Chapitre 1. L'apprentissage artificiel et la classification automatique des documents ayant un contenu *monolingue*

1.1 Introduction

L'apprentissage artificiel ou automatique (en anglais, « *Machine Learning ») selon* [Mitchell, 1997] est défini comme,

> *« Un sous-domaine de l'intelligence artificielle qui s'intéresse à conférer aux machines la capacité de s'améliorer à l'accomplissement d'une tâche, en interagissant avec leur environnement. »*

L'apprentissage artificiel se divise principalement en deux façons : l'apprentissage *supervisé* et l'apprentissage *non-supervisé*.

- La classification (ou catégorisation[8]) automatique des documents d'une manière supervisée est le processus d'assigner d'une façon autonome et automatique des documents à une ou plusieurs catégories prédéfinies (ex. Politique, Economie, Sport, etc.).
- La manière dite non-supervisée, quant à elle, ignore les catégories de sortie et c'est à l'algorithme d'apprentissage d'analyser les documents pour les concevoir.

C'est dans l'approche dite supervisée que s'inscrit la façon dont on aborde aujourd'hui le problème de la catégorisation automatique de documents, et notamment ceux écrits en caractères arabes.

L'apprentissage artificiel est un processus d'induction général qui permet la construction automatique des classifieurs en essayant de trouver automatiquement une liaison fonctionnelle, que l'on appelle également *modèle de prédiction*, entre les documents à catégoriser et l'ensemble des catégories. Autrement dit, [Jalam, 2003] indique dans la page 52 de sa thèse que pour catégoriser des documents « *on fournit à la machine des exemples sous la forme (Document, Classe*[9]*). Cette méthode de raisonnement est appelée inductive car on induit de la connaissance (le modèle) à partir des données (l'échantillon de Documents) et des sorties (leurs Classes). Grâce à ce modèle, on peut alors estimer les classes des nouveaux documents : le modèle est utilisé pour "prédire". Le modèle est bon s'il permet de bien prédire ».* Bien sûr, les

[8] Dans ce travail, les deux termes *classification* et *catégorisation* sont interchangeables et désignent le même concept.

[9] Dans le contexte de la classification automatique des documents, les deux termes « classe » et « catégorie » sont interchangeables et désignent la même chose.

nouveaux documents à catégoriser (i.e. ceux non utilisés lors de l'apprentissage et inconnus au modèle de prédiction) doivent appartenir aux mêmes domaines que ceux utilisés lors de l'apprentissage i.e. on ne saurait, par exemple, essayer de classer[10] un article scientifique à partir d'un modèle construit sur un ensemble d'apprentissage constitué d'articles de journaux de mode [Sebastiani, 2002].

Contrairement aux systèmes experts classiques, où des experts humains fournissent et maintiennent les règles de raisonnement, cette approche économise considérablement les ressources humaines. De plus, elle est simple à porter envers les différents domaines i.e. on pourra entrainer un algorithme d'apprentissage sur un corpus composé d'un ensemble de documents pour construire un modèle de classification automatique, et ensuite l'entrainer sur un autre corpus météorologique pour prédire la météo. La seule chose à changer est le corpus d'apprentissage. Cela rend cette approche très efficace et plus simple à utiliser et maintenir.

Dans ce paradigme, l'apprentissage s'effectue à partir d'un ensemble d'exemples déjà classés, dit *espace d'apprentissage,* où chacun d'eux est constitué d'un objet d'entrée (ex. un document) et d'une valeur de sortie désirée pour cet objet (ex. la catégorie : politique, sport, économie, etc.). En connaissant les sorties prévues, l'algorithme peut généraliser les exemples afin d'identifier les différents attributs des objets qui justifient une sortie particulière, pour devenir en mesure de traiter de nouveaux exemples. Il faut noter qu'on cherche toujours à construire le modèle de prédiction qui produit le moins d'erreurs.

1.2 Définition de la classification automatique des documents : quoi et comment ?

La catégorisation des documents est définie comme la tâche de trouver une liaison entre un ensemble de documents et un ensemble de catégories. Formellement, la catégorisation consiste à assigner une valeur de l'ensemble $\{0, 1\}$ à chaque entrée du tableau présentée ci-dessous (que l'on appelle « *la matrice décisionnelle* »).

