Chapitre 2

Codage des textes : Etat de l'art

Table des matières

2.1- Introduction

L'explosion de la quantité d'informations textuelles provoquée par l'évolution à grande échelle des outils de communication essentiellement Internet qui est sorti de l'aspect réservé à un milieu restreint à un aspect de vulgarisation au grand public, a rapidement, fait sentir le besoin de recherche de mécanismes et outils de traitement automatique des quantités d'informations diffusées sur le Web.
Ainsi, avec les bases de données multimédia, les dépêches d'agences de presse, les publications scientifiques, les bibliothèques électroniques, etc... Qui sont consultés habituellement sur le réseau, on dispose de plus en plus de grandes masses de documents non ou faiblement structurées, en particulier les documents textuels qui sont considérés comme étant des documents non structurés, surnommés « documents plats » par quelques auteurs, c'est-à-dire comme une séquence ou un ensemble de mots sans informations complémentaires sur le document.
Différents formats HTML, SGML, XML, DOC, PDF, … peuvent être des moyens pour stocker et représenter ces documents.
Le manque de structure au sein de ces collections volumineuses rend difficile l'accès à l'information qu'elles contiennent, d'où la nécessité aujourd'hui, de chercher comment structurer automatiquement ces corpus pour les rendre utilisables d'une façon rapide et optimale pour y faciliter leurs traitements automatiques et notamment la classification.
Pour pouvoir y appliquer les différentes techniques et algorithmes d'apprentissage, une transformation de ces documents non ou peu structurés est indispensable.
La transformation ou le codage de ces documents est une préparation à « l'informatisation » de ces derniers, chaque type de documents comme les images, les vidéos et notamment les textes dispose de ses propres techniques de codage.
Plusieurs approches de représentation des documents textuels ont été proposées dans ce contexte, la plupart étant des méthodes vectorielles.
Les principales méthodes de représentation de textes n'utilisent pas d'information grammaticale ni d'analyse syntaxique des mots : seule la présence ou l'absence de certains mots est porteuse d'informations.

Un état de l'art des différentes approches de représentation et codage de textes est développé dans ce chapitre.

2.2- Le texte

La numérisation est une opération qui s'est élargie pour atteindre toutes formes de documents et notamment les textes, et ce dans le but de leur exploitation sur les réseaux. Cet élargissement a entraîné derrière lui beaucoup de travaux qui ont un rapport surtout avec le formatage et la normalisation des textes qui ont été développés pour être à la fois rapide et efficace suite au développement de l'Internet.
Depuis les débuts de la numérisation des données textuelles, le texte a été considéré, et c'est encore vrai aujourd'hui dans la plupart des cas, comme tout simplement une séquence de caractères. Ces caractères peuvent être représentés dans différents espaces de codage, le plus courant étant le codage ASCII admettant 256 caractères différents, mais en dépit ce codage ne prenait pas en charge les langues comme l'arabe ou le chinois. Afin de pouvoir représenter ces langues, différentes normes de codages sont crées et plus largement utilisées aujourd'hui comme la norme UNICODE qui permet la représentation de 65536 caractères.

Pour la plupart des langues (occidentales et orientales font partie), l'espace de codage au niveau caractère n'est pas un espace très informatif car un caractère seul ne présente pas une information sémantique riche. Un texte est plutôt considéré comme une séquence de mots (un

mot lui même étant une séquence de caractères) et représenté dans un espace de mots dont la dimension est plus grande que celle du caractère (Le nombre de caractères possibles est limité mais en revanche le nombre de mots qu'on peut avoir est énorme), mais dont chaque dimension est beaucoup plus informative.
Ainsi la représentation informatique de ces textes nécessite un traitement spécifique.
Mais Très vite, les méthodes de se sont heurtées au fait qu'un texte n'est pas un sac dans lequel seraient mélangées en vrac ses propres éléments. Le moins qu'on puisse dire sur un texte qu'il est une chaîne linéaire, donc un espace ordonné. (« *Voile du bateau* » et « *Bateau à voile* » ont des sens complètement différents).
Evoquer la composition d'un texte fait appel à deux définitions de la composition : il s'agit à la fois de déterminer les unités qui vont constituer le texte, tels les atomes qui composent les molécules, et de constituer un texte c'est-à-dire de distribuer, d'organiser ces unités afin d'atteindre certaines idées, comme une molécule qui possède certaines propriétés en raison de sa structure.
Plusieurs approches de représentation dans cet espace sont proposées dans la littérature. Nous détaillons par la suite ces différentes représentations.

2.3- Prétraitements

Nous allons aborder ultérieurement les différentes méthodes de représentation des documents. Ces représentations sont toutes effectuées à base de mots qui sont eux-mêmes une séquence de caractères. Il est donc nécessaire d'effectuer, au préalable du codage d'un document dans un espace de mots, une transformation permettant le passage de l'espace du caractère à un espace de mots.
Le prétraitement des textes est une phase capitale du processus de classification, puisque la connaissance imprécise de la population peut faire échouer l'opération.
Après la première opération que doit effectuer un système de classification à savoir la reconnaissance des termes utilisés, nous devons expurger le plus possible les informations inutiles des documents afin que les connaissances gardées soient aussi pertinentes qu'il se peut. En effet dans les documents textuels de nombreux mots apportent peu (voir aucune) d'informations sur le document concerné. Les algorithmes dits de *"*Stop Words*"* s'occupent de les éliminer. Un autre traitement nommé *"*Stemming*"* permet également de simplifier les textes tout en augmentant leurs caractères informatifs comme d'autres méthodes qui proposent de supprimer des mots de faible importance.
Toutes ces transformations et méthodes font partie de ce qu'on appelle le prétraitement. Plusieurs d'entre elles sont spécifiques à la langue des documents (on ne fait pas le même type de prétraitement pour des documents écrits en anglais qu'en français ou encore en arabe).

Le prétraitement est généralement effectué en six étapes séquentielles :

1. La segmentation
2. Suppression des mots fréquents
3. Suppression des mots rares
4. Le traitement morphologique
5. Le traitement syntaxique
6. Le traitement sémantique

2.3.1- La segmentation

La première opération que doit effectuer un système de classification est la reconnaissance des termes utilisés. La segmentation consiste à découper la séquence des caractères afin de regrouper les caractères formant un même mot.

Habituellement, cette étape permet d'isoler les ponctuations (reconnaissance des fins de phrase ou de paragraphe), ensuite découper les séquences de caractères en fonction de la présence ou l'absence de caractères de séparation (de type « espace », « tabulation » ou « retour à la ligne »), puis regrouper les chiffres pour former des nombres (reconnaissance éventuelle des dates), de reconnaitre les mots composés.
Eventuellement, nous pouvons unifier les écritures en lettre majuscules ou en lettres minuscules avant ou après les opérations déjà indiquées.
C'est un traitement de surface assez simple dans le principe, mais particulièrement difficile à réaliser de manière exacte sur les documents ayant beaucoup de bruits et des représentations assez variées.
Notons que pour des corpus multilingues, une technique de segmentation moins intuitive a été proposée : la segmentation en n-grammes.

2.3.2- Suppression des mots fréquents ou élimination des "Mots Outils"

Les mots qui apparaissent le plus souvent dans un corpus sont généralement les mots grammaticaux, mots vides (empty words) ou mots outils (stop words) : les articles, les prépositions, les mots de liaisons, les déterminants, les adverbes, les adjectifs indéfinis, les conjonctions, les pronoms et les verbes auxiliaires etc.., qui constituent une grande part des mots d'un texte, mais malheureusement sont faiblement informatifs, sur le sens d'un texte puisqu'ils sont présents sur l'ensemble des textes.
A titre d'exemple on peut citer en dans la langue Française, le cas des articles « le », « la », « les » ou de certains mots de liaison « ainsi », « toutefois » etc..
Ou en Anglais : Les prépositions (about, after,through.), les déterminants (the, no, one.), les conjonctions (though, and, or.), les adverbes (above,almost, yet.), les pronoms (who, another, few.) et certains verbes (are, can, have, may,will.).
Et en Arabe : … حروف الجر، حروف العطف، أسماء الإشارة، أخوات كان، أخوات إن، الخ

Ces termes très fréquents peuvent être écartés du corpus pour en réduire la dimension. Cette possibilité de réduire la taille des entrées de l'index en éliminant les mots vides s'explique par le fait que ces termes sont présents dans la quasitotalité des documents et ont donc un pouvoir discriminant faible en comparaison avec d'autres termes.
D'après la loi de Zipf (Voir Section 2.6.1.4). Leur élimination lors d'un pré-traitement du document permet par la suite de gagner beaucoup de temps lors de la modélisation et l'analyse du document.
Ces mots doivent être supprimés de la représentation des textes pour deux raisons :

- D'un point de vue linguistique, ces mots ne comportent que très peu d'informations. La présence ou l'absence de ces mots n'aident pas à deviner le sens d'un texte. Pour cette raison, ils sont communément appelés « mots vides ».
- d'un point de vue statistique, ces mots se retrouvent sur l'ensemble des textes sans aucune discrimination et ne sont d'aucune aide pour la classification.

Une répartition des mots outils par rapport les mots utiles dans un corpus est représentée dans la figure 2.1.

