Université de Montréal

Apprentissage Automatique pour la détection de relations d'affaire

par
Grâce CAPO-CHICHI

Département d'Informatique et de Recherche Opérationnelle
Université de Montréal

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de Maîtrise
en Informatique

Avril 2012

Université de Montréal
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé :

Apprentissage Automatique pour la détection de relations d'affaire

Présenté par :
Grâce

a été évalué par un jury composé des personnes suivantes :

Philippe Langlais, président-rapporteur
Jian-Yun Nie, directeur de recherche
Pascal Vincent, membre du jury

Résumé

Les documents publiés par des entreprises, tels les communiqués de presse, contiennent une foule d'informations sur diverses activités des entreprises. C'est une source précieuse pour des analyses en intelligence d'affaire. Cependant, il est nécessaire de développer des outils pour permettre d'exploiter cette source automatiquement, étant donné son grand volume. Ce mémoire décrit un travail qui s'inscrit dans un volet d'intelligence d'affaire, à savoir la détection de relations d'affaire entre les entreprises décrites dans des communiqués de presse.

Dans ce mémoire, nous proposons une approche basée sur la classification. Les méthodes de classifications existantes ne nous permettent pas d'obtenir une performance satisfaisante. Ceci est notamment dû à deux problèmes : la représentation du texte par tous les mots, qui n'aide pas nécessairement à spécifier une relation d'affaire, et le déséquilibre entre les classes. Pour traiter le premier problème, nous proposons une approche de représentation basée sur des mots pivots c'est-à-dire les noms d'entreprises concernées, afin de mieux cerner des mots susceptibles de les décrire. Pour le deuxième problème, nous proposons une classification à deux étapes. Cette méthode s'avère plus appropriée que les méthodes traditionnelles de ré-échantillonnage.

Nous avons testé nos approches sur une collection de communiqués de presse dans le domaine automobile. Nos expérimentations montrent que les approches proposées peuvent améliorer la performance de classification. Notamment, la représentation du document basée sur les mots pivots nous permet de mieux centrer sur les mots utiles pour la détection de relations d'affaire. La classification en deux étapes apporte une solution efficace au problème de déséquilibre entre les classes.

Ce travail montre que la détection automatique des relations d'affaire est une tâche faisable. Le résultat de cette détection pourrait être utilisé dans une analyse d'intelligence d'affaire.

Mot clés : Relation d'affaire, classification supervisée, sélection de caractéristiques, déséquilibre de classes.

Abstract

Documents published by companies such as press releases, contain a wealth of information on various business activities. This is a valuable source for business intelligence analysis; but automatic tools are needed to exploit such large volume data. The work described in this thesis is part of a research project on business intelligence, namely we aim at the detection of business relationships between companies described in press releases.

In this thesis, we consider business relation detection as a problem of classification. However, the existing classification methods do not allow us to obtain a satisfactory performance. This is mainly due to two problems: the representation of text using all the content words, which do not necessarily a business relationship; and the imbalance between classes. To address the first problem, we propose representations based on words that are between or close to the names of companies involved (which we call pivot words) in order to focus on words having a higher chance to describe a relation. For the second problem, we propose a two-stage classification. This method is more effective than the traditional resampling methods.

We tested our approach on a collection of press releases in the automotive industry. Our experiments show that both proposed approaches can improve the classification performance. They perform much better than the traditional feature selection methods and the resampling method.

This work shows the feasibility of automatic detection of business relations. The result of this detection could be used in an analysis of business intelligence.

Keywords: business relation, supervised classification, feature selection, unbalanced data.

Table des matières

Liste des tableaux

Liste des figures

A mes Parents

Remerciements

Ce mémoire n'aurait pu être réalisé sans la précieuse collaboration et le constant soutien de plusieurs professeurs, collègues, parents et amis. Je remercie donc tous ceux et celles qui, durant les dernières années, ont contribué à la réussite de ce travail.

Je tiens plus particulièrement à remercier chaleureusement mon directeur de recherche, Monsieur Jian-Yun Nie, Professeur-chercheur à l'université de Montréal dans le domaine de la recherche d'information, pour son encadrement, son aide, ses conseils et encouragements tout au long de ce travail. Il a su me transmettre son insatiable curiosité, sa passion pour le travail intellectuel rigoureux, ainsi que pour la recherche universitaire.

Je souhaite également remercier Massih-Reza Amini chercheur au CRTL, pour m'avoir fait l'honneur d'accepter de m'aider dans l'élaboration des approches et d'avoir été disponible pour répondre à toutes questions et inquiétudes de ma part.

Je remercie également 'le comité de correction' pour avoir participé à l'évaluation de mon travail.

Je remercie aussi les membres passées et présents des équipes du laboratoire de recherche RALI de l'Université de Montréal pour leur accueil chaleureux. Merci en particulier à Lixin, Florian Boudin, Pierre-Paul Monty, Fabrizio Gotti, Fadhila Hadouche, Alessandro, Emmanuel, pour leurs discussions porteuses d'idées et tous leurs conseils.

Je remercie bien sûr ma famille et mes amis pour m'avoir soutenu pendant toute la durée de ce travail.

Chapitre 1

Introduction

1.1 Contexte

Dans le domaine d'intelligence d'affaire, les gens posent constamment des questions comme : Quels sont les clients de telles compagnies? Quelle entreprise est leader dans tel domaine? Quelles relations existe t-il entre telles ou telles autres entreprises ? L'entreprise qui peut répondre efficacement à ces questions aura indubitablement une longueur d'avance sur ses concurrents. Dans un environnement devenu de plus en plus complexe et compétitif, les réponses à ces questions deviennent primordiales pour l'élaboration d'une stratégie gagnante.

