles peuvent être ensemblistes, algébriques, statistiques (Miller & all, 1999).
La recherche d'informations est généralement effectuée en indexant préalablement tous les documents de la base selon les mots qu'ils contiennent ; la recherche consiste à trouver, le plus rapidement possible, les documents ayant des mots communs avec la requête de l'utilisateur.

- La catégorisation de textes, consiste à trouver dans un flux de documents, ceux qui sont relatifs à un sujet défini par avance. L'une des applications consiste à fournir à un utilisateur, en temps réel, toutes les informations importantes pour l'exercice de son métier. Dans ce cas, l'utilisateur n'exprime pas son intérêt par une requête, mais par un ensemble de documents

pertinents. Cet ensemble de documents pertinents définit ce que l'on appelle, un *thème* ou une *catégorie*.

La recherche d'information se différencie de la classification ou la catégorisation par le très grand nombre de réponses possibles, qui peut être infini. L'application classique serait la réponse d'un moteur de recherche ou d'intelligence artificielle à une demande. La distinction entre ces deux disciplines peut être simplifiée de la manière suivante : ***dans le premier cas, la base de documents est fixe et l'interrogation est variable, alors que, dans le deuxième cas, la source de documents est variable et l'interrogation est fixe.***
Dans la pratique, la catégorisation de textes bénéficie de deux avantages par rapport à la recherche d'information : la stabilité dans le temps de la classe sélectionnée et la quantité réduite de documents à traiter dans le temps. La stabilité de la classe laisse le temps de construire des modèles performants permettant de rechercher la façon dont l'information est codée dans un texte. Le fait de traiter les textes un à un, au lieu de s'attaquer à une base importante de textes, est moins pénalisante pour un système moins performant, et rend possible l'utilisation de modèles plus complexes.

SALTON recommandait, à la fin des années 60, le regroupement des documents des corpus pour permettre une recherche d'information plus rapide en ne calculant plus les distances entre la requête et chaque document mais seulement entre la requête et chaque classe : « Clearly in practice it is not possible to match each analysed document with each analysed search request because the time consumed by such operation would be excessive » (Salton, 1968).
Une autre étude menée par (Bellot, 2002) montre que la classification automatique permet d'améliorer l'efficacité des systèmes de recherche. Un système de recherche documentaire, comme on a vu précédemment, donne, en réponse à une requête, une liste de documents. La liste des documents trouvés est souvent si longue que les utilisateurs ne peuvent l'examiner entièrement et laissent de côté certains documents pertinents mal classés. L'étude a démontré qu'une classification automatique des seuls documents retrouvés permet d'améliorer la qualité de la recherche documentaire.

1.11- Démarche à suivre pour la catégorisation de textes

Pour réaliser l'opération de catégorisation automatique de textes comme nous l'avons défini, la démarche commune est la suivante : la première phase consiste donc à formaliser les textes afin qu'ils soient compréhensibles par la machine et utilisables par les algorithmes d'apprentissage. La catégorisation des documents est la deuxième phase, cette étape est bien entendu décisive car c'est elle qui va permettre ou non aux techniques d'apprentissage de produire une bonne généralisation à partir des couples (Document, Classe).
Pour améliorer la performance des modèles, une évaluation de la qualité des classifieurs et la comparaison des résultats fournis par les différents modèles est effectuée en fin de cycle.

La démarche d'une approche standard de classification automatique de textes peut être résumée de la manière suivante :

- Eliminer les caractères de séparation, les signes de ponctuations, les mots vides, etc..;
- Les termes restants sont tous des attributs
- Un document devient un vecteur <terme, fréquence>
- Entrainer le modèle de classification à partir des couples (Document, Classe).
- Évaluer les résultats du classifieur

La figure 1.2 illustre la démarche de catégorisation de textes avec ses trois étapes qui peuvent être schématisées comme suit :