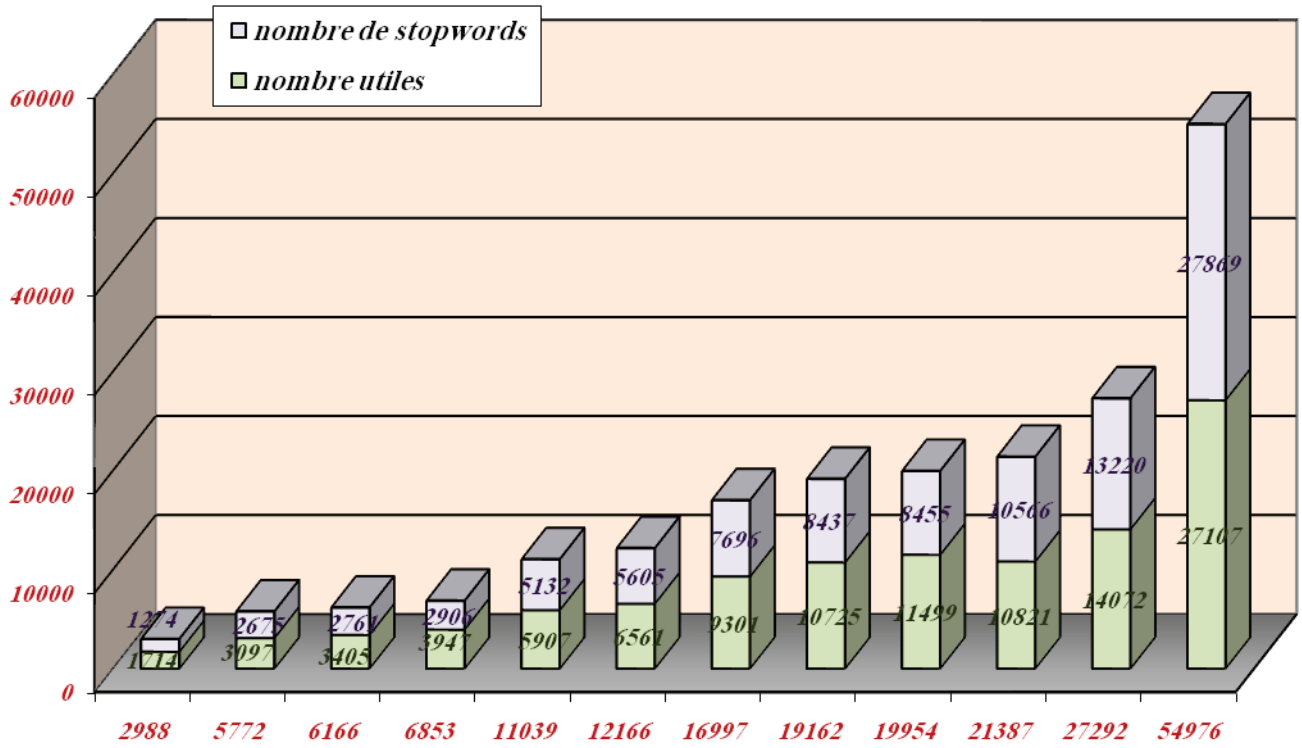

Figure 2.1 : Répartition des mots utiles et des mots vides dans un corpus

L'élimination systématique du corpus des mots vides peut se faire par l'intermédiaire d'une liste prédéfinie de mots pour chacune des langues étudiées.
Cependant, l'établissement d'une telle liste peut poser des problèmes. D'une part, il n'est pas facile de déterminer le nombre de mots exacts qu'il faut inclure dans cette liste. D'autre part, cette liste est intimement liée à la langue utilisée et n'est donc pas transposable directement à une autre langue.
Par exemple Sahami.M dans dans sa thése de PHD (Sahami, 1999) définit une liste de 570 mots courant en anglais, plus une liste de 100 mots très fréquents sur le web.
Comme on peut les écarter en fixant un seuil maximal de fréquence, pour ne pas sélectionner les mots présents dans une grande partie du corpus.
Une autre manière d'éliminer les mots vides d'un texte passe par l'utilisation d'un étiqueteur syntaxique (Part of Speech Tagger) – Les mots sont écartés en fonction de leur étiquette syntaxique sans avoir besoin de liste prédéfinie.

Enfin, un dernier point concernant les opérateurs de négation (ex : pas, ne, non) qui peuvent être supprimés sans gravité. Dans un contexte de classification de textes, une notion affectée par un opérateur de négation reste inchangée contrairement à une négation dans un contexte de recherche d'information qui peut être déterminante pour les résultats attendus. Dans le cadre d'une recherche documentaire, le but à atteindre pour l'utilisateur de rechercher l'information en lien avec la requête. En revanche, dans le cade d'une classification de textes en plusieurs catégories, les opérateurs de négation ne vont guère influencer les résultats puisque l'on cherche à distinguer les thèmes les uns des autres. Par exemple les deux phrases suivantes : il est malade et il n'est pas malade traitent toutes les deux le même sujet de santé, et le terme malade, avec ou sans négation, est un terme décrivant cette notion de santé. En évidence, elles ont un sens opposé mais sont toutes liées au sujet de santé.

2.3.3- Suppression des mots rares

En général, les auteurs cherchent également à supprimer les mots rares, qui n'apparaissent qu'une ou deux fois sur un corpus, afin de réduire de façon appréciable la dimension des vecteurs utilisés pour représenter les textes, puisque, d'après la loi de Zipf (Voir Section 2.6.1.3), ces mots rares sont très nombreux.
D'un point de vue linguistique, la suppression de ces mots n'est pas nécessairement justifiée : certains mots peuvent être très rares, mais très informatifs. Néanmoins, ces mots ne peuvent pas être utilisés par des méthodes à bases d'apprentissage du fait de leur très faible fréquence ; il n'est pas possible de construire de statistiques fiables à partir d'une ou deux occurrences ;
Une des méthodes communément retenues pour supprimer ces mots consiste à ne considérer que les mots dont la fréquence totale est supérieure à un seuil fixé préalablement.

Notons enfin, que les mots ne contenant qu'une seule lettre sont généralement écartés pour les mêmes raisons précédentes, comme par exemple le mot « D » dans la « Vitamine D » ou le mot « C » dans le « langage C ».

2.3.4- Le traitement morphologique

Consiste à effectuer un traitement au niveau de chacun des mots en fonction de leurs variations morphologiques : flexion, dérivation, composition afin de rassembler les mots de sens identiques. Donc, le but est de regrouper par exemple les termes «manger» et «mangent» ou les termes « cheval » et « chevaux » car ils ont la même signification. L'intérêt de cette opération est la réduction des dimensionnalités de l'espace de codage des textes afin d'améliorer davantage la performance du système de classification en matière d'espace mémoire et vitesse de traitement.

Plusieurs traitements morphologiques existent :

- ***Le stemming*** ou ***la désuffixation*** regroupe sous un même terme (stem) les mots qui ont la même racine. L'extraction des stems se fait par la technique de racinisation (ou stemming) qui utilise à la place des dictionnaires, des algorithmes simples basées sur des règles de remplacement de chaînes de caractères pour supprimer les suffixes les plus utilisés.
Le stemming est un traitement linguistique moins approfondie que la lemmatisation, ayant deux avantages : Plu rapide que la lemmatisation (algorithmes simples ne faisant pas référence aux dictionnaires et règles de dérivation) et la possibilité de traiter les mots inconnus sans traitement spécifique. (Clech, 2004)
Néanmoins, sa précision et sa qualité sont naturellement inférieures, du fait qu'elle ne gère que les règles principales et ne peut pas prendre en compte les nombreuses exceptions des règles de dérivations. Par exemple, en français l'une des règles préconise de supprimer le « e » final de chaque mot, le mot *« fraise »* est alors transformé en « *frais »* ce qui suppose une relation entre les deux mots qui n'existe pas. Qui fait de cette opération dépendante de la langue, nécessitant une adaptation pour chaque langue utilisée.
Plusieurs stemmers ont été développés pour déterminer les racines lexicales, l'algorithme le plus couramment utilisé pour la langue anglaise est celle de PORTER (Porter, 1980).

- ***La lemmatisation*** conserve, non pas les mots eux-mêmes, mais leur racine ou lemme. Ce principe permet de prendre en compte les variations flexionnelles (singulier/pluriel, conjugaisons,…) ou dérivationnelles (substantifs, verbes, adjectifs,…) en regroupant sous le même terme tous les mots de la même famille et donc d'améliorer la classification.
La lemmatisation est donc une tâche plus compliquée à mettre en œuvre que la recherche de racines, puisqu'elle elle s'appuie sur des outils de TALN, ce qui nécessite beaucoup de

ressources linguistiques (dictionnaires, règles de dérivation, etc..). De plus les résultats contiennent encore des erreurs à cause des problèmes de polysémie (ambiguïté) et d'incomplétude des dictionnaires.
Un algorithme efficace, nommé TreeTagger (Schmid, 1994) a été développé pour les langues anglaise, française, allemande et italienne. Cet algorithme utilise des arbres de décision pour effectuer l'analyse grammaticale, puis des fichiers de paramètres spécifiques à chaque langue.

Toutes les études montrent que les performances des systèmes de classification, après lemmatisation, sont plus nettement supérieures à celles avant lemmatisation.

2.3.5- Le traitement syntaxique

La syntaxe traite les combinaisons et l'ordre des mots dans la phrase.
Le traitement ***syntaxique*** identifie et regroupe un ensemble de mots dont la sémantique dépend de leur association. Par exemple, les mots « casque bleu » ne signifient habituellement pas qu'on a affaire à un casque qui est bleu, mais plutôt à une organisation militaire dépendante de l'ONU. L'analyseur syntaxique a pour but d'identifier ce type de cas. La phase d'analyse syntaxique consiste aussi à éliminer des ambiguïtés comme par exemple les problèmes d'homographie.