Les réponses à ces questions existent, mais souvent sous forme cachée dans des documents. Par exemple, les rapports annuels d'entreprise, les communiqués de presse, les articles de journaux nous renseignent tous sur différentes entreprises sur leurs états financiers, des activités, des changements importants du personnel, etc. Ces informations sont en effet des données de base utilisées dans des analyses d'affaire. Cependant, pour que ces informations brutes deviennent utiles, il est nécessaire de les traiter. Notamment, on doit extraire des informations pertinentes pour une analyse donnée. Cette tâche d'extraction est traditionnellement effectuée manuellement. Cette pratique est devenue de plus en plus difficile étant donné l'explosion des informations actuelles. Des outils automatiques doivent être développés pour aider à extraire des informations utiles.

Parmi les informations utilisées en analyse d'affaire, les relations existantes entre les entreprises ont une valeur importante. Ces relations peuvent souvent dévoiler des intentions des entreprises à poursuivre une direction dans leurs développements futurs. Elles peuvent aussi dresser un portrait d'un domaine : les leaders, tous les autres joueurs dans le domaine ainsi que les relations entre eux.

Le communiqué de presse est une forme de communication largement utilisée par des entreprises pour annoncer leurs activités et associations avec d'autres entreprises. Voici un exemple qui illustre un fragment de communiqué de presse :

« *WASHINGTON, July 12 /PRNewswire/ -- Ford Motor and a subsidiary of the Chrysler Corp. have won federal contracts to develop zero-pollution fuel cell engines, according to Fuel Cells 2000, a non-profit educational organization* »

Cet exemple montre que la compagnie *Ford Motor* et *Chrysler Corp* sont en relation d'affaire du type *partenariat* et ont un contrat de développement en commun. L'exemple suivant signale un lien de type *copropriétaire* entre les deux compagnies :

« Ford Motor Company has reached agreement to purchase the outstanding shares of the Hertz Corporation, giving Ford sole ownership of the world's largest car rental company. Ford has agreed to purchase the 26 percent share of Hertz owned by AB Volvo and an additional 20 percent from a Hertz management group ».

Notre travail décrit dans ce mémoire vise à identifier si un communiqué de presse décrit une relation d'un certain type entre deux entreprises données. Les types de relations sont prédéfinis par des chercheurs en analyse d'affaire. Nous pouvons alors considérer la détection des relations comme un problème de classification : les types de relations correspondent aux classes, et on tente de classer chaque document dans la classe correspondante (y comprise dans une classe correspondant à aucune relation d'affaire).

En effet la classification de texte consiste à chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de classes (étiquettes, classes). Cette liaison fonctionnelle, que l'on appelle également *modèle de prédiction*, est estimée par un apprentissage automatique. Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement étiquetés, dit *ensemble d'apprentissage*, à partir duquel nous estimons les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire le modèle qui produit le moins d'erreurs en prédiction.

Malgré de nombreux algorithmes de classification disponibles, quand on traite une application concrète, on doit résoudre plusieurs problèmes spécifiques de l'application. Par exemple, comme dans n'importe quelle tâche de classification de textes, il est nécessaire de représenter les documents et les classes à l'aide d'un même formalisme et celui généralement utilisé est un espace vectoriel formé par des mots [Sebastiani 2002].

Un problème souvent rencontré est que les mots contenus dans un texte ne décrivent pas très bien le contenu du texte. Par exemple, dans le premier exemple de communiqué de presse, la relation de *partenariat* n'est décrite qu'implicitement. Aucun mot (comme *partenaire*, *associé*, etc.) qui peut décrire cette relation explicitement n'existe dans le communiqué de presse. Par contre, plusieurs mots non pertinents pour la détection de la relation sont utilisés dans le texte. Une approche de base qui utilise tous les mots du texte comme une représentation n'est évidemment pas appropriée.

La tâche de classification peut s'avérer encore plus complexe et délicate lorsque les données disponibles pour l'apprentissage n'ont pas une distribution équilibrée entre les différentes classes. Dans ce cas les classificateurs auront du mal à bien fonctionner et certaines classes seront trop favorisées par rapport à d'autres, ce qui risquerait de biaiser les résultats. Dans le cas de détection de relations dans les communiqués de presse, ce problème est très prononcé. Nous devons donc trouver une solution à ce problème.

Dans la sous-section suivante, nous allons donner un aperçu rapide de notre application pour bien situer le travail.

1.2 Problématique

Face aux demandes du milieu d'affaire pour mieux connaître les entreprises et leurs relations, les chercheurs de HEC-Montréal ont initié un projet de recherche qui tente d'extraire les relations entre les entreprises à partir des communiqués de presse. Pour ce faire, ils ont collecté les communiqués de presses dans les domaines automobile et pharmaceutique entre (1994) et (2005). Le domaine automobile est utilisé pour une étude de faisabilité. Des préparations manuelles ont été faites : un ensemble d'entreprises

d'intérêts ont été identifiées, les relations pertinentes pour les analyses d'affaire subséquentes ont été définies. Le projet a commencé avec une approche basée sur des règles : un ensemble de règles portant sur les apparitions de certains mots importants, les agences de publication (certaines agences ne publient que certains types de communiqués de presse pouvant contenir ou non des relations d'affaire), sont définies pour faire une présélection des communiqués de presse susceptibles de contenir une relation d'affaire. Cependant, cette approche basée sur des règles manuelles s'est avérée rapidement limitée. Le nombre de communiqués de presse restant est encore trop important pour un examen manuel, et les règles n'étaient pas précises. Ainsi, le besoin d'identification des relations de façon automatique s'est fait sentir. C'est dans ce contexte que notre travail s'inscrit.

Notre tâche consiste à utiliser d'une part, les relations d'affaire que les experts ont définies (contract, client, owner, association…etc.), et d'autres parts les compagnies déjà identifiées manuellement comme Chrysler, Volvo, General Motors, Ford … etc. pour déterminer si un communiqué de presse présente ou non une relation d'affaire spécifique.