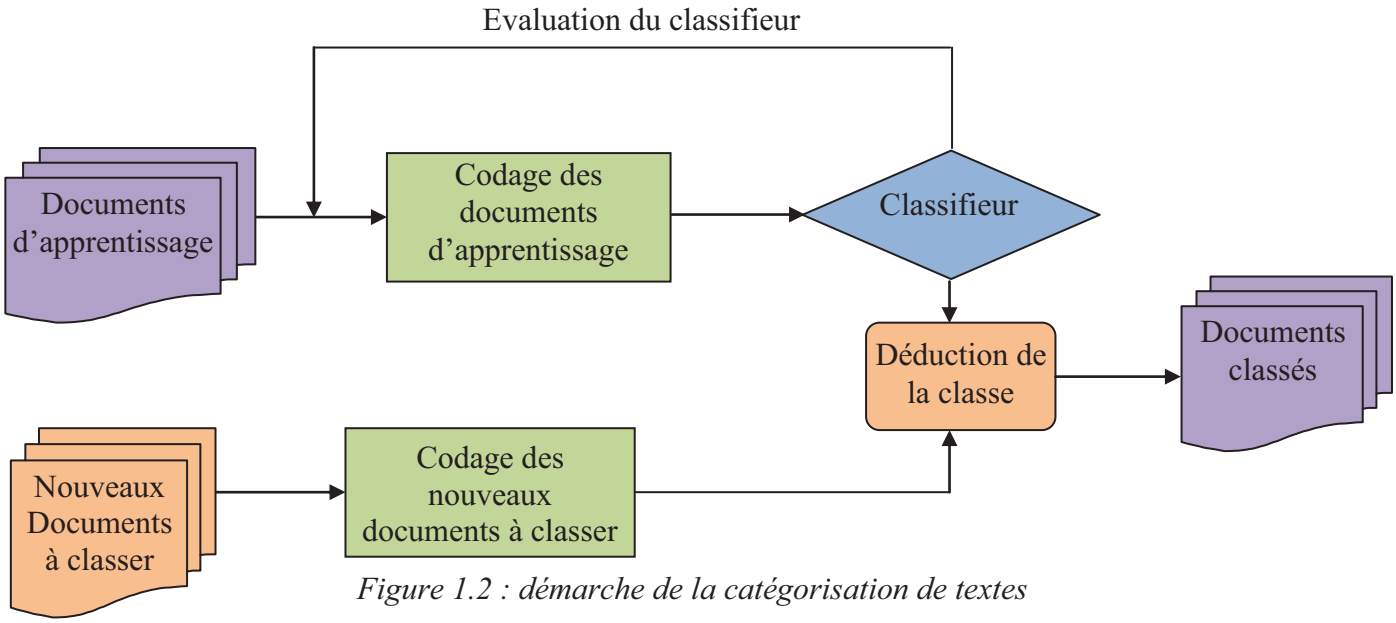

Figure 1.2 : démarche de la catégorisation de textes

Toutes ces étapes seront développées dans les chapitres 2,3 et 4.

1.12- Problèmes de la catégorisation de textes

Plusieurs difficultés peuvent s'opposer au processus de catégorisation de textes. Des problèmes connus dans la discipline liés à l'apprentissage automatique supervisé comme la subjectivité de la décision prise par les experts, le sur-apprentissage, etc.. mais aussi des problèmes particuliers liés à la nature des données traitées à savoir des données textuelles comme la polysémie, la redondance, Les variations morphologiques ou même L'homographie, etc..