2.3.6- Le traitement sémantique

Le traitement ***sémantique*** consiste à extraire la signification des expressions et traiter la polysémie à savoir les différents sens possibles d'un même mot. Par exemple, cette phase permet de différencier le mot « *base* » qui peut correspondre à une *base* militaire ou à une *base* de données. C'est une opération laborieuse, qui fait appelle aux ontologies, et qui n'est pas aujourd'hui bien maîtrisée et dont l'intérêt en terme de meilleures performances, dans les systèmes de classification, n'est pas toujours démontré.

- Notons, en fin de cette section, que les différents traitements appliqués sur un texte avant sa représentation informatique ne sont pas toujours nécessaire pour toutes les méthodes de représentation d'un texte, notamment le codage en n-grammes, qu'on va étaler par la suite, qui s'en passe d'une bonne partie de ces prétraitements en s'attaquant aux documents, pratiquement, dans leurs états bruts.

2.4- Définition de descripteurs

La définition ou l'extraction de caractéristiques au sein d'un texte est une phase décisive puisque la représentation déduite doit conserver au mieux l'information contenue dans le texte.
Ces caractéristiques constituent les éléments informationnels composant le document. Le plus petit élément informationnel étant le caractère, à un niveau supérieur on a le mot, regroupant un ensemble de caractères, puis à un niveau plus global nous pouvons définir les phrases, les paragraphes,…et pour finir le document lui-même.
La difficulté est donc le choix de cet élément de base : descripteur, terme ou caractéristique, puisque le processus de classification de textes en dépend directement.
Différentes méthodes sont proposées pour le choix des termes et les poids attribués à ces termes, des auteurs utilisent les mots comme descripteurs, d'autres utilisent les groupes de mots comme les mots composés, les expressions ou les collocations, comme d'autres qui préfèrent les techniques des n-grams, etc…
Dans la section suivante, nous allons définir les différentes sortes de termes, utilisés dans la littérature, pour la représentation d'un document texte.

2.4.1- Représentation en « sac de mots » « bag of words »

Le choix des mots comme descripteurs d'un document c'est le choix le plus intuitive, ainsi un texte sera représenté dans l'espace des mots par un vecteur dont chaque composante correspond au nombre d'apparition d'un mot dans le document, cette représentation est connue par « sac de mots » , « bag of words » .
(Salton & McGill, 1983), (Lewis, 1992), (Dumais & all, 1998), (Yang, 1999), (Vinot & all, 2003), (Pessiot & all, 2004), (Pothin & Richard, 2007) (Trinh, 2008), (Gotab, 2009), (Jégou & all, 2010), et bien d'autres ont préféré l'utilisation des mots comme descripteurs pour le codage des documents.
Pour clarifier la notion de mot, Y. Gilly dans son ouvrage « Texte et fréquence » (Gilly, 1988) l'a considéré comme étant une séquence de caractères appartenant à un dictionnaire, ou formellement, comme étant une suite de caractères séparés par des espaces ou des caractères de ponctuations (Cette définition n'est pas valable pour toutes les langues).

Ces descripteurs ont un vrai privilège de posséder un sens explicite cependant la représentation en « bag of words » affiche plusieurs anomalies :

- Le problème des mots composés comme : Arc-en-ciel, peut-être et le problème des sigles comme : APN, FAF, IBM.
- La présence parmi les descripteurs, des mots outils, qui constituent une grande part des mots d'un texte, mais qui portent peu d'informations utiles pour classer un texte.
- La distinction des mots d'une même famille en raison de leur variation morphologique (ex : écrire, écriture, écrit, écrivain,…) est en général un handicap, car chaque variation a une fréquence très faible alors que les regrouper permettrait d'avoir des fréquences importantes et d'amoindrir le phénomène d'imprécision des fréquences évoqué dans (Jalam ,2003).
- Enfin cette représentation est un regroupement en vrac de tous les mots du document « sac de mots » sans prendre en compte les combinaisons et l'ordre des mots dans la phrase entraine une perte dans la sémantique du texte.

Pour y remédier, un prétraitement linguistique amené par l'application des procédures de lemmatisation et de stemming, avant la représentation des documents, est indispensable.

Marionnaud: Union et Etudes Investissement franchit 5% des droits de vote
PARIS, 31 juil (AFP) - La société Union et Etudes Investissement (caisse Nationale de Crédit Agricole) a franchi en hausse le seuil de 5% des droits de vote du groupement français de parfumerie Marionnaud et détient désormais 292.157 actions, soit 8,09% du capital et 5,05% des droits de vote, a indiqué vendredi le Conseil des Marchés Financiers.
Ce franchissement de seuil résulte de l'acquisition de 11.460 actions, précise le CMF.

Mot	Occur.	Mot	Occur.	Mot	Occur.
a	2	détient	1	le	3
acquisition	1	en	1	marchés	1
actions	2	et	4	marionnaud	2
agricole	1	études	2	nationale	1
caisse	1	financiers	1	parfumerie	1
capital	1	franchi	1	précise	1
ce	1	franchissement	1	résulte	1
cmf	1	franchit	1	seuil	2
conseil	1	français	1	société	1
crédit	1	groupement	1	soit	1
de	9	hausse	1	union	2
des	4	indiqué	1	vendredi	1
droits	3	investissement	2	vote	3

du	2	l	1		
désormais	1	la	1		

Tableau 2.1 : Exemple de la représentation en « sac de mots »
Les chiffres et dates sont supprimés de la représentation

2.4.2- Représentation des textes par des collocations

Cette approche proposée ici, consiste à regrouper certains mots (collocations) afin d'obtenir des descripteurs ou expressions plus porteurs de sens au lieu d'utiliser des mots isolés composant le texte.
Identifier des collocations consiste à trouver des mots qui "vont ensemble" et qu'il est naturel de trouver proches dans le langage. Pour former ces groupes de mots, on n'a pas besoin de syntagmes nominaux, juste des paires de mots qui peuvent être séparés par des mots vides

Le but n'est pas ici de chercher à analyser les textes d'un point de vue syntaxique, mais les représenter selon un ensemble d'usages de la langue, qui ont une influence sur le système classification. (Par exemple : *repas-bien-garni*, *parler-en-connaissance-de-cause, tout-à-fait-normal*).

Rémi Lavalley, Patrice Bellot et Marc El-Bèze dans (Lavalley & all, 2009), expliquent pourquoi le fait de considérer une suite de mots comme une seule unité informative permet d'améliorer les performances d'un système de classification dans leur article « Interactions entre le calcul de collocations et la catégorisation automatique de textes ».
Tout d'abord pour augmenter la significativité du terme : par exemple, si nous arrivons à repérer l'expression « *effet particulièrement désagréable* » dans un texte, on pourra préjuger que la critique est négative, alors qu'un système classique aurait pris les mots séparément et aurait pu décider autrement, par exemple :
– *effet* : fait pencher vers une critique positive (comme dans l'expression "*cette odeur fait bon effet*") ;
– *désagréable* vers une critique négative.
Ainsi, en traitant l'expression dans son intégralité nous pensons accroître son pouvoir discriminant.
La seconde raison qui pousse à penser que l'on peut améliorer les résultats, selon (Lavalley & all, 2009), vient du fait qu'on peut envisager de créer des collocations propres à une catégorie pendant la phase d'apprentissage.
Pour extraire les collocations présentes dans le corpus d'apprentissage, ils se sont appuyés sur la méthode du Rapport de Vraisemblance.
Néanmoins, parmi les grands problèmes dans cette approche est de savoir lesquelles garder : toutes n'ayant pas la même pouvoir discriminant sur la classification finale, certaines peuvent en effet se recouper.
Les algorithmes d'extraction des collocations, appliquent les règles dans l'ordre dans lequel se trouvent les mots (parcours gauche-droite de la phrase).
Un autre problème posé est de savoir où s'arrêter (nombre de termes associés), car créer des combinaisons trop grandes entraîne des problèmes de fréquence faible (faible probabilité d'apparition de ces collocations).
Aussi, ils ne traitent que des règles assemblant les mots deux à deux (et pas directement trois à trois ou quatre à quatre par exemple)
De plus, il faut trouver un moyen de trouver des collocations "à trous" (mots non obligatoirement consécutifs).

Enfin, malgré qu'il s'agit d'une méthode nouvelle, pour laquelle de nombreuses améliorations sont envisageables, comme même, il existe un certain nombre de travaux de développement de méthodes pour trouver des collocations évoquées dans (Lavalley & all, 2009), par exemple celle de Yu J., Jin Z.,Wen Z.(2003),ou celle de Smadja F. A., McKeown K. R(1990) ou encore Seretan V., Nerima L., Wehrli E (2004), pour une méthode utilisant des filtres syntaxiques appliqués au corpus du Web).