[10] ou catégoriser

	c_1	c_2	…	…	…	…	c_n
d_1	v_{11}	v_{12}	…	…	…	…	v_{1n}
d_2	v_{21}	v_{22}	…	…	…	…	v_{2n}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
d_m	v_{n1}	v_{n2}	…	…	…	…	v_{mn}

Tableau 1.1 Matrice décisionnelle

On représente ce processus, plus formellement, comme la fonction

$$\Phi : D \times C \rightarrow \{1, 0\}$$

où,

$C = \{c_1 \dots c_n\}$ est l'ensemble *prédéfinie* des catégories,

et $D = \{d_1, \dots, d_n\}$ est l'ensemble des documents à catégoriser.

Une valeur de 1 pour v_{ij} signifie que le document d_j doit être placé dans la catégorie c_i, bien qu'une valeur de 0 signifie le contraire.

Selon [Sebastiani, 2002] page 3, il y a deux observations fondamentales à respecter pour obtenir une catégorisation généralisable :

- *« Les catégories ne sont que des labels symboliques. Ce qui veut dire que le sens ou la signification du nom de la catégorie n'a rien à voir avec la construction du classifieur. Autrement dit, on ne peut jamais utiliser le texte qui constitue le label (*par exemple, Economie ou Sports*) dans le processus de la catégorisation.*
- *L'affectation d'un document à une catégorie doit, en général, se baser sur le contenu du document plutôt que ses métadonnées (par exemple, date de publication, le type de document, le nom de l'auteur, etc.). »*

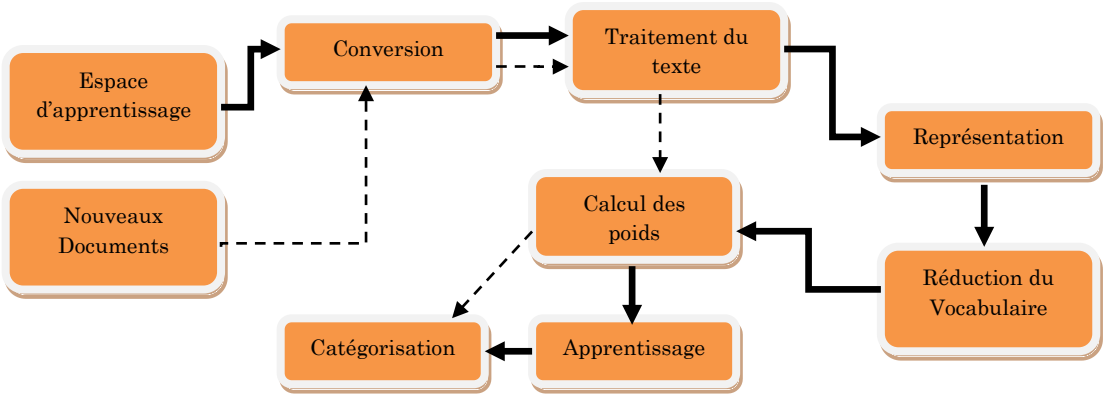

Figure 1.1 Le schéma général de la classification automatique des documents

La Figure 1.1 présente le schéma général de la classification automatique de documents. Le Chapitre 2. chapitre 2 présente une description détaillée de chacun de ses éléments.

1.2.1 Comment représenter un document ?

« Puisqu'il n'existe pas actuellement une méthode d'apprentissage capable d'exploiter un texte directement sans que ce dernier soit bien structuré, une transformation préliminaire est alors indispensable » [Sebastiani, 2002] page 10. Une transformation adoptée dans beaucoup des travaux du domaine se base sur le modèle du « *sac-de-mots* » (en anglais, « *bag-of-words* ») proposé par [Salton et McGill, 1983] pour représenter chaque document sous forme d'un vecteur composé de nombres réels représentant les poids des mots constituants le document. Tous ces vecteurs seront fusionnés ensemble en évitant les doublons. Le résultat sera reparti dans un tableau croisé ayant comme lignes les différents documents i.e. $d_i | \, i \in \{1, \dots, m\}$ présents dans l'espace d'apprentissage et comme colonnes l'intégralité des attributs i.e. $t_j | \, j \in \{1, \dots, n\}$ qui existent dans le corpus. Le contenu du tableau est le poids p_{ij} (le nombre réel indiqué ci-dessus) de l'attribut t_j qui existe dans le document d_i comme le démontre le Tableau 1.2.