Dans ce qui suit nous allons signaler les dix principales difficultés qui s'opposent à la catégorisation de textes :

1.12.1- Redondance(Synonymie)

La redondance et la synonymie permettent d'exprimer le même concept par des expressions différentes, plusieurs façons d'exprimer la même chose.
Cette difficulté est liée à la nature des documents traités exprimés en langage naturel contrairement aux données numériques. LE FEVRE illustre cette difficulté dans l'exemple du chat et l'oiseau : *mon chat mange un oiseau, mon gros matou croque un piaf et mon félin préféré dévore une petite bête à plumes* (Lefèvre, 2000).
La même idée est représentée de trois manières différentes, différents termes sont utilisés d'une expression à une autre mais en fin compte c'est bien le malheureux oiseau qui est dévoré par ce chat.
Lors d'une représentation vectorielle d'un document, ces termes sont représentés séparément, et les occurrences du concept sont dispersées. Il est alors important de rassembler ces termes en un groupe sémantique commun.
Pour y remédier, il est alors intéressant de concevoir une ontologie afin de cerner les sens des termes, naturellement, cela engendre des coûts supplémentaires pour sa réalisation et sa maintenance.

1.12.2- Polysémie (Ambiguïté)

A la différence des données numériques, les données textuelles sont sémantiquement riches, du fait qui sont conçues et raisonnées par la pensée humaine. Contrairement des langages informatiques, le langage naturel, autorise des violations des règles grammaticales engendrant plusieurs interprétations d'un même propos. Un mot possède, dans différents cas, plus d'un sens et plusieurs définitions lui sont associées.
Par conséquent, à cause de la polysémie, les mots seuls sont parfois de mauvais descripteurs.. Le mot *livre* peut désigner une unité monétaire, ou un bouquin. Le mot *avocat* peut désigner le fruit, le juriste, ou même au sens figuré, la personne qui défend une cause. Le mot *table* de cuisine ce n'est pas le même que dans *table* de multiplication. Le mot *pièce* peut correspondre à une pièce de monnaie par exemple, ou à une pièce dans une maison, de même pour *pavillon, bloc*, *glace*, etc..

1.12.3- L'homographie

Deux mots sont dits homographes si 'ils s'écrivent de la même façon sans forcément avoir la même prononciation. L'homographie est une sorte d'ambigüité supplémentaire. (Ex : avocat en tant que fruit et avocat en tant que juriste)
L'homographie et l'ambiguïté génère du bruit qui va causer une dégradation de précision (indicateur nécessaire pour mesurer la performance du classifieur). Il sera alors préférable d'ôter ces ambigüités.

1.12.4- La graphie

Un terme peut comporter des fautes d'orthographe ou de frappe comme il peut s'écrire de plusieurs manières ou s'écrire avec une majuscule. Ce qui va peser sur la qualité des résultats. Parce que si un terme est orthographié de deux manières dans le même document (Ghelizane, Relizane), la simple recherche de ce terme avec une seule forme graphique néglige la présence du même terme sous d'autres graphies, ce qui va influencer les résultats puisque les différentes graphies vont être traitées séparément. Néanmoins du point de vue pratique, le fait qu'un terme inconnu est proche d'un autre terme prouve qu'il a été mal orthographié.
Toujours dans ce contexte, (Loupy & El-Bèze, 2000) et (Loupy, 2000) affirment que la graphie peut donner une information relative au sens du terme employé. Prenons par exemple le cas pour le mot Histoire dont la majuscule indique qu'il s'agit de la discipline étudiant le passé et non d'un roman ou une blague. La prise en compte de toutes ces variations morphologiques pour la classification automatique de textes n'est pas étudiée dans ce mémoire.

1.12.5- Les variations morphologiques

Les conjugaisons, pluriels, influent négativement sur la qualité des résultats puisque les différentes variations morphologiques vont êtes considérées séparément et chacune va être prise comme un élément à part comme par exemple les trois termes : maître, maîtresse, maîtriser sont traités indépendamment quoique en réalité ça pivote sur la même idée. Pour y remédier soit on applique la lemmatisation ou le stemming, à notre texte soit carrément on opte pour une représentation en n-grammes qui peut nous éviter ces pré-traitements, tout ceux-ci va être étalé dans le chapitre 2.

1.12.6- Les mots composés

La non prise en charge des mots composés comme : comme Arc-en-ciel, peut-être, sauve-qui-peut, etc.. Dont le nombre est très important dans toutes les langues, et traiter le mot Arc-en-ciel par exemple en étant 3 termes séparés réduit considérablement la performance d'un système de classification néanmoins l'utilisation de la technique des n-grammes pour le codage des textes atténue considérablement ce problème des mots composés.