2.4.3- Représentation des textes par des phrases

Ces techniques sont venues pour remédier à la déstructuration syntaxique causée par la représentation en "sacs de mots". Les résultats fournis par ce type de représentation « sac de mots » se basent finalement sur des mots éparpillés composant des textes qui sont très éloignés de ceux qu'ils sont censés représenter. Mais les techniques de traitement statistique (Bayes, Markov...) s'apprêtent mal des représentations à partir de phrases en raison de leurs caractéristiques irrégulières et exceptionnelles (longueur, redondance, bruit, structure compliquée...). Beaucoup de chercheurs s'y sont cassé les dents ayant abouti souvent à des solutions dérivées plus simples.
Logiquement, une telle représentation doit obtenir de meilleurs résultats que ceux obtenus via les mots en raison de la richesse sémantique de la phrase, cependant leurs propriétés statistiques ne permettent pas de définir des hypothèses statistiques fiables (Lewis, 1992) car, le grand nombre d'assemblages possibles de mots engendre des faibles fréquences et trop aléatoires, ne permettant pas d'approximer le risque réel de manière correcte grâce au risque empirique.
L'utilisation de "sac de phrases" entraîne évidemment un problème de taille (pour n mots il y existe potentiellement n^k combinaisons de longueur k).
Pour y remédier, on ne considère pas toutes les séquences possibles mais on tente d'effectuer une sélection des phrases, en privilégiant celles qui sont sémantiquement riche. Dans la phrase "*Le gentil lapin orange mange la carotte bleue*" par exemple, on peut dire que des séquences comme "*gentil lapin orange*", "*carotte bleue*", "*lapin orange*", ... sont porteuses de sens. Alors que les séquences "*orange mange*", "*le gentil*"…etc. sont insignifiantes.
Une autre approche de (Caropreso & all, 2001) qui proposent d'utiliser des phrases statistiques comme descripteurs au lieu des phrases grammaticales qui ont amélioré considérablement la performance du classifieur. Une phrase statistique est définie par (Jalam ,2003), comme une collection de mots adjacents (mais pas forcément classés) qui apparaissent ensembles mais qui ne respectent pas forcement les règles grammaticales.
Une autre étude menée par (Scott & all, 1999), démontre que ce type de représentation améliore la qualité des résultats par rapport aux méthodes de type « sac de mots » lorsque les documents étudiés sont limités en nombre et taille. De plus, ce type de représentation présente de grandes variations dans la qualité de ses résultats en fonction du type de documents à classifier. En effet, lors de ces tests il a utilisé une base de documents issus de Reuters et une base de textes de chansons folkloriques américaines. Les résultats sont bons sur Reuters mais très mauvais sur les textes de chansons.

Toutefois, la représentation des textes par des phrases est un domaine dont les recherches restent toujours actives.

2.4.4- Représentation des textes avec des racines lexicales (stemming)

L'opération de stemming expliquée précédemment, vise de ramener un mot à une de ses parties qui le caractérisent (racine) ainsi que tous les mots qui lui sont linguistiquement liés plutôt que considérer les mots entiers (on parle de *stem* en anglais).

Ainsi, le stem de numériseur serait *numéris* regroupant aussi : *numériseur, numériseurs*, *numérisée, numériser, numérisation, numérisations*, etc... On voit donc bien, l'intérêt du stemming puisque qu'il rassemble plusieurs mots de significations très proches dans un même groupe. La lemmatisation n'aurait pas pu regrouper *numériseur* et *numérisation* dans le même ensemble. Malheureusement, cette opération n'étant pas basée sur des contraintes linguistiques puissantes, peut conduire, à une amplification du bruit et des confusions sémantiques, en regroupant par erreur des mots de différentes significations, peuvent être générées. Comme la racine lexicale *port* qui regroupe dans le même ensemble le verbe *porter* et le nom *port* (Lieu pour les bateaux) alors que sémantiquement sont très distincts. De plus, le terme œil et son pluriel yeux ne pourront jamais être ramenés au à la même racine avec une simple opération de stemming. Pour éviter ces confusions, cette opération doit être utilisée avec précaution. (De Loupy, 2000)
La représentation des textes par ces stems peut apporter des résultats supérieurs à ceux obtenus par les lemmes (que nous allons voir dans ce qui suit), démontré et approuvé par de Loupy, et bien meilleur que le codage de type « sac de mots » ou chaque variation d'un mot est considéré comme une nouvelle composante du vecteur. Alors, on peut facilement imaginer combien on va gagner en question de dimensionnalité en optant pour les stems comme descripteurs.

2.4.5- Représentation des textes avec des lemmes (lemmatisation)

La lemmatisation décrite auparavant, consiste à utiliser l'analyse grammaticale afin de remplacer les verbes par leur forme infinitive et les noms par leur forme au singulier.
La substitution des mots par leur lemme réduit également l'espace des descripteurs comme pour les racines, et permet de représenter par un même descripteur des termes de même signification. Par exemple, le remplacement des mots *banking, bank*, *banks*, par l'unique racine *bank* semble être rentable tout comme le remplacement des formes conjuguées *rebondit* et *rebondi* par le lemme *rebondir*.
Les mêmes confusions d'ambiguïté peuvent être entrainées en représentant par un même descripteur des mots avec des sens différents, comme par exemple *glace* qui peut être *un miroir* ou « *une glace aux chocolats"*.
L'ambiguïté peut être causée aussi par le simple remplacement de la forme plurielle d'un mot par sa forme singulier comme *actions* qui est représenté par le descripteur *action*. Dans un contexte économique, le mot *actions* se réfère couramment à des actions d'entreprises et n'a rien à voir avec la notion *action* employée par exemple dans la phrase : « Le plan d'*action* de l'état ».

2.4.6- Représentation des textes avec la méthode des n-grammes

Une autre approche pour coder les documents émerge : les n-grammes (Shannon, 1948). On définit un n-gramme (n-gram) par est une séquence de n caractères : bi-grammes pour n=2, tri-grammes pour n=3, quadri-grammes pour n=4, etc..On n'a plus besoin de chercher les délimiteurs(les espaces ou les caractères de ponctuations) comme c'était le cas pour les mots.
Quelques auteurs admettent les n-grammes comme une chaine non ordonnée de caractères; par exemple un tri-grammes peut être constitué du 2éme, 4éme et 1ér caractère, d'autres auteurs n'autorisent pas ce désordre. Pour notre cas, on va admettre qu'un n-grammes désignera une chaîne de n caractères consécutifs.
Pour un texte quelconque, les n-grammes correspondants sont générés en faisant déplaçer un masque de n caractères sur tout le texte. Ce déplacement s'effectue caractère par caractère, à chaque déplacement la séquence de n caractères est enregistrée, l'ensemble de ces séquences constitue l'ensemble des n-grammes représentant le texte. (Miller & all, 1999)

Par exemple, pour générer les 3-grammes de la phrase "Tu es libre", on obtient : "Tu " , "u e" , " es" , "es " , "s l" , " li" , "lib" , "ibr" , "bre".
Historiquement, les n-grammes étaient conçus pour la reconnaissance de la parole, pour prédire l'apparition de certains caractères en fonction des autres caractères mais par la suite, le concept n-grammes a été bénéfique, pour le domaine de recherche d'information et la classification de textes, avec plusieurs travaux qui ont démontré que cette segmentation ne faisait pas perdre d'information.
Il est à noter qu'il existe une autre utilisation de cette notion, où le n-gramme est une suite de n mots et dont l'intérêt est de détecter les liens entre les mots pour en déduire quel mot va apparaître conditionnellement à la présence des n-1 mots précédents (Brown & all, 1992) (dans ce qui suit nous allons considérer le n-gramme dans le sens de séquence de n caractères).
Ainsi pour un alphabet de 26 lettres on obtient $26^2 = 676$ bi-grams ou $26^4 = 456\ 976$ quadri-grams, concernant les mots, pour un dictionnaire de 20 000 mots on obtient $20\ 000^2 = 400$ millions de bi-grams et $20\ 000^3 = 8\ 000$ milliards de tri-grams.
Plusieurs spécialistes récents dans le domaine, ont employé ce type de codage pour repésenter leurs documents, par exemple : (Ralaivola, 2006), (Gotab, 2009), (Généreux, 2010). Cette technique a servi pour coder, même de textes en langue chinoise (Wei, 2009).

Enfin, notons que la technique des n-grammes est tojours trés utilisée, pour représenter les textes, en raison de ses avantages qui vont être dévoilés dans le chapitre 6, qui nous ont incité nous aussi, d'ailleurs, pour l'adopter dans notre étude.

2.4.7- Représentation des textes par des combinaisons de termes

Au lieu de prendre les termes un par un comme descripteurs, l'idée ici est de combiner linéairement des termes pour améliorer la qualité des résultats. L'intérêt est corriger les anomalies liés aux ambigüités et redondances du vocabulaire en combinant plusieurs termes pour avoir des nouvelles variables artificielles, jouant le rôle de nouveaux « termes » (Jalam ,2003).
Une approche typique a été proposé par Deerwester, S.Dumais et autres dans (Deerwester & all, 1990), appelée Latent Semantic Indexing (LSI) appuyée sur l'Analyse des Correspondances Factorielles, introduite dans plusieurs recherches dans le cadre du traitement des données textuelles.
On va revenir sur cette méthode ultérieurement dans le cadre de la sélection de termes dans la section 2.5.3.3 puisque c'est une, parmi les techniques de réduction de dimensionnalité

2.4.8- Représentation des textes basée sur les concepts

Les approches précédentes n'extraient pas la sémantique d'un document mais simplement une comparaison morphologique. Si on peut supposer que chaque terme a un sens, il est plus difficile de prouver que deux documents étant composés des mêmes termes aient forcément le même sens. Les auteurs proposent donc, une nouvelle approche de représentation textuelle « plus sémantique » basée non pas sur les termes présents sur le texte à traiter mais sur les concepts correspondants. Ainsi, au lieu de définir un espace vectoriel dont chaque composante représente un terme (mot, stem, lemme, ou n-gram), on projette l'ensemble de termes du texte sur un ensemble fini de concepts.
Un concept peut représenter un objet matériel, une notion, une idée (Uschold & King, 1995). Trois éléments constituent la notion de concept { terme(s), notion, objet(s)}, un terme ou plusieurs, une notion et un ensemble d'objets. La notion, également appelée intention du concept, contient la sémantique du concept, exprimée en termes de propriétés et d'attributs.