	$\mathbf{t_1}$	$\mathbf{t_2}$	…	…	…	…	$\mathbf{t_n}$
$\mathbf{d_1}$	p_{11}	p_{12}	…	…	…	…	p_{1n}
$\mathbf{d_2}$	p_{21}	p_{22}	…	…	…	…	p_{2n}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
$\mathbf{d_m}$	p_{n1}	p_{n2}	…	…	…	…	p_{mn}

Tableau 1.2 Le tableau croisé issu de la restructuration de l'espace d'apprentissage

Si t_j existe dans le document d_i son poids aura une valeur supérieure à 0 et 0 autrement. Les différents techniques du calcul de p_{ij} seront abordées dans le paragraphe §2.3. « *Bien que cette approche ne tienne pas en compte l'ordre des mots dans les documents, elle produit des résultats satisfaisants* » [Forman, 2007]. Pour cela, nous l'adoptons dans cette thèse.

1.2.2 La réduction de la dimension de l'espace d'apprentissage

L'exploitation directe du corpus d'apprentissage constitue un vrai obstacle à l'algorithme d'apprentissage. Si on utilise directement tous les mots des documents d'entraînement[11] et qu'on crée une colonne dans le tableau croisé pour chacun, on se retrouve face à un espace vectoriel énorme. L'exploration et l'analyse d'un tel espace nécessitent beaucoup de ressources, en fonction de mémoire, de puissance et de temps de calcul. Ces ressources ne sont pas toujours disponibles et accessibles à la majorité des chercheurs de ce domaine. En plus, utiliser tous ces mots influencera négativement la fidélité de la classification car plusieurs mots sont vides de sens ex. les « mots vides » *ou « stop words »* comme, par exemple, les prépositions à, de, la, donc, etc. Il est donc inutile de les garder d'autant plus que la présence d'un mot dans un nombre élevé de documents affaiblira son pouvoir de discrimination. Dans certains cas, le problème peut devenir beaucoup plus grave puisqu'un tel espace d'apprentissage peut nuire à l'efficacité et au bon fonctionnement des algorithmes de classification, voire parfois les rendre dysfonctionnels. C'est en bonne partie pour ces raisons que certaines techniques ont été mises en place pour réduire la dimension du vocabulaire.

La réduction de la dimension est l'une des plus vieilles approches permettant d'apporter des éléments de réponse à ces problèmes. Son objectif est de sélectionner ou d'extraire un sous-ensemble optimal des mots les plus pertinents, selon un critère fixé auparavant, et

[11] Dans le contexte de la classification automatique par apprentissage artificiel, ce terme indique la phase d'apprentissage où on cherche à construire le modèle de prédiction.

l'élimination des mots non-pertinents et redondants. Cette méthode permet donc de réduire la dimension de l'espace d'apprentissage et de rendre l'ensemble des données plus représentable.

Les techniques de réduction de dimension se divisent en deux grandes familles :

- La *sélection d'attributs* qui conserve uniquement les mots utiles à la classification selon un critère fixé préalablement tandis que les autres sont rejetés.
- L'*extraction d'attributs* qui, par contre, crée de nouveaux attributs en faisant des regroupements ou des transformations des attributs déjà existants.

Nous constatons que la sélection d'attributs est meilleure quant à l'élimination d'attributs réellement inutiles ou même d'attributs erronés, ex. mots mal orthographiés, tandis que l'extraction d'attributs est plutôt axée sur la réduction du nombre d'attributs redondants. Nous abordons ces deux approches d'une manière plus détaillée respectivement dans les paragraphes §2.4, §2.5, et §2.6.