Dans cette étude de faisabilité, les chercheurs de HEC-Montréal ont analysé tous les communiqués de presse dans le domaine automobile. Nous pouvons donc considérer le problème comme un problème de classification, en utilisant une partie des communiqués de presse analysés comme des exemples d'entraînement et une autre partie pour le test.

Une bonne classification ne peut se faire sans avoir trouvé un meilleur ensemble de caractéristiques (features) discriminatoires servant pour représenter efficacement les documents. Dans la littérature de l'indexation et classification de textes, on utilise généralement des mots pondérés formant un vecteur pour représenter le contenu d'un texte. Cependant, une méthode d'indexation de base utilise tous les mots présents (à part les mots outils) et donc est susceptible d'utiliser des mots non significatifs (bruits) au même titre que des mots significatifs. Ainsi le premier problème auquel nous nous sommes confrontés est de trouver une meilleure méthode de sélection des caractéristiques pouvant mieux représenter le contenu du texte en vue de l'identification des relations d'affaires. Pour cela nous allons nous appuyer sur la structure de communiqué de presse, notamment les

positions auxquelles les noms d'entreprises apparaissent dans le texte. Notre idée de base est de s'appuyer plus fortement sur les mots de contexte proche des noms d'entreprises. Cette solution s'avère plus efficace que les méthodes de ré-échantillonnage.

Le deuxième problème auquel nous nous confrontons est le déséquilibre entre les classes, c'est-à-dire que les communiqués de presse dont nous disposons sont répartis inégalement entre les différentes classes. En effet, pour certaines classes, il existe beaucoup plus de communiqués de presse que pour d'autres. Notamment, si on considère les textes dans lesquels aucun lien d'affaire n'est décrit (*no-link*), les instances dans cette classe sont nettement plus nombreuses que les autres classes présentant un lien d'affaire. Si aucun traitement n'est fait par cela, les classificateurs risquent de présenter un biais vers la classe majoritaire au détriment de la précision dans les classes minoritaires. Ce problème est souvent observé par la communauté scientifique dès lors qu'on se retrouve face à des données réelles. Une approche pour tenter le résoudre en adoptant des techniques de ré-échantillonnages. Cette solution s'avère limitée dans notre contexte d'application.

Ainsi le deuxième problème de recherche abordé dans ce mémoire est le traitement du déséquilibre accru entre les classes en apprentissage supervisé. Nous proposons une approche en deux étapes : d'abord on va séparer les classes contenant une relation de la classe qui ne contient pas de relation ; ensuite on va séparer les classes de relations. Nos expérimentations montrent que cette approche est plus adaptée que les méthodes de ré-échantillonnage traditionnelles.

Dans nos expérimentations, nous comparons notre méthode de sélection de caractéristiques et de traitement du déséquilibre de classes, avec les méthodes classiques. Nous allons constater que nos méthodes s'avèrent généralement plus performantes que les méthodes de base existantes. Ceci montre qu'il est nécessaire d'adapter les méthodes existantes à notre problème afin d'arriver à une performance élevée.

1.3 Contributions

Cette étude vise à tester la faisabilité de l'identification des relations d'affaire à partir des documents dans le domaine d'affaire, notamment, les communiqués de presse. Nos expérimentations montreront que le problème de détection de liens d'affaire est faisable avec des algorithmes de classification existante. Toutefois il est nécessaire d'effectuer certains traitements adaptés pour la sélection de caractéristiques et le déséquilibre entre les classes. Ce sont ces deux problèmes que nous traitons en particulier pour tenter d'améliorer la performance de classification par rapport à une approche de base.

Pour le premier problème de sélection des caractéristiques, nous proposons une représentation des documents tenant compte de la position des mots relatifs aux mots pivots, c'est-à-dire les noms des compagnies qui nous intéressent. Cette nouvelle représentation apporte une amélioration du résultat en comparaison avec une représentation vectorielle de base.

Pour le second problème qui concerne le déséquilibre des classes, nous proposons une classification à deux niveaux. Cette approche est différente de celle de rééquilibrage typiquement utilisée pour résoudre ce problème dans la communauté scientifique et elle s'avère plus adaptée à notre problème.

1.4 Organisation du mémoire

La suite du mémoire se structure comme suit : En premier lieu, le chapitre 2 présentera un état de l'art sur l'apprentissage supervisé et la classification. Le problème de sélection de caractéristiques et le problème de déséquilibre entre les classes seront exposés.

Au chapitre 3, nous présentons de façon détaillée les approches que nous proposons pour la détection automatique des liens d'affaire à partir des communiqués de presse. Dans ce chapitre, nous décrivons les différentes approches que nous avons utilisées, ainsi que les résultats obtenus.

Le dernier chapitre proposera une synthèse critique des travaux présentés dans ce mémoire et exposera les pistes de recherche pouvant être explorées dans la continuité de nos travaux.

Chapitre 2

Classification supervisée de textes : État de l'art

L'objet d'étude de ce mémoire est la détection automatique de relations d'affaire en utilisant une approche d'apprentissage supervisé. Ainsi, nous allons faire un tour d'horizon sur certains travaux auxquels s'apparente ce dernier thème. Il y a eu plusieurs travaux portant sur ce sujet depuis au moins le début des années 1960. Malgré le fait que des avancées importantes ont été observées depuis, la recherche dans ce domaine est toujours très nécessaire. Les résultats obtenus aujourd'hui peuvent encore être améliorés et il y a toujours de nouveaux types de données à traiter. Les classificateurs automatiques performent presque aussi bien que les humains pour certaines tâches, mais pour d'autres, l'écart est encore grand. Ceci sans compter que le besoin d'un traitement efficace de l'information est grandissant, car l'immense bassin de données à notre portée est sans intérêt s'il n'est pas bien géré.