L'ensemble d'objets, également appelé extension du concept, regroupe les objets manipulés à travers le concept ; ces objets sont appelés instances du concept. Par exemple, le terme « stylo », a pour intention « instrument nécessaire pour écrire ou dessiner », et a plusieurs réalisations : « marqueur, stylo à bille, stylo à encre, etc... ».
Un concept est ainsi doté d'une sémantique référentielle (celle imposée par son extension) et d'une sémantique différentielle (celle imposée par son intension).
Un concept ayant une extension vide est appelé concept générique, ces concepts génériques correspondent généralement à des notions abstraites (par exemple, la « vérité »).
Le stylo qui est lui-même composé d'un bouchon et de l'encre et autre chose montre que cette notion ne peut se définir qu'en utilisant d'autres concepts comme « bouchon », « encre » etc..
Les concepts manipulés dans un langage donné sont organisés au sein d'un réseau de concepts liés par des propriétés conceptuelles sous forme d'ontologie lexicale appelée thesaurus.
Un thesaurus est un ensemble de termes normalisés basé sur une structuration hiérarchisée. Les termes y sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques.
L'avantage d'une telle méthode est en particulier de réduire les problèmes d'ambigüité et de synonymie dans le vocabulaire et de la construction syntaxique (restituer l'ordre des termes). Cette nouvelle approche de représentation permet donc, une factorisation des termes par regroupement de leur champ sémantique. (Jaillet & all, 2003). En effet plusieurs synonymes seront représentés par un seul concept, d'où une réduction de dimensionnalité considérable dans l'espace de codage.

Une expérience a aboutit, celle du Langage Universel d'Echanges (Universal Networking Language) défini dans (Uchida & Zhu, 1999). UNL est un formalisme permettant de modéliser la sémantique de chaque texte par un graphe. Toute expression en langage naturel peut être modélisée en UNL puis traduite dans n'importe quelle langue cible. En UNL, chaque phrase d'un document est définie par un hyper graphe ou les nœuds sont les concepts et les arcs orientés les relations entre les concepts. Puisque la comparaison de graphes n'est pas toujours évidente, une méthode de représentation de ces graphes pour être exploitée dans un processus de catégorisation est décrite dans (Shah & all, 2002).
Par ailleurs, d'autres approches utilisent une représentation de type conceptuelle, c'est le cas de WCM (Word Category Map) (Kohonen & all, 2000).

Cependant, les deux inconvénients majeurs de ce type de codage restent, que les noms propres du texte ne sont pas pris en compte (Absents du thésaurus puisque ces derniers sont sémantiquement vides par définition), et le coût excessif pour la conception, la réalisation et la maintenance d'une telle solution appuyée sur les ontologies.

2.5- Sélection de descripteurs

2.5.1- Besoin de la sélection de descripteurs

Pour une problématique de classification, l'ensemble des descripteurs est constitué de l'ensemble des termes du corpus, un terme pouvant être un mot, un stem ou un n-gramme, etc.., ce qui peut représenter plusieurs centaines de milliers de termes, même après les prétraitements appliquées dans la première phase qui ont procédé à l'élimination des mots les plus fréquents et les plus rares, soit parce qu'ils n'étaient pas discriminants (Mots vides très faiblement informatifs), soit parce qu'ils n'étaient pas exploitables statistiquement (très faible fréquence), le nombre de termes s'avère encore très élevée (De quelques dizaines de milliers à plusieurs centaines de milliers de termes). Parmi l'ensemble des descripteurs restants, on peut penser que tous ne sont pas nécessairement discriminants voire nuisible pour le système.

Ainsi, Il est nécessaire de diminuer davantage et choisir les descripteurs les plus appropriés (ceux qui assureraient les meilleures performances au classifieur), qui vont être utilisés comme vecteurs d'entrées avant de pouvoir utiliser un modèle d'apprentissage.

La sélection de descripteurs est un des principaux enjeux du processus, puisque du choix des descripteurs et de la connaissance précise de la population va dépendre la mise au point du classifieur. L'information nécessaire à la construction d'un bon modèle de prédiction peut être disponible dans les vecteurs d'entrées mais une sélection inappropriée de descripteurs ou d'exemples d'apprentissage peut faire échouer l'opération. (Zighed & Rakotomalala, 2002).

Evidemment, que quel que soit le modèle statistique utilisé ultérieurement, si la représentation des textes n'inclut pas certains descripteurs ou si les descripteurs retenus sont trop nombreux ou mal choisis, le classifieur aura des performances médiocres.

Les entrées non discriminantes doivent être supprimées pour deux raisons différentes :

- Pour réduire le temps de calcul : Plus le nombre d'entrées est grand, plus le nombre de paramètres à déterminer est élevé, ce nombre intervient dans l'expression de la complexité de l'algorithme qui va exiger un temps de traitement plus important. (Pour les modèles tels que les réseaux de neurones, le nombre de poids du réseau croît linéairement avec le nombre de descripteurs utilisés en entrée du modèle).

- Pour diminuer le sur-apprentissage : Comme les bases d'apprentissage sont limitées, des associations inattendues peuvent apparaître entre des descripteurs non informatifs et des classes ; elles peuvent avoir une influence négative sur la qualité du modèle. Il faut alors disposer d'une base d'exemples plus grande afin de diminuer le sur-apprentissage résultant du nombre trop important de paramètres, dont certains sont de très faible fréquence, par rapport aux textes du corpus d'apprentissage : on ne peut pas construire des règles stables à partir de quelques apparitions d'un terme dans l'ensemble d'apprentissage.

Le sur-apprentissage dépend aussi beaucoup du modèle d'apprentissage utilisé, en effet certains sont capables de sélectionner les termes informatifs et ne sont pas affectés par un pléthore d'informations inutiles alors que d'autres considèrent que tous les termes sont discriminants, une sélection préalable est donc indispensable.

Les méthodes de sélection de descripteurs ont donc pour but de choisir parmi un ensemble de descripteurs possibles, les descripteurs les plus importants, c'est-à-dire ceux qui vont permettre d'obtenir de bonnes performances sur une base différente de la base d'apprentissage.

2.5.2- Le nombre de descripteurs conservés

Les méthodes de sélection de descripteurs fournissent, en général, une liste de descripteurs classées par degré d'importance, cette notion d'importance qui dépend de la méthode de classement; l'intérêt des différentes approches de réduction de dimensionnalité est d'avoir un ensemble de descripteurs plus réduit mais informatif. Il reste ensuite à fixer le nombre de descripteurs à garder dans cet ensemble. (Stricker, 2000)
La méthode de classification va être forcément, très décisive dans le seuil à fixer pour le nombre de descripteurs à conserver. Comme par exemple, dans un réseau de neurones, réduire la dimension des vecteurs d'entrées est très recommandé alors qu'une approche SVM est capable de traiter des listes plus longues de termes.
Nous cherchons donc, à supprimer des termes de la représentation des textes, tout en sachant que chaque suppression de terme entraîne une perte d'information ; il faut trouver le bon compromis entre, d'une part, la nécessité de réduire l'espace des descripteurs avec moins de redondances possibles et, d'autre part, le nécessité de garder suffisamment d'informations.

Plusieurs chercheurs dans le domaine ont essayé de réaliser ce bon compromis, comme par exemple (Dumais & all, 1998) construit son modèle à base des SVM en prenant en considération seulement 300 termes sur le corpus Reuters, par contre (Joachims, 1998) pense autrement en considérant que tous les termes du corpus, fournis après les prétraitements (Presque 10000 termes), sont informatifs et sa conservation en entrée dans son modèle est nécessaire, sauf que les résultats fournis sont moins bons que pour (Dumais & all, 1998), ce qui amène à dire que les 10000 termes gardés par Joachims n'étaient pas tous utiles.
Plusieurs auteurs, dans leurs travaux, ont proposé différents nombres de descripteurs pour représenter les textes, de 180 à 100 jusqu' aux 20 premiers descripteurs, sans atteindre les qualités des classifieurs. Une synthèse des différentes expérimentations portantes sur le nombre adéquat de descripteurs retenus, qui n'impliquent pas qu'une grande dimension est nécessaire pour avoir des meilleurs résultats, est évoquée dans (Jalam ,2003).

2.5.3- Les méthodes de sélection de descripteurs

La majorité des méthodes sont basées sur le calcul d'une statistique pour chaque terme qui représente son importance pour le document où il figure ou pour le corpus complet, puis à sélectionner les termes les plus importants. Il existe plusieurs formules statistiques pour mesurer la quantité d'information apportée à partir du nombre du nombre d'apparitions du terme dans la classe et hors de la classe.
Dans cette phase, il s'agit aussi de générer un profil pour chaque catégorie. Le profil d'une catégorie doit contenir tout les termes qui caractèrisent cette catégorie par rapport aux autres. Pour mesurer ces statistiques et construire ces profils, il est nécessaire d'utiliser une méthode de sélection de termes.
En effet, il existe plusieurs méthodes de selection de termes :

2.5.3.1- Principales méthodes

Dans ce qui suit nous allons présenter les principales formules utilisées pour mesurer la quantité d'informations contenue dans les termes pour les documents ou les classes, dont les performances sont comparées dans plusieurs études.