1.2.3 Le choix du classifieur

La catégorisation de documents comporte un grand choix des techniques d'apprentissage souvent utilisées dans la littérature. Parmi ces techniques figurent *les arbres de décisions* [Govindarajan, 2007], *les réseaux bayésiens naïfs* [Rich, 2007], *les machines à vecteurs de support* [Pilászy., 2005] et, récemment *la technique du dopage* proposée par [Schapire, 2002] que nous avons appliquée sur des documents écrits en caractères arabes dans le chapitre 5.

Le choix du classifieur dépend de l'objectif final à atteindre et de la taille du corpus. Selon [Jalam, 2003] « *La différence entre ces classifieurs est leur mode de construction : est-ce que ces classifieurs sont construits manuellement* (ex. les systèmes experts classiques) *ou bien automatiquement par induction à partir des données*? » En plus, est-ce que leur modèle de prédiction construit est symbolique (ex. les arbres de décision[12]) pouvant être interprété par des experts humains ou bien il s'agit d'une fonction numérique calculée à partir de données servant d'exemples? [Jalam, 2003]. En plus, la taille de l'espace d'apprentissage joue un rôle très important dans le choix de l'algorithme d'apprentissage. Comme nous le verrons dans le

[12] Voir §3.4

chapitre 5, les arbres de décision sont capables de traiter un corpus ayant une petite ou moyenne taille. Par contre, si le corpus est d'une grande taille on peut utiliser les machines à vecteurs de support. De plus, si l'on cherche à améliorer la fidélité de l'algorithme utilisé par le dopage on peut l'associer avec un des algorithmes de la famille de boosting.

1.2.4 L'évaluation des résultats

Puisqu'on adopte l'approche supervisée pour la classification automatique des documents écrits en caractères arabes, on sait d'avance la catégorie à laquelle appartient chacun des documents de l'espace d'apprentissage. Ainsi, nous utilisons une approche simple et largement appliquée dans la littérature pour évaluer les résultats d'un classifieur. Cette approche isole un sous-ensemble des documents de l'espace d'apprentissage avant l'entrainement et l'utilise uniquement pour évaluer la fidélité du classifieur où on fournit un document en entrée à l'algorithme d'apprentissage et on compare la sortie avec la catégorie attendue. Une fois qu'on a ce résultat on met à jour le **Tableau 5.1** page 112 (que l'on appelle le « tableau de contingence ») duquel on se servira pour calculer la performance du classifieur. Dans la majorité des travaux portant sur la classification automatique des documents, la performance d'un classifieur est mesurée via les trois mesures classiques : le *rappel*, désigné dorénavant par ρ, la *précision*, désignée dorénavant par π, et la mesure surnommée F1[13] qui cherche un équilibre entre ces deux mesures. Selon [Sebastiani, 2002] « *la précision est définie comme la probabilité conditionnelle qu'un document choisi aléatoirement soit correctement classé par l'algorithme d'apprentissage tandis que le rappel mesure la largeur de l'apprentissage et correspond à la fraction des documents dites pertinents, parmi ceux proposés par le classifieur* ». Le paragraphe §5.2 élabore en détail ces trois mesures.

1.3 Quelques applications de la classification automatique des documents

Selon [Sebastiani 2002], les travaux sur la catégorisation automatique des documents remontent jusqu'aux années 1960 avec les travaux séminaux de [Maron, 1961] qui

[13] Voir §5.2

portaient sur l'indexation automatique des documents. Dans ce qui suit, nous présentons les exemples[14] du contexte général des applications de la classification de documents.

1.3.1 L'indexation automatique

Comme nous le savons déjà, l'indexation automatique des documents sert à assigner un ou plusieurs mot(s)-clé(s) à un document pour décrire son contenu. Chaque mot-clé appartient à un ensemble prédéfini de mots que l'on appelle *un dictionnaire contrôlé*.

Si les données du dictionnaire contrôlé sont considérées comme des catégories, l'indexation de documents devient une instance de la catégorisation automatique de documents et peut être ainsi adressé par les techniques de l'apprentissage artificiel.