Au premier abord, le problème de classification est facile à décrire. D'un côté, on est en présence d'une banque de documents textuels et de l'autre, d'un ensemble prédéfini de classes. L'objectif est de déterminer de façon automatique dans quelle classe classer chacun des textes, à partir de leur contenu ou d'autres indicateurs, tel qu'illustré à la figure 1. La solution à ce problème n'est pas simple et plusieurs facteurs sont à prendre en considération, tels que la sélection des caractéristiques utilisées pour décrire les données et le traitement des données (classes) déséquilibrées.

Ce chapitre vise à définir plus en profondeur le problème de classification et les solutions suggérées dans la littérature.

- En premier lieu, ce chapitre précisera le processus de la tâche de classification, les principales difficultés qu'elle engendre et les principales façons d'aborder le problème de déséquilibre des classes.
- Le deuxième point qui sera présenté est le choix d'un mode de représentation adéquat des instances à traiter, en l'occurrence les textes. Il s'agit d'une étape incontournable en apprentissage automatique : on doit opter pour une façon judicieuse de représenter les données avant de les soumettre à un algorithme. Nous décrivons les méthodes de représentation couramment utilisées dans ce domaine.
- Par la suite, il sera question de la méthode de sélection et d'extraction des caractéristiques (ou des features), presque toujours impliquée en classification de textes et par laquelle on élimine les caractéristiques jugés inutiles à la classification.

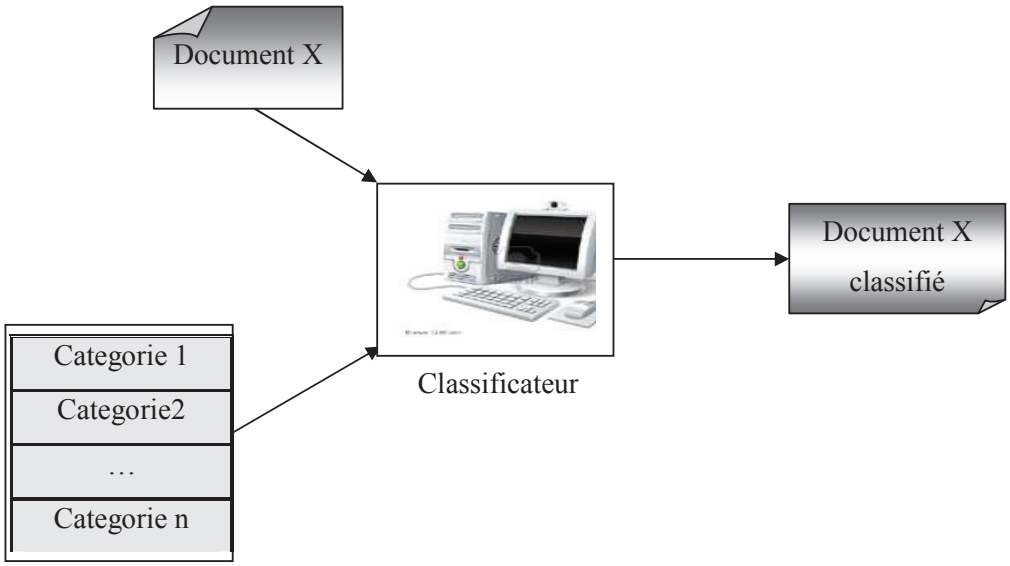

Figure 1 : Illustration de la tâche de classification

2.1 La Tâche de classification de textes

L'on distingue généralement deux façons d'aborder le problème de la classification automatique [Sebastiani 1999].

1- Jusqu'à la fin des années 1980, l'approche dominante pour le résoudre s'inscrivait dans une optique d'ingénierie des connaissances (*«knowledge engineering»*). On construisait un système expert comportant un ensemble de règles définies manuellement, par des experts du domaine, et qui ensuite pouvait procéder automatiquement à la classification. Ces règles prenaient généralement la forme d'une implication logique où l'antécédent portait, typiquement, sur la présence ou l'absence de certains mots, et où le conséquent désignait la classe d'appartenance du texte. Par exemple : Si 'fusion' est présent, alors classer dans la classe 'copropriétaire'.

Cette approche peut s'avérer très efficace mais l'inconvénient est que l'édition des règles de décision peut s'avérer très longue. Dans beaucoup de cas, il est difficile d'établir de telles règles avec une grande précision. Et surtout, si des classes s'ajoutent ou si on désire utiliser le classificateur dans un autre domaine, on doit répéter l'exercice. C'est donc la pertinence de cet ensemble de règles, qui évolue dans le temps, qui mine l'intérêt envers cette façon de faire. Dans le projet dans lequel s'inscrit cette étude, les chercheurs de HEC-Montréal ont commencé avec cette approche. C'est justement en observant la difficulté d'enrichir la base de règles qu'ils ont abandonné l'idée, et tournent vers une approche automatique.

2- Avec le développement de techniques d'apprentissage automatique (*«machine learning»*), on voit le problème sous un autre angle. En fait, même en 1961, Maron [Maron 1961] avait proposé un classificateur bayésien qui se distinguait de l'édition de règles en se basant plutôt sur un calcul de probabilités. Cependant, ce n'est pas avant le début des années 1990 que cette approche a pris son envol.

En effet avec des techniques d'apprentissage machine, l'on essaie d'apprendre à la machine à généraliser, en lui soumettant un ensemble d'exemples déjà associés à leurs classes préalablement établies par un humain. Cet ensemble est appelé *Ensemble*

d'apprentissage ou *Ensemble d'entrainement*. Ainsi lorsqu'on présente un nouveau texte à la machine, elle serait capable de le classer en se basant sur ce qu'elle a appris avec l'ensemble d'apprentissage.

La figure 2 présente une vue schématique du processus général de classification. Les différentes étapes et méthodes utilisées pour mener à bien la tâche de classification seront expliquées plus en détail dans les sections qui suivent.