- La *Fréquence-document* (Document Fréquency) : Une première méthode de sélection qui peut être considérée comme une méthode de prétraitement approfondie : elle est très simple puisqu'elle correspond simplement au pourcentage de documents dans lesquels le terme apparaît, cette méthode conduit à supprimer les termes très fréquents et très rares afin de conserver les mots les plus importants avec le risque de supprimer des termes très riches et informatifs pour le système. Pour écarter les mots les plus fréquents, nous fixons un seuil maximal de fréquence n'autorisant pas de sélectionner les termes présents dans une très forte proportion de textes (ex : un terme qui apparaisse dans 180 textes d'un corpus de 200 textes n'est pas sélectionné), de même un seuil minimal est fixé pour éliminer les termes très rares (ex : un terme qui a moins de 5 apparitions dans tout le corpus n'est pas sélectionné)

  La Fréquence-document du terme T : $\mathcal{P}$ (T)

- Le *Gain d'Information* : C'est une mesure nécessaire pour prédire la catégorie d'un document selon la présence ou l'absence d'un mot dans un texte, on peut interpréter cette statistique par la quantité d'information apportée par la présence ou l'absence d'un terme dans un document. Un gain d'information important indique que le terme contient plus d'information pour le texte, en revanche, une perte d'information indique que le terme contient moins d'information nécessaire pour classer les textes avec ce terme.

Le Gain d'Information apporté par T=$\mathrm{P}(T,C)\ log\ \frac{P(T,C)}{P(T)P(C)} + \mathrm{P}(T,C)\ log\ \frac{P(\overline{T},C)}{P(\overline{T})P(C)}$

- L'*Information Mutuelle* : Représente la corrélation entre deux variables aléatoires; pour notre cas, les deux variables sont le terme et la classe ; Cette mesure a été fréquemment utilisée pour la catégorisation de textes pour effectuer la sélection de descripteurs, utilisée par (Lewis, 1992), (Moulinier, 1996) et (Dumais et all, 1998).

  L'Information Mutuelle du terme T pour la classe C = $\log\ \frac{P(T,C)}{P(T)P(C)}$

- Le *Chi-deux univarié* ($\mathcal{X}^2_{\text{univarié}}$) : mesure le degré d'indépendance de deux variables aléatoires (présent ou absent), ici le terme (*T*) et la classe (*C*). Cette mesure se calcule à partir d'une table de contingence (2x2), indiquant le nombre de textes associés à la classe *C* où le terme *T* est soit présent (i textes dans le tableau), soit absent (k textes dans le tableau). La même chose est faite pour les textes non associés à *C,* nous calculons j et l. Ainsi de suite chaque terme candidat aura son tableau.

	Classe *C* présente	Classe *C* absente
Terme *T* présent	i	j
Terme *T* absent	k	l

Tableau 2.2 : Table de contingence selon le nombre de documents

Ainsi la formule du $\mathcal{X}^2_{\text{univarié}}$ sera comme suit :

$$\mathcal{X}^2_{\text{univarié}}(\mathrm{T,C}) = \frac{(i+j+k+l)\ (il-jk)^2}{(i+k)(j+l)(i+j)(m+l)}$$

Cette mesure a été utilisée pour la sélection des descripteurs dans (Schütze et all, 1995) et (Wiener et all, 1995).

- Le *Chi-deux mutivarié* ($\mathcal{X}^2_{\text{mutivarié}}$) est une méthode supervisée permettant la sélection de termes en prenant compte de leurs fréquences dans chaque classe comme l'univarié ajoutées aux interactions termes/termes et termes/classes. L'idée consiste à extraire les meilleurs termes qui caractérisent une classe par rapport aux autres. Les principales caractéristiques de cette méthode sont évoquées dans (Jalam ,2003) et (Clech, 2004).
  Un tableau de contingence (termes - classes) de dimension N x M sera construit, N étant le nombre total de termes, M le nombre de documents.
  Cette mesure a été utilisée pour la sélection des descripteurs dans (Jalam ,2003) dans le cadre de la catégorisation de textes multilingues.

2.5.3.2- Inconvénient commun (Association de termes)

Chacune de ces techniques sélectionne un terme pour son d'importance, mais deux termes peuvent ne pas avoir d'importance pour la classification du document pris indépendamment, alors que la présence simultanée de ces deux termes peut avoir un rôle important. Par exemple, les deux termes *droits* et *homme* ne sont pas des termes très informatifs pris un par un, mais l'association de ces deux mots compose le concept de *droits de l'homme* qui a une signification très précise, reste à prendre en considération la distance entre les termes : *droits* et *homme* qui peuvent être présents dans le même document séparément sans la présence d'une interaction entre ces deux termes. Aucune des méthodes présentées ci-dessus ne résout ce problème qui représente un défaut commun à toutes les statistiques précédentes, une

méthode dite « l'orthogonalisation de Gram-Schmidt » issue des méthodes utilisées pour trouver la solution des moindres carrés d'un problème linéaire par rapport à ses paramètres résout ce problème partiellement. Cet algorithme itératif classe les termes par ordre décroissant de leur pouvoir discriminant, du plus important au moins important tout en tenant compte de ceux déjà classés. Une description détaillée de cette procédure et son application peut être trouvée dans (Dumais & Chen, 2000).
Ajoutons au problème d'association de termes, des problèmes de synonymie et polysémie du vocabulaire qui ne sont pris en charges par ces formules statistiques classiques.

2.5.3.2- Autres approches

Les méthodes de représentation des textes à base de concepts ou les combinaisons de termes, sont naturellement des techniques pour diminuer le nombre de termes, qui peuvent solutionner les problèmes de synonymie et polysémie. Ainsi comme on a vu précédemment, le texte sera représenté de telle manière que le descripteur ne sera plus un terme simple mais une combinaison de termes ou il va correspondre à un concept sémantique. Cela dépasse donc la racinisation qui ne cherche qu'à regrouper les mots de même famille et non pas les mots de même sens.
Une première approche appelée *Term Clustering* (Lewis, 1992) consiste à regrouper plusieurs termes pour former un nouvel attribut. Chaque attribut est donc censé représenter un concept sémantique. Cette association de plusieurs termes avec un concept permet de gérer la synonymie. La polysémie des mots est également prise en compte en permettant à un terme d'appartenir à plusieurs groupes.
Une autre technique nommée *Indexing by Latent Semantic Analysis* proposée par S.Deerwester et S.Dumais (Deerwester & all, 1990). LSI est basée sur le principe d'une structure latente des termes qu'on peut retrouver à l'aide de l'analyse factorielle, qui décompose la matrice d'occurrence $[M_{ij}]$ en valeurs singulières (M_{ij} est le nombre d'occurrences du terme *j* dans le document *i*). La décomposition revient à changer la représentation par un changement de base. Chaque descripteur est donc représenté par une combinaison linéaire de termes. Seulement les *j* axes de plus grandes valeurs singulières seront préservés. LSI a été utilisée pour sélectionner les entrées d'un réseau de neurones dans (Wiener & all, 1995).
Ces deux méthodes, LSI et Term Clustering, améliorent les performances de quelques pourcents par rapport les autres méthodes de sélection de termes moins complexes, mais elles nécessitent un temps de calcul supplémentaire pendant l'apprentissage et impossibilité de traiter un nouveau document sans relancer tout le processus. L'utilisation de ces deux techniques dépend du contexte de l'application et de la vitesse des changements susceptibles d'intervenir dans le corpus.

2.5.4- Sélection des termes par rapport la classe ou tout le corpus

(Jalam ,2003) a rappelé ce qui a été évoqué par (Sebastianni, 2002) sur la réduction des dimensions qui peut être localement ou globalement :

- Réduction locale : Chaque classe est caractérisée par un profil composé d'un ensemble de termes, et chaque texte sera représenté par une liste de termes dépendante de la catégorie.

- Réduction globale: Contrairement au cas précédent, un texte est représenté par une seule liste de termes dans tous le corpus indépendamment des classes.

2.6- Pondération ou calcul de poids

Le tableau (termes x documents) est constitué par le nombre d'apparitions du terme dans le document du corpus. Cette information de base doit être pondérée en fonction de divers paramètres liés au document lui même (ex : le nombre de termes par document) ou au corpus en intégralité (ex : le nombre de termes du corpus). L'intérêt de cette pondération est mieux exploiter l'information contenue dans le document pour améliorer les performances d'un système de classification de textes (SPARCK-JONES, 1972).

Plusieurs systèmes de pondération ont été développés dans la littérature, qui se reposent, tous sur les deux hypothèses suivantes :

- Plus le nombre d'apparitions d'un terme dans un texte est important, plus ce terme est discriminant pour la classe associée.
- Plus le nombre d'apparitions d'un terme dans le corpus est important, alors moins ce terme peut discriminer les textes.