1.3.2 L'organisation des documents

L'indexation automatique en se basant sur un dictionnaire contrôlé est une instance de l'organisation de documents. En général, « *plusieurs applications portant sur la question de l'organisation des documents peuvent être réalisées par les techniques de la classification automatique* » [Sebastiani, 2002] page 6. Citons, à titre d'exemple, l'organisation du grand nombre d'annonces publiées dans des périodiques, comme par exemple *Paruvendu*[15] ou *TopAnnonces*[16], qui peuvent profiter des techniques de la classification automatique pour faciliter leur classement.

1.3.3 Le filtrage et le routage de documents

[Belkin 1992] définit *le filtrage (ou routage) de documents* comme « *la catégorisation d'un ensemble dynamique, plutôt que statique, de documents sous forme d'un flux de documents asynchrone envoyé par un émetteur d'information à un consommateur d'information* ». Par contre, [Sebastiani, 2002] observe qu'une confusion entre les deux termes *filtrage* et *routage* existe chez certains auteurs. Ce point de vue est

[14] L'idée générale de ces exemples est prise de [Sebastiani, 2002] et ensuite élaborée par nous.
[15] http://www.paruvendu.fr/ [Date de dernière visite : Mai 2010].
[16] http://www.topannonces.fr/ [Date de dernière visite : Mai 2010].

confirmé par [Liddy et al. 2004] qui utilisent le niveau de la sélection pour établir la différence :

- *« Si la sélection de l'information est faite chez l'émetteur, par exemple une agence de presse* (Reuters, Associated Press, etc.) *avant qu'elle soit envoyée chez le récepteur*, par exemple un journal de sport, *on parle alors de routage car, dans ce cas là, le système de routage doit bloquer la livraison de toutes les informations qui n'intéressent pas le récepteur »* (i.e. tous les informations n'appartenant pas à la catégorie Sport). On peut trouver des exemples de l'utilisation des techniques de classification automatique pour le routage de documents chez [Schütze et al. 1995, Bhagwat et al. 2007, Iyer et al. 2000] et d'autres.

- « *Si la sélection de l'information est faite au niveau du récepteur, on parle alors de filtrage. »* Un exemple de filtrage d'information est la détection de *spams* (les courriers indésirables) [Zhou et al. 2007, Gordillo et al. 2007]

On peut considérer le routage et le filtrage comme des cas spéciaux de catégorisation de documents dans des catégories non-chevauchantes i.e. la catégorisation des documents dans deux catégories : les pertinents et les non-pertinents. En plus, un système de routage peut aller un peu plus loin et opère une sous-catégorisation idéologique des documents jugés pertinents au profil du récepteur ; dans l'exemple ci-dessus les articles sont catégorisés selon le sport dont ils traitent et cela permettra à des journalistes spécialisés de travailler uniquement avec les documents auxquels ils/elles s'intéressent [Sebastiani, 1999]. [Schütze et al. 1995] présentent une comparaison plus détaillée entre des classifieurs utilisés pour le routage de documents.

1.3.4 La désambiguïsation sémantique automatique (DSA)

La désambiguïsation sémantique d'un mot ambigüe (ex. polysémique[17] ou homonymique[18]) dans un texte (en anglais, « *Word Sense Disambiguation »)* consiste à déterminer le sens correct de ce mot dans ce texte. Par exemple, prenons le mot « *but* » dans les deux phrases suivantes :

« *Le but de Maradona* » et « *Le but de notre politique* »

[17] Un mot ayant plusieurs sens (ex. *mémoire* humaine, *mémoire* d'ordinateur, le *mémoire* de thèse, etc.)
[18] Mot ayant la même forme, la même graphie mais des sens différents (ex. Je *porte* la *porte*).

C'est clair que le mot *but*, pourtant écrit et prononcé de la même façon, a un sens différent dans chacune des deux phrases. C'est dans des cas pareils qu'interviennent les techniques de DSA pour décider lequel des sens convient le plus. La DSA est très importante pour un nombre d'applications comme, par exemple, l'indexation de documents basée sur le sens du mot plutôt que sur le mot lui-même [Schütze and Pedersen, 1995], pour améliorer la traduction automatique [Vickrey et al. 2005, Carpuat et Wu 2007], ainsi que d'autres applications de gestion de documents basées sur le contenu. En fait, les techniques de DSA fonctionnent de la même façon qu'un classifieur une fois que l'on considère le mot dans un tel contexte étant un document et ses propres sens étant des catégories. [Gale et. al, 1993 et Schütze, 1998] donnent plus de détails sur ce point de vue.