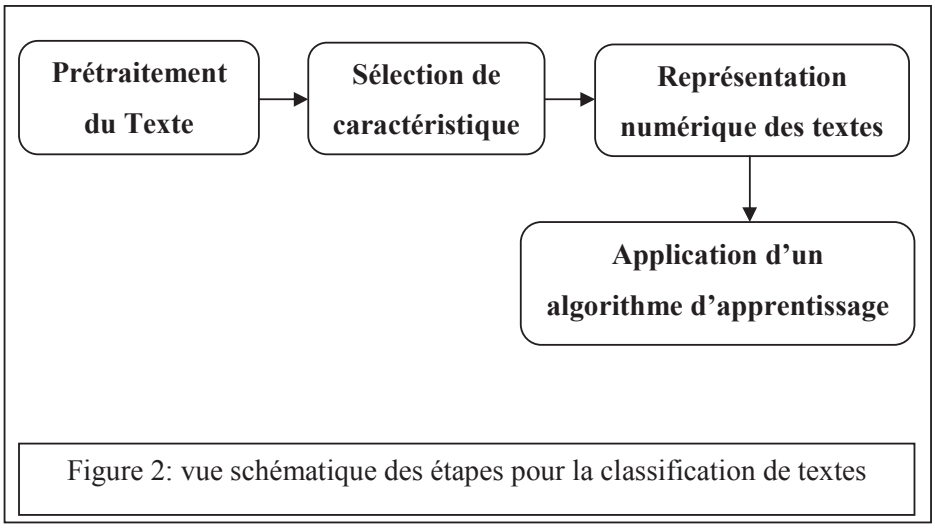

Figure 2: vue schématique des étapes pour la classification de textes

2.1.1 Les textes vus comme des unités mesurables

Dans le processus de classification, il existe une étape fondamentale qui consiste à représenter les textes de façon manipulable par un classificateur. En général, l'on considère les textes comme des sacs de mots. Ils sont ensuite projetés directement dans un espace vectoriel des mots que l'on utilise comme des caractéristiques. L'approche utilisée couramment par la communauté est celle basée sur le modèle vectoriel de Salton "Vector Space Model" [Salton et al. 1975]. Dans ce modèle, un corpus est représenté par une matrice dans laquelle les lignes sont relatives aux caractéristiques et les colonnes aux documents. Une cellule d'une telle matrice comptabilise la fréquence d'apparition d'une caractéristique dans un document ou une mesure dérivée de ceci.

Différentes représentations des documents existent du point de vue de la comptabilisation de la fréquence d'apparition d'une caractéristique.

La représentation binaire considère qu'un document est représenté par un vecteur dans un espace dont les composantes informent sur la présence (valeur égale à 1) ou l'absence (valeur égale à 0) d'un terme dans un texte. C'est une représentation très simple, ce qui est un avantage important pour les systèmes nécessitant un temps de calcul très faible. Cette représentation simple est encore largement utilisée dans le domaine de classification de textes car elle présente un bon compromis entre complexité et performance des systèmes. Néanmoins cette représentation est peu informative car elle ne renseigne ni sur la fréquence d'un mot qui peut constituer une information importante pour la classification, ni sur la longueur des documents.

La représentation fréquentielle est une extension naturelle de la représentation binaire qui prend en compte le nombre d'apparitions d'un mot dans un document. Ainsi, un document est représenté dans un espace où chaque composante correspond au nombre d'apparitions du terme correspondant dans le document.

Une autre représentation qui est couramment utilisé par la communauté est celle basée sur la pondération TF-IDF, tel que décrite dans la section suivante.

2.1.2 Représentation TF-IDF

La représentation présentée ici est aussi une représentation vectorielle qui tente d'être plus informative que les représentations précédentes. La pondération TFIDF a démontré une bonne efficacité dans des tâches de classification de textes, et en plus son calcul est simple [Sebastiani 1999]. Les termes sont pondérés selon un modèle de pondération qui tient généralement compte de deux facteurs, local et global. La composante TF (term frequency) reflète l'importance du terme dans le document. Elle est souvent mesurée par la fréquence brute du terme dans le document ou par son log. La composante IDF (Inverse Document Frequency weighting) consiste à attribuer les plus grands poids aux termes d'index apparaissant dans un faible nombre de documents. Ces derniers termes étant considérés comme portant le plus fort pouvoir de discrimination et correspondent à une partie d'information spécifique au document. La mesure IDF est définie comme suit:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

où t est un terme d'index, N est le nombre total de documents dans la collection, et df(t) est le nombre de documents qui contiennent le terme d'index t. Le poids du terme d'index t dans un document d est alors défini comme suit dans la pondération TFI-DF:

$$W(d,t) = TF(d,t).IDF(t)$$

où $TF(d,t) = \log(f(t,d))$, et f(t, d) représente la fréquence d'apparition du terme d'index t dans le document d.

Ainsi, un terme qui a une valeur de TFI-DF élevée doit être à la fois important dans le document auquel ce terme est associé, et doit apparaitre peu souvent dans les autres documents. Cette pondération sera appliquée dans la suite de nos travaux comme une pondération de base étant donné sa performance prouvée dans la RI et classification de textes. Cependant, nous allons développer des mesures supplémentaires pour répondre plus efficacement à notre tâche.

2.1.3 Prétraitements des données

Un mot dans un document peut avoir une signification spécifique ou non par rapport au contenu du document. Par exemple, un terme de spécialisation correspond à une partie du contenu du document, mais un terme générique du langage (e.g. « de » en français) ne décrit pas le contenu. Ainsi, nous ne devons pas retenir tous les mots présents dans un document comme des caractéristiques utiles dans la représentation d'un document. Afin d'éliminer les mots vides de sens, on procède à un filtrage en utilisant un anti-dictionnaire ou stopliste : Une stopliste contient tous les mots outils vides de sens que l'on ne garde pas comme index. Il existe des stoplistes standard (souvent utilisées) dans plusieurs langues. Une telle liste contient généralement quelques centaines à quelques milliers de mots outils, selon la langue traitée.