Voici, un petit aperçu sur les pondérations les plus habituellement utilisées signalées dans (Clech, 2004):

- Commençant par le choix le plus simple, qui ne s'intéresse que sur la présence ou la non-présence d'un terme dans le texte, il consiste à utiliser une pondération binaire : 1 si le terme est présent une ou plusieurs fois dans le document, 0 dans le cas contraire .Cette représentation binaire est historiquement la plus ancienne et la plus simple.
  Néanmoins, cette fonction est moins utilisée pour les méthodes statistiques, car ce codage supprime de l'information qui peut être utile : l'apparition du même mot plusieurs fois dans un texte peut constituer un élément de décision important.
- Les représentations fréquentielles sont aujourd'hui les plus utilisées et sur lesquelles notre étude va être basée. Plusieurs variantes s'illustrent :
  - Une première approche consiste à utiliser seulement le nombre d'occurrences des termes. Cette pondération ne peut être valable que dans les documents de même taille, sinon elle avantage les termes qui se répètent souvent dans les documents les plus longs.
  - Une autre approche simple consiste à utiliser la fréquence d'un terme par rapport au nombre de termes composant le texte.
  - La pondération TFxIDF corrige la fréquence du terme(Term Frequency) en fonction de sa fréquence dans le corpus(Inverse Document Frequency). La correction se fait en multipliant du rapport des n textes du corpus sur le nombre de documents contenant le terme. Le logarithme est utilisé pour lisser les résultats.
  - Le TFC normalise le TFxIDF en fonction de l'ensemble des termes du document.
  - La pondération LTC est du TFC sur lequel on applique du logarithme afin d'atténuer les différences de nombre d'occurrences pour diminuer les effets des différences de fréquences.
- La pondération basée sur l'entropie est la plus performante comparée à six autres méthodes, approuvée par les expérimentations de (Dumais, 1991) et confirmée par d'autres auteurs récémment. Selon l'auteur l'entropie devance le TFxIDF sur les cinq corpus testés. Néanmoins, cette méthode est assez complexe du fait qu'elle fait intervenir l'ensemble des autres textes, sollicitant un temps de traitement plus grand.

D'autres pondérations sont en pleine étude :

- La représentation séquentielle n'a fait l'objet de travaux que récemment car elle nécessite l'utilisation de modèles plus complexes et que son intérêt n'est pas toujours démontré.

Cependant, c'est une représentation naturelle qui permet de conserver l'ordre des mots d'un document.

- D'autres représentations plus riches existent notamment dans le domaine du Traitement de la Langue Naturelle (TALN) comme par exemple les représentations qui prennent en compte le rôle du mot dans une phrase (Nom, Verbe, Sujet, etc…). Ces représentations s'avèrent efficaces, cependant, les modèles qui les utilisent sont peu performants de ceux qui travaillent dans des espaces plus « simples ». Ces dernières ne sont pas présentées ici, pour plus d'informations sur celles-ci, il est intéressant de se reporter à (Chandra, 1998).

Un processus de classification automatique de textes employant des méthodes essentiellement statistiques peut-être représenté selon :

- Le modèle vectoriel
- Le modèle probabiliste
- Représentation séquentielle

2.6.1- Le modèle vectoriel

2.6.1.1- Représentation binaire

Historiquement, la première représentation d'un texte était sous forme de vecteur binaire, et malgré l'apparition de nouvelles formules pour pondérer les termes, cette façon de représenter un document, est restée toujours largement utilisée en raison du bon compromis fourni entre performance et complexité. Effectivement, en raison de sa simplicité, son temps de traitement est faible et en contrepartie ses résultats ne sont pas mauvais. Elle est appelée représentation « par mots clés ». La méthode consiste à transformer le texte en un vecteur dont les éléments renseignent sur la présence (valeur égale à 1) ou l'absence (valeur égale à 0) d'un terme dans le texte. Deux exemples de textes sont indiqués dans la figure 2.2, leurs vecteurs binaires correspondants sont représentés dans le tableau 2.3

« ce groupe est le groupe de la mort du mondial 2010»	« cette équipe est une équipe favorite au mondial 2010»
Document 1	*Document 2*

Figure 2.2 : Deux exemples de documents

Cette façon de représenter un texte, est peu informative car elle ne donne pas les informations nécessaires ni sur les occurrences d'un terme dans le document qui peut être une information importante pour l'opération de classification, ni sur la longueur du texte.

2.6.1.2- Représentation fréquentielle

Cette représentation consiste à présenter le texte sous forme de vecteur dont les éléments renseignent non seulement sur la présence ou l'absence d'un terme comme dans un vecteur binaire mais aussi informe sur le nombre de présences du terme dans le texte.
Ainsi, un document est transformé en un vecteur dont les composantes vont correspondre au nombre d'occurrences des termes dans le document.
Pour chaque document, un poids est attribué à chacun des termes qu'il contient. Une matrice « documents /termes » représente l'ensemble des documents (un vecteur est associé à chaque document, les composantes des vecteurs sont les poids des termes) (Salton & McGill, 1983).

Cette méthode conçoit le calcul du poids proprement dit des termes.
Aucune analyse linguistique n'est utilisée, ni aucune notion de distances entre les mots.
Ainsi les deux inconvénients majeurs de cette représentation, sont la non prise en charge des interactions des termes entre eux traduit par une indépendance de ces termes d'une part et d'autre part la déstructuration syntaxique du document causée par le fait que le modèle ne permet pas de conserver l'ordre des mots.
Un exemple de cette représentation est montré dans le tableau 2.3

2.6.1.3- Représentation fréquentielle normalisée

Du point de vue statistique**,** la représentation fréquentielle confronte un problème majeur du fait qu'un texte long sera représenté par un vecteur dont la norme sera supérieure à celle de la représentation d'un document plus court. Il est donc recommandé de normaliser la représentation fréquentielle par rapport à la taille du document. Ainsi le poids du terme sera le nombre d'occurrences de ce dernier dans le texte sur le nombre d'occurrences de tous les termes du texte. Le tableau 2.3 contient une représentation de ce type.

2.6.1.4- Vecteur TF-IDF

Dans le but d'avoir des représentations plus riches en informations que la représentation fréquentielle basique ou même sa version normalisée, une autre variante des représentations vectorielles s'illustre appelé le codage TF-IDF. Cette représentation se base principalement sur une certaine loi appelée loi de Zipf qui montre la façon dont les mots sont distribués dans un corpus.

a- Loi de Zipf :

La répartition des fréquences des termes dans un corpus a été étudiée empiriquement par Zipf (Zipf, 1949). Zipf est parti d'un principe général avant qu'il énonce cette loi mathématiquement. Cette loi réaffirme que la distribution des occurrences des mots dans un corpus donné n'est pas uniforme, certains mots apparaissent très fréquemment, tandis que d'autres apparaissent très rarement.
Les termes les plus informatifs d'un corpus de textes ne sont pas :

- Les termes qui se répètent souvent dans le corpus, qui sont généralement des mots outils. Les mots les plus fréquents en français sont les mots grammaticaux comme *le*, *la*, *les*, *et*... Sur le corpus Reuters, les cinq mots qui se répètent le plus sont : the, of, to, and, in. Cependant les mots non outils qui apparaissent fréquemment contiennent sûrement des informations fortes sur la sémantique du texte.
- Ni les termes les moins fréquents du corpus présents dans un seul ou quelques textes rédigés par des auteurs utilisant un vocabulaire très particulier ou même des termes issus de fautes d'orthographe.

Ces deux observations précédentes ne se contredisent pas et peuvent être récapitulées de la manière suivante : ***« Un mot est informatif dans un document si il y est présent souvent mais qu'il n'est pas présent trop souvent dans les autres documents du corpus ».*** (Denoyer, 2004)

La figure 2.3 illustre de manière graphique la loi de Zipf, qui montre clairement l'évolution de l'importance des mots par rapport leurs fréquences dans un corpus.

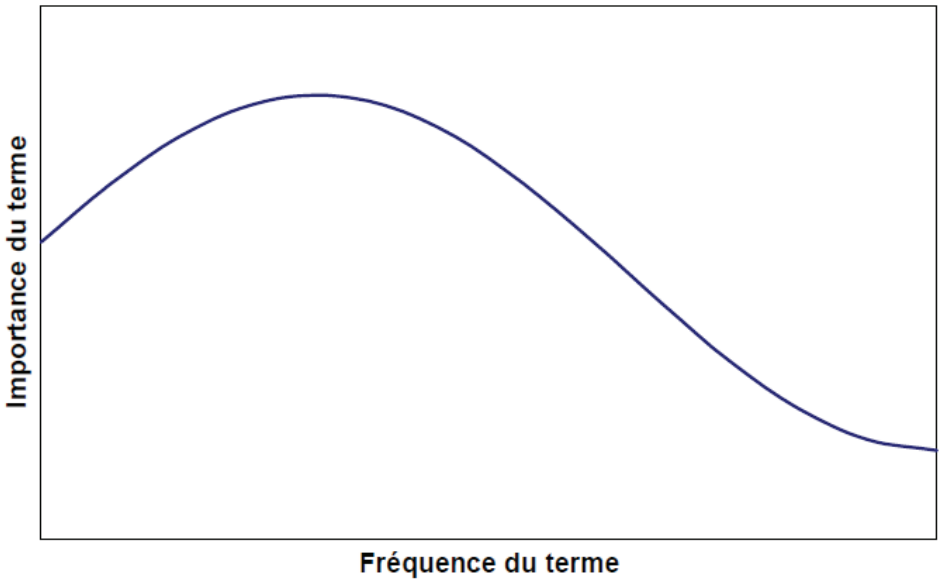

Figure 2.3 : Loi de Zipf

« Un mot est important si il n'est ni trop fréquent ni trop rare ».