1.3.5 La catégorisation des pages et des sites web

L'utilisation des techniques de la classification automatique de documents a dernièrement fait l'objet de beaucoup d'intérêt dans le domaine de la catégorisation des pages ou sites Web où l'on cherche à placer ces derniers dans une ou plusieurs catégories. Cette répartition sert à concevoir des catalogues hiérarchisés utilisés dans le domaine de la recherche d'information. Ce genre d'organisation a permis de faire des recherches que l'on appelle des « recherches verticales » puisqu'elles sont faites dans un domaine spécifique. Nous citons, entre autres, les sites du commerce électronique http ://www.shopwiki.com/. ou http ://www.nextag.com/, le site de la recherche scientifique http ://www.scirus.com/, et le site de la recherche dans le domaine médicale http ://www.medstory.com/[19].

En fait, les techniques de l'apprentissage artificiel ont été utilisées dans de nombreuses autres applications et domaines comme, par exemple, l'identification de la langue [Jain et al. 1994, Liu, et al. 2005, Cavnar and Trenkle, 1994], la reconnaissance des caractères manuscrits [Feng et. al, 2005] et [Jaeger et. al, 2005], la classification des documents multimédia [Denoyer 2003] bien que d'autres.

[19] [Date de dernière visite de ces 4 sites: Mai 2010]

1.4 Les difficultés rencontrées lors de la classification automatique des documents

Contrairement au traitement des données numériques, le traitement des données textuelles par les méthodes d'apprentissage est beaucoup plus difficile. Cela est dû, entre autres, aux pièges de la langue naturelle, la grande dimension de l'espace d'apprentissage, le sur-apprentissage, la subjectivité de la décision prise par le classifieur, et l'imprécision des fréquences des mots. Dans ce qui suit nous décrivons brièvement chacun de ces points[20].

1.4.1 Les pièges de la langue naturel

On précise rarement le sens des mots que l'on emploie d'autant plus qu'un mot peut avoir plusieurs sens : «*pollachôs legetaï*» disait Aristote, les choses « *se disent en plusieurs sens* ». Selon, [Lefèvre, 2000] la langue naturelle est équivoque d'où naissent les problèmes de :

- L'*implicite* puisque tout n'est pas exprimé dans le discours et par la suite il est impossible de le prendre en compte par des logiciels.

- La *redondance* puisqu'il existe plusieurs façons d'exprimer la même idée dû à des mots ou des expressions différents ayant le même sens, ou des sens voisins (la synonymie), des expressions équivalentes mais de structure ou de termes différents (la paraphrase), et à l'incompatibilité entre le sens propre d'un mot (dénotation) et son sens dans un contexte particulier (connotation).

- L'*ambigüité* car ce qui est exprimé possède souvent plusieurs interprétations comme dans les cas de l'homonymie, la polysémie, et l'homotaxie (une même syntaxe recouvrant des réalités différentes).

1.4.2 La malédiction de la grande dimension de l'espace d'apprentissage

Pareil à beaucoup d'autres travaux portant sur la classification automatique des documents, l'approche adoptée dans cette thèse pour la représentation d'un document est celle du modèle vectoriel élaboré dans §2.3. Puisque l'espace d'apprentissage est composé dans la plupart du temps d'un très grand nombre de documents, cette

[20] Les idées des points 1, 2, 4, 5 ont été proposées par [Jalam, 2003, pages 13-16].

approche génère une énorme matrice (tel que présenté dans le Tableau 1.2 page 28). A présent, il n'existe aucun algorithme d'apprentissage capable d'exploiter cette gigantesque matrice telle qu'elle est sans que cette dernière affecte négativement sa performance et sa fidélité et même parfois le rendre inopérable. Pour cela, il est indispensable de réduire la taille de cette matrice, avant de pouvoir l'utiliser, en appliquant une des mesures présentées dans le paragraphe §2.5. Néanmoins, la réduction de dimension ne doit pas être sur-appliquée afin d'éviter de supprimer des attributs pertinents [Sebastiani, 2002]. Comme il n'existe aucune règle générale qui décrit combien il faut réduire et quelle est la dimension qui donne les meilleurs performances et résultats, seules les expérimentations par tâtonnement peuvent nous l'indiquer.