En plus des mots fonctionnels dans une langue, certains mots très fréquents dans une collection de documents peuvent être également ajoutés dans la stopliste [Stricker 2000]. Ces mots sont considérés peu informatifs pour la collection de textes donnée.

En plus de filtrer les mots fréquents par une stopliste, on peut aussi filtrer les mots rares, en considérant que les mots rares ne représentent pas une sémantique importante. Pour cela, on peut établir un seuil, et seuls les mots dont la fréquence atteint le seuil sont gardés dans l'index.

De la même façon, les ponctuations et les symboles tels que $, #, *, etc. peuvent aussi être supprimés.

Rappelons que le but ultime d'une représentation est de représenter adéquatement le contenu sémantique du texte. Or, un mot (ou plutôt une forme de mot) n'est souvent qu'une forme qui peut représenter un concept. D'autres mots ou d'autres formes de mot peuvent souvent représenter le même concept. Cependant, il est difficile de procéder à une analyse sémantique pour reconnaître le concept représenté par un mot. Une façon simple et robuste consiste à transformer une forme de mot en une représentation standard selon la morphologie. L'hypothèse est que les formes similaires, qui sont différentes seulement en

leurs suffixes, correspondent généralement au même concept. Par exemple, « information », « informations », « informationnel », voire « informatique » sont fortement relié dans leurs significations, et ces mots peuvent être représentés par le même index (ou caractéristique). Ainsi, nous pouvons utiliser le processus de troncature (*stemming*) qui supprime les suffixes de mots pour transformer une forme de mot en sa représentation standard. Une autre transformation possible est la lemmatisation, qui transforme un mot en sa forme standard (e.g. nom dans sa forme singulier, verbe à l'infinitif, etc.).

Dans le domaine de RI, il existe quelques algorithmes de *stemming* standard : dont celui de Porter et celui de Krovetz. Ces algorithmes sont largement utilisés dans les expérimentations en RI et classification de texte, et les résultats montrent que le processus de *stemming* peut non seulement réduire le nombre d'index (caractéristiques) utilisés, donc la taille de l'espace vectoriel, mais aussi augmenter la performance de recherche et de classification. Nous allons utiliser ces différentes techniques de réduction de vocabulaire dans les étapes préliminaires de notre étude.

La suppression de ces classes de mots et la standardisation des formes de mot s'avèrent utiles mais insuffisantes car le nombre de termes reste encore très élevé. Il est donc nécessaire d'utiliser une méthode statistique pour déterminer les mots utiles pour la discrimination entre les documents. Cela nous amène à décrire les méthodes de sélection de caractéristiques utilisées pour y remédier.

2.2 Méthode de Sélection des caractéristiques en classification

La sélection de caractéristiques consiste à choisir un sous-ensemble minimum de P caractéristiques à partir d'un ensemble original en contenant Q ($P < Q$), de sorte que l'espace de caractéristiques soit réduit de façon optimale selon certains critères d'évaluation [Liu et *al*, 2005]. L'objectif principal est la réduction du nombre de caractéristiques

utilisées tout en essayant de maintenir ou d'améliorer les performances de classification du système.

La sélection de caractéristiques a été largement étudiée dans plusieurs domaines comme la bioinformatique, la classification de texte, le data mining, le traitement d'images etc. [Dash et *al*, 1997] [Jensen, 2005]. Dans [Mendez et *al*, 2007], les auteurs présentent une analyse détaillée montrant l'influence du changement de la dimension de la représentation d'un message sur certaines techniques classiques de filtrage de spams.

La nécessité d'avoir une bonne sélection de caractéristique est une chose indubitable, en ce sens qu'une bonne classification dépend largement des attributs utilisés pour représenter les documents. Il a été souligné dans la littérature qu'une caractéristique non pertinente n'apporte aucune information permettant de discriminer les classes entre elles. De même qu'une caractéristique bruitée porte des informations incorrectes et pouvant fausser la classification. Des problèmes de corrélations entre des caractéristiques peuvent être identifiés également si cette sélection n'a pu être faite avec une bonne stratégie. Dans ce cas-ci une même information sera redondante et cela impliquera une saisie des données sur de nombreuses caractéristiques dont le coût en termes de temps de calcul sera très élevé. Il est donc nécessaire de déterminer quelles sont les caractéristiques indispensables pour classifier les données. C'est pourquoi certains critères de sélection de caractéristique ont été mis en place par des chercheurs.

Parmi les méthodes les plus souvent utilisées citons :

- le calcul de l'information mutuelle [Lewis, 1992] [Mouliner, 1997] [Dumais *et al.*, 1998];
- la méthode du chi-2 [Schütze *et al.*, 1995] [Wiener *et al.*, 1995] ou des méthodes plus simples utilisant uniquement les fréquences d'apparitions [Wiener, 1993] [Yang et Pedersen, 1997];
- le gain d'information.

Nous décrivons brièvement ces méthodes de sélections dans les sous sections suivantes.

2.2.1 Information mutuelle

Cette méthode est basée sur le nombre de fois qu'un mot apparaît dans une certaine catégorie. En effet plus un mot va apparaître dans une catégorie, plus l'information mutuelle entre ce mot et cette catégorie sera jugée élevée, et dans le cas contraire celle-ci va être jugée moins élevée. Il s'agit d'un des critères très souvent utilisés. Une forme de calcul d'information mutuelle (pointwise mutual information) souvent utilisée est comme suit:

$$IM\ (w, clas) = \frac{P(w, clas)}{P(w) * P(clas)}$$

où P(*w*, *clas*) est la probabilité jointe de la classe *clas* et de la caractéristique *w* (qui est estimée par la proportion des documents dans la collection qui appartiennent à la classe et qui contiennent le mot) , P(*w*) est la probabilité de *w* (le DF document frequency relative de la caractéristique *w*) et P(*clas*) est la probabilité de la classe (la proportion de documents de classe *clas*).