Les observations de cette loi sur la distribution des fréquences des termes nous conduit à la suppression des mots fréquents et des mots rares en prenant en considération un seuil fixé préalablement, qui fait de cette loi une première méthode de réduction de dimensionnalité de l'espace des descripteurs. Donc il s'agit d'éliminer des termes inutiles pour les algorithmes d'apprentissage. Cette phase du processus est critique, car les mots écartés sont irrécupérables. De_ce_fait il faut être très attentif pour ne pas supprimer des mots sémantiquement riches.

b- Représentation TF-IDF :

Le codage *TF.IDF* a été introduit pour la prise en compte de la loi de Zipf dans le cadre du modèle vectoriel et qui donne parfois son nom à la méthode vectorielle. Le principe de base est que l'élément du vecteur représentant un texte se calcule en multipliant un facteur qui concerne l'importance du terme T dans le texte avec un autre qui concerne l'importance de ce terme dans tout le corpus :

$$\mathbf{L_{zipf}\ (T) = Poids\ dans\ le\ document\ *\ Poids\ dans\ le\ corpus}$$

Donc la formulation de la pondération s'appuie sur deux notions :

- La Fréquence du Terme (ou *Term Frequency* : *TF*) qui prend en compte le nombre d'occurrences du terme dans le document.
- et l'Inverse de la Fréquence en Document (ou *Inverse Document Frequency* : *IDF*) qui prend en compte le nombre d'occurrence du terme dans le corpus.

Ces deux notions sont combinées multiplicativement de façon à attribuer un poids d'autant plus fort que le terme apparaît souvent dans le document et rarement dans le corpus complet.

*Term Frequency * Inverse Document Frequency*

c- Variantes du TF-IDF :

Il est admis que l'occurrence d'un terme est une information importante, mais cependant, on considère généralement que *TF.IDF* ne doit pas être l'identité. Si, par exemple, un mot apparaît deux fois dans un texte, son importance n'est pas nécessairement deux fois plus

grande que s'il n'apparaissait qu'une seule fois. Pour améliorer la pondération des termes plusieurs variantes du *TF* ont été proposées :
Le *TF* peut être égal au nombre d'occurrence du terme, à son log, au log de son log, etc. L'utilisation du logarithme permet de diminuer l'importance des termes fortement répétés. En effet un terme à plus d'importance s'il passe de 1 à 2 occurrences que s'il passe de 20 à 21 occurrences. Contrairement au *TF*, un large consensus existe pour l'*IDF*.
Le modèle le plus classique est celui pour lequel la première valeur est égale à la fréquence du mot dans le document notée $TF(d, w_i)$ pour *term frequency* et la seconde valeur est égale à $Log\left(\frac{nbre\_doc}{DF(w_i)}\right)$ où nbre_doc est le nombre de documents du corpus et $DF(w_i)$ est le nombre de documents qui contiennent le mot *i (DF* signifie *document frequency)*.
Particulièrement dans ce cas là, cette représentation sera appelée représentation *TF-IDF* , Elle correspond à la représentation suivante :

$$TFIDF(d, w_i) = TF(d, w_i) \times Log\left(\frac{nbre\_doc}{DF(w_i)}\right)$$

Le tableau 2.3 inclut un exemple de représentation *TF-IDF*.

En général, on utilisera une version du vecteur *TF-IDF* normalisé, comme pour les représentations fréquentielles, afin d'éviter les problèmes posés par les différentes longueurs de textes. On aura ainsi, une représentation *TF-IDF* normalisée.
Plusieurs normalisations sont proposées, parmi elles, une mesure de pondération appelée *TFC* qui est similaire à celui de *TF-IDF* mais qui corrige les longueurs des textes par la normalisation en cosinus, afin de ne pas favoriser les documents les plus longs.

$$TFC(t_k, d_j) = \frac{TF \times IDF(t_k, d_j)}{\sqrt{\sum_{s=1}^{|r|} (TF \times IDF(t_s, d_j))^2}}$$

D'autres pondérations sont aussi utilisées, comme par exemple le LTC (Buckley & all, 1994) dont l'intérêt est la réduction des effets des différences de fréquences, ou encore le codage à base d'entropie utilisé par (Dumais, 1991).

Pour conclure(Salton & Buckley, 1988) et (Joachims, 1999) confirment que la représentation *TF-IDF* avec toutes ses variantes est la représentation la plus utilisée en recherche d'information aussi bien en recherche documentaire qu'en classification.

	Doc1	***Doc2***	***Doc1***	***Doc2***	***Doc1***	***Doc2***	***DF***	***Doc1***	***Doc2***
Ce	1	0	1	0	0.09	0	1	0.09	0
Cette	0	1	0	1	0	0.11	1	0	0.11
Groupe	1	0	2	0	0.18	0	2	0.09	0
Equipe	0	1	0	2	0	0.22	2	0	0.11
Le	1	0	1	0	0.09	0	1	0.09	0
Est	1	1	1	1	0.09	0.11	2	0.045	0.055
De	1	0	1	0	0.09	0	1	0.09	0
La	1	0	1	0	0.09	0	1	0.09	0
Une	0	1	0	1	0	0.11	1	0	0.11
Mort	1	0	1	0	0.09	0	1	0.09	0
Favorite	0	1	0	1	0	0.11	1	0	0.11
Du	1	0	1	0	0.09	0	1	0.09	0
Au	0	1	0	1	0	0.11	1	0	0.11
Mondial	1	1	1	1	0.09	0.11	2	0.045	0.055
2010	1	1	1	1	0.09	0.11	2	0.045	0.055
Vocabulaire	***Vecteur binaire***		***Vecteur Fréquentiel***		***Vecteur Fréquentiel Normalisé***		***DF et Vecteur TF-IDF***		

Tableau 2.3 : Représentations vectorielles des documents de la figure 2.2

2.6.2- Le modèle probabiliste

La pondération des termes dans ce modèle est estimée par les probabilités d'apparitions des termes dans les textes. Ces techniques admettent l'indépendance entre les différents termes pour une simplification des calculs. Malgré que cette supposition soit peu réaliste, elle donne, néanmoins, des résultats intéressants. (Robertson & Sparck-Jones, 1976).
Dans l'approche du modèle probabiliste, le coefficient de similarité entre un document et les différentes classes du corpus est la probabilité que le document soit assigné à la classe.
Dans le modèle probabiliste, on considère que les documents sont générés par tirage aléatoire des différents termes qui les composent, les valeurs de probabilité de chaque tirage, sont estimées à partir des occurrences trouvées sur les documents du corpus. Cela revient en général à estimer la probabilité d'apparition du terme T sachant que le document appartient à la classe C.
Parmi les modèles probabilistes les plus utilisés le modèle Naïve Bayes, que nous verrons par la suite dans la section 3.2.6.

2.6.3- Représentation séquentielle

La représentation séquentielle d'un texte est une représentation sémantiquement plus riche et conceptuellement plus simple mais elle exige des modèles plus complexes comme les modèles de Markov Cachés. Le texte dans un tel codage, n'est pas représenté par un vecteur dans un espace donné, mais par une séquence de mots. Cette représentation est en fait une représentation naturelle d'un texte mais à cause de sa conception séquentielle qui nécessite le développement de modèles plus évolués, beaucoup moins d'applications se sont imposés par rapport aux représentations vectorielles.
De plus, il s'est avéré que, bien que cette représentation fût plus informative pour sa conservation de l'ordre des mots dans le texte, les classifieurs basés sur ce codage ne livraient

pas toujours de meilleurs résultats que des classifieurs basés sur des codages plus simples. Enfin, de par la nature non vectorielle de cette représentation, le stockage et l'indexation des textes restent une tâche beaucoup plus compliquée que dans les représentations précédentes. Notons néanmoins qu'il est possible à partir d'une représentation séquentielle de construire une représentation vectorielle tandis que l'inverse n'est pas vrai. (Denoyer, 2004)

2.7- Conclusion

Pour pouvoir appliquer les différents algorithmes d'apprentissage sur les documents de type textuels, un ensemble de techniques ont été développé pour montrer comment l'information textuelle est habituellement prise en compte pour la représentation « informatique » de ces documents. Les différentes approches de représentation informatique de textes sont exposées dans ce chapitre.

Ainsi avant la codification des documents, un ensemble d'opérations préliminaires doivent être faites pour épurer le texte de tous les mots inutiles et conserver seulement ceux qui sont porteurs d'informations et utiles pour le processus de classification. Mais malgré tous les prétraitements appliqués sur le document, l'espace des descripteurs, qui peuvent être des n-grammes, des stems, des phrases, des concepts ou tout simplement des mots, reste très grand et très creux, d'où la nécessité d'une diminution préalable de cet espace.

Plusieurs techniques de sélection des descripteurs ou réduction de dimensionnalité sont proposées dans la littérature, une bonne partie de ces approches sont étalées dans ce chapitre.

Une fois la liste des descripteurs arrêtés, un degré d'importance ou poids est attribué à tous les termes présents dans la représentation vectorielle ou probabiliste puisque chaque terme possède un certain nombre d'occurrences dans le document ou dans le corpus qui est différents aux autres.

Finalement on peut qualifier notre texte, par fichier « informatique » apte à être employé dans les différentes méthodes d'apprentissage automatique.