1.4.3 Le sur-apprentissage (*Overfitting)*

Un des problèmes majeurs qu'on peut rencontrer lors de l'apprentissage est le phénomène de « sur-apprentissage », où l'algorithme d'apprentissage s'entraîne et se sur-adapte aux exemples d'entraînement d'une façon qui affaiblit sa généralisation et son pouvoir de catégoriser correctement de nouveaux exemples.

La grande dimensionnalité de l'espace d'apprentissage présente un des cas menant au sur-apprentissage que l'on peut résoudre en appliquant une des techniques de réduction de la taille de l'espace d'apprentissage. Dans les autres cas, il faut surveiller la performance du classifieur lors de son entraînement. Par exemple, dans le cas du dopage (c.f. §3.7) on peut échapper le piège de sur-apprentissage en surveillant la performance du classifieur à chaque itération et arrêter directement le processus de l'entrainement une fois que sa fidélité commence à se détériorer.

1.4.4 La subjectivité de la décision

Un des défis difficiles à relever lors de la classification automatique est celui du choix de la catégorie à attribuer à un tel document du fait que cette décision dépend de son contenu sémantique d'autant plus que l'interprétation de ce dernier varie d'un expert à un autre. Selon [Sebastiani, 2002], les experts sont souvent en désaccord étant donné que la sémantique d'un document est une notion *subjective.* Il cite l'exemple d'un article de journal parlant de la visite de l'ancien Président des Etats Unis Bill Clinton aux funérailles de Dizzy Gillespies (un fameux trompettiste) et indique que ce

document peur être classé sous la catégorie *politique*, ou la catégorie *Jazz*, ou bien les deux, voire même aucune d'eux. En fait, tout dépend de la subjectivité des experts.

1.4.5 L'imprécision des fréquences

Puisqu'on traite des milliers de documents, on se retrouve face à un très grand nombre d'attributs appartenant à des documents. Ces attributs se reproduisent rarement dans chacun de ces documents. Par conséquence, les cellules du tableau croisé résultant du paragraphe §2.3 contiennent souvent des petites valeurs, voire dans la plupart du temps une valeur de 0. Comme on l'a déjà indiqué dans le paragraphe §2.3, ces valeurs correspondent aux poids de chaque mot dans le document qui le contient. Le calcul du poids d'un mot dépend souvent de sa fréquence. Selon [Jalam, 2003] page 15 les poids *« suivent approximativement des lois de Poisson; le coefficient de variation (CV = écart-type/moyenne) donne une indication sur la précision relative de l'estimation de la fréquence dans une cellule du tableau croisé; pour une loi de Poisson de moyenne m, la variance est aussi égale à m; le coefficient de variation est donc : CV=* $\frac{\sqrt{m}}{m} = \frac{1}{m}$*; si m est petit, le CV est grand et la fréquence est donc imprécise »*.

1.5 Conclusion

Ce chapitre introductif a élaboré et défini l'apprentissage artificiel et la classification automatique des documents ainsi que son objectif qui sont devenus des aspects majeurs et distingués dans le domaine de la recherche d'information, et cela est dû à plusieurs raisons :

- Le grand nombre des domaines dans lesquels ces techniques peuvent être appliquées,
- Leur indispensabilité pour des applications de classification où une réponse est requise dans un temps très minimal,
- Leur utilité pour les experts humains travaillants sur des applications qui ne peuvent pas se disposer du facteur humain mais en même temps profitant d'un outil de suggestion de décisions plausibles.
- Leur efficacité qui a prouvée d'être comparable à celle des experts humains.

Dans ce qui suit nous décrivons les différents éléments du processus de la classification automatique : la représentation d'un document, la réduction de la taille du vocabulaire, le calcul du poids, le choix du classifieur et les critères d'évaluation.