La faiblesse de cette mesure est qu'elle est beaucoup trop influencée par la fréquence des mots. Pour une même probabilité conditionnelle sachant la catégorie, un terme rare va être avantagé, car il risque moins d'apparaître en dehors de la catégorie.

2.2.2 Le gain d'information (GI)

Le gain d'information est utilisé pour déterminer en quelque sorte le pouvoir discriminatoire d'un mot. Elle mesure la diminution de l'entropie, ou la quantité de connaissance gagnée si une caractéristique est présente ou absente. Mathématiquement le gain d'information (GI) pour une caractéristique *w* est défini par :

$$GI(w) = \sum_{X\epsilon\{w,\bar{w}\}} \sum_{clas\epsilon\{c_i\}} P(X, clas) log\left(\frac{P(X, clas)}{P(X) * P(clas)}\right)$$

Où P(X, $clas$) est la proportion de documents dans la collection appartenant à la classe $clas$ et où la caractéristique X est présente;

P(X) est le DF (Document-frequency) relative qui représente le nombre de document dans lequel la caractéristique X apparaît par rapport au nombre total de documents.

P($clas$) est la proportion de documents de classe $clas$ dans le corpus.

2.2.3 Méthode du chi2 (X^2)

Il s'agit d'une mesure statistique bien connue, qui s'adapte bien à la sélection de caractéristiques. Elle évalue le manque d'indépendance entre un mot et une classe. Elle utilise les mêmes notions de cooccurrence mot/catégorie que l'information mutuelle. Une différence importante est qu'elle est soumise à une normalisation, qui rend plus comparable les termes entre eux. Elle perd quand même de la pertinence pour les termes peu fréquents.

Le calcul de Chi2 nécessite de construire le tableau de contingence (2×2) pour chaque caractéristique w du corpus et pour chaque classe (voir tableau 1).

	Caractéristique w présente	Caractéristique w absente	
Classe présente	a	c	a+c
Classe absente	b	d	b+d
	a+b	c+d	N= a+b+c+d

Tableau 1 : Tableau de contingence pour l'absence ou la présence d'une caractéristique dans les classes.

La statistique du X^2 peut se calculer sous la forme :

$$X^2\,(w, clas) = \frac{N(ad - cb)^2}{(a + c)(b + d)(a + b)(c + d)}$$

Quelle que soit la mesure de sélection de caractéristiques utilisée, il est toujours nécessaire de définir un seuil à partir duquel l'on juge qu'un mot est gardé ou est supprimé, ou définir un nombre de caractéristiques à garder (les premières caractéristiques les plus significatives sont à garder). Le choix de cette valeur du seuil ou du nombre doit être fait judicieusement car elle est aussi très déterminante pour la performance de la classification.

Une étude comparative de ces méthodes de sélection de caractéristique est effectuée dans [Yang et Pedersen, 1997]. Il semble en résulter que l'information mutuelle est légèrement supérieure aux autres.

Il s'agit de critères de sélection heuristiques. De meilleurs résultats sont sans doute possibles avec des techniques de sélection avant ou arrière (forward selection or backword elimination) en mettant le classificateur dans une boucle. Mais c'est coûteux. Il y a aussi le LASSO (least absolute shrinkage and selection operator).

Précisons qu'il existe une autre famille de techniques alternatives à la sélection de caractéristiques, qui consistent à extraire des caractéristiques, c'est-à-dire de définir de nouvelles caractéristiques synthétisées à partir des caractéristiques de base.

Pour bien classifier des textes, on pourrait croire qu'il n'est pas nécessaire d'avoir un grand nombre de caractéristiques pour obtenir de bons résultats. Cependant, ceci n'est pas toujours le cas. Par exemple, certains chercheurs [Dumais et *al*., 1998] ont démontré que les machines à vecteurs de supports (SVM) qui supportent de grandes dimensions n'ont pas nécessairement besoin de beaucoup de caractéristiques pour donner de bons résultats. Même avec un petit nombre ils répondent bien. Ceci nous fournit une justification à écarter des caractéristiques qui n'introduisent que du bruit dans la représentation. Dans notre travail, comme dans beaucoup d'autres travaux sur la classification, nous allons donc utiliser un sous-ensemble de caractéristiques les plus significatives. En plus des techniques de sélection de caractéristiques standard, nous allons aussi tester différentes approches spécifiques à nos documents et notre tâche.

Une fois que l'on a effectué les différents prétraitements sur le corpus et représenter les documents, l'on utilise des algorithmes d'apprentissage pour effectuer la classification, quelques uns de ces algorithmes seront présentés dans la suite de notre travail. Mais avant cela nous faisons un tour d'horizon sur l'extraction de relation dans le domaine médical et entre entités nommées qui est une tâche qui s'apparente à notre étude.

2.3 Travaux similaires : Extraction de relations dans le domaine médical et entre entités nommées

L'extraction de relation tente en général d'identifier diverses relations dans un ensemble de documents. Dans le domaine médical, il peut s'agir de trouver diverses interactions entre gênes, de trouver la relation existante entre une maladie et un médicament. Un exemple de question typique à laquelle on cherche à répondre est : «est-ce-que tel médicament guérit telle maladie?»; ou au contraire «donne des effets indésirable?».

On peut aussi s'intéresser à l'extraction de relations entre entités nommées. Une entité nommée peut être par exemple une personne, une date, une organisation… Dans ce contexte, on peut tenter de trouver une relation entre une personne et une organisation, à répondre la question : «est-ce-que telle personne travaille au sein de telle organisation ? » ou une relation existante entre une personne et une date : «telle personne est née à telle date ?».