1.12.7- Présence-Absence de termes

La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoique on sait très bien qu'il ya plusieurs façons d'exprimer les mêmes choses, dés lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document.
Cette réflexion pointue nous amène à être attentifs quant à l'utilisation des techniques d'apprentissage se basant sur l'exclusion d'un mot particulier.

1.12.8- Complexité de l'algorithme d'apprentissage

Plus tard, Dans le chapitre 2 : représentation et codage des documents, nous verrons qu'un texte est représenté généralement sous forme de vecteur contenant les nombres d'apparitions des termes dans ce texte. Or, le nombre de textes qu'on va traiter est très important sans oublier le nombre de termes composant le même texte donc on peut bien imaginer la dimension du tableau (textes * termes) à traiter qui va compliquer considérablement la tâche de classification en diminuant la performance du système. De ce fait, une réduction de la taille du tableau, comme nous allons voir par la suite, est primordiale avant d'entamer l'apprentissage.

1.12.9- Sur-apprentissage

Le nombre de termes très important et très varié qui ne se répètent dans tous les textes va causer énormément de creux dans le tableau de grande dimension (textes*termes) qui peut provoquer du sur-apprentissage qui s'explique par le fait que le modèle n'arrive pas à bien classer les nouveaux textes, pourtant il l'a bien fait dans la phase d'apprentissage en classant correctement les textes de la base d'apprentissage.
Pour limiter le sur-apprentissage, on doit sélectionner des termes pour réduire la dimensionnalité. D'après les expériences antérieures, le nombre de termes doit être limité par rapport au nombre de textes de la base d'apprentissage.
Quelques auteurs recommandent d'utiliser au moins 50 à100 fois plus de textes que de termes.
En général le nombre de textes d'apprentissage est limité, c'est pour cela on cherche à agir sur le nombre des termes utilisés en les diminuant, pour éviter ce sur-apprentissage.
Sans bien sûr pénaliser le système en supprimant des termes pertinents (Sebastiani, 2002).

1.12.10- Subjectivité de la décision

Parmi les problèmes classiques usuels dans le domaine de l'apprentissage supervisé c'est la subjectivité de la décision prise par les experts qui décident de la classe à laquelle le texte va être attribué.
Certainement après la lecture du texte à classer, l'expert va trancher à quelle(s) catégorie(s) ce texte appartient en se basant sur le contenu sémantique et le contexte du texte et même en consultant d'autres textes préalablement associés à certaines classes, pour valider la décision prise qui ne peut être que subjective.

Les experts humains ne lisent pas de la même manière ! Ne réfléchissent pas de la même manière ! Donc ne classent pas de la même manière !
Ainsi un même document peut être classé différemment par deux experts, ou encore un même document peut être classé différemment par le même expert, soumis à deux instants différents (Clech & Zighed, 2004)
D'après les expériences : Lorsque deux experts humains doivent déterminer les classes d'une collection de textes, il y a souvent désaccord sur plus de 5 % des textes. Il est donc illusoire de rechercher une classification automatique parfaite.

1.13- Conclusion

La catégorisation de textes s'est avérée au cours des dernières années comme un domaine majeur de recherche pour les entreprises comme pour les particuliers.
Ce dynamisme est en partie dû à la demande importante des utilisateurs pour cette technologie. Elle devient de plus en plus indispensable dans de nombreuses situations où la quantité de documents textuels électroniques rend impossible tout traitement manuel.

La catégorisation de textes a essentiellement progressé ces dix dernières années grâce à l'introduction des techniques héritées de l'apprentissage automatique qui ont amélioré très significativement les taux de bonne classification.

Nous avons tenté dans ce chapitre1 de définir la classification ainsi que les notions nécessaires pour l'entame de la suite de ce mémoire.