De façon générale, ces diverses extractions de relations sont pour la création ou l'enrichissement des bases de connaissances ou des ontologies. Ces techniques sont beaucoup utilisées dans le domaine du Traitement Automatique de la Langue et servaient également dans l'accomplissement des systèmes de questions-réponses. Vue l'importance de l'extraction de relation, plusieurs études ont été faites dans différents cadre de travail.

Dans le domaine médical, [Rindflesch et *al*, 2000] et [Srinivasan et Rindflesch, 2002] ont conçu le système SemRep pour extraire des relations de branchement artériel

dans des comptes-rendus opératoires et des relations entre des problèmes médicaux et leurs traitements. De même [Friedman et al. 2001] ont élaborés le système MedLEE qui permet l'extraction des relations de compte-rendu radiologique, et [Chen et Friedman 2004] pour la mise en place d'un système d'extraction des interactions biomoléculaires et des relations gène-phénotype. Dans ce cas les approches sont basées sur une définition manuelle des patrons d'extraction. En effet, si l'on se place dans un système de question-réponse, pour répondre à la question « Quand a été créé le vaccin contre la rage nommé 'Anti-Rage'? », il est nécessaire de retrouver la relation entre 'Anti-Rage'et sa date de création dans les textes. Or, cette relation peut être exprimée de différentes manières :

- «Anti-Rage, créé le 10 mai 1958 à Neuilly-sur-Seine» ;
- «Anti-Rage est créé le 10 mai 1978 à Neuilly-sur-Seine» ;
- «Anti-Rage est créé en mai 1978» ;
- «Anti-Rage. 10 mai 1958 création à Neuilly-sur-Seine».

Afin de retrouver sa date de création, il faut pouvoir généraliser la façon dont cette relation est exprimée dans les textes. Ainsi, à partir des deux premières phrases, on peut déduire le patron d'extraction suivant :

<vaccin> (,|est) créé le <date>.

Créer ces patrons manuellement est fastidieux et peu facilement généralisable.

D'autres approches utilisent l'apprentissage automatique. Par exemple [Roberts et *al*. 2008] proposent une approche classique fondée sur des SVM pour extraire des relations dans le corpus du projet CLEF (the Clinical E-Science Framework project). [Zhou et *al*, 2005] et [Uzuner et *al*. 2010] utilisent les dépendances syntaxiques entre les concepts sous forme d'attributs dans une approche vectorielle basée sur des SVM pour identifier des relations entre des personnes, des organisations, et des lieux.

Plusieurs autres travaux similaires ont été faits tels que la conception d'un système de question-réponse [Iftene et *al*. 2008], l'extraction d'information [Banko et *al*. 2007] ou l'extraction de réseaux sociaux [Matsuo 2006].

Notre tâche pourrait être vue sous cet angle d'extraction de relation étant donné que nous essayons de détecter les différentes relations d'affaire liant des compagnies. Si l'on considère les compagnies comme des entités nommées, nous pourrions évidemment utiliser ces différentes méthodes existantes pour notre tâche.

2.4 Algorithmes d'apprentissage pour la classification

Plusieurs algorithmes de classification ont été développés dans le domaine de l'apprentissage automatique. Nous présentons ici quelques unes de ces approches.

2.4.1 Méthodes Bayésiennes naïves

La méthode bayésienne naïve tente de déterminer la probabilité $P(C|T)$ de classer le texte T dans la classe C. Elle utilise le théorème de Bayes suivant:

$$P(C|T) = \frac{P(T|C) * P(C)}{P(T)}$$

où $P(C|T)$ est la probabilité de la classe C sachant le texte T ; $P(T|C)$ la probabilité du texte T sachant la classe C ; $P(C)$ et $P(T)$ sont les probabilités à priori de la classe C et du texte T. $P(T)$ est une constante, ainsi :

$$P(C|T) \propto P(T|C) * P(C)$$

Dans le cadre de la classification de textes, les classificateurs bayésiens naïfs cherchent à prédire la valeur d'un nouvel objet (document, texte, évènement) à partir d'une estimation de probabilités prenant en compte des connaissances ou observations existantes.

Ainsi, pour $P(T|C)$, on suppose que les termes présents dans le textes T sont indépendants (c'est là l'hypothèse naïve) et on estime cette probabilité par la formule suivante.

$$P(T|C) = \prod_{t_i \in T} P(t_i|C)$$

Si on suppose que T est représenté par un vecteur binaire.
Dans cette formule,

- $P(t_i|C)$ est estimée selon la fréquence de présence de t_i dans les documents de la classe C.
- $P(C)$ peut être estimée par la proportion des documents de la classe C dans la collection.

Pour réellement déterminer à quelle classe un document appartient, il faut calculer $P(C_i|T)$ pour chacune des classes. Étant donné que P(T) reste constant pour toutes les classes, déterminer $P(C_i|T)$ se résume juste au calcul de $P(T|C_i) * P(C_i)$.

2.4.2 La méthode des k-plus-proches voisins

La méthode des k-plus proches voisins (k-ppv) connue en anglais sous l'appellation (k-nearest neighbour ou k-NN) est l'une des techniques d'apprentissage supervisé les plus simple à mettre en œuvre et qui a vu son origine avec [Fix et Hodges 1951]. Il diffère des autres approches de classification en ce sens qu'il ne nécessite pas en tant que telle une phase d'entrainement. Les données sont simplement emmagasinées en mémoire. Pour prédire la classe d'un nouvel objet, on le compare à ses voisins les plus proches par une mesure de similarité. Il est recommandé de ne pas considérer le voisin le plus proche de l'objet à classer, mais ses k-plus proches voisins afin de minimiser les risques d'erreurs.

D'autres versions améliorées de l'algorithme essayent de pondérer les voisins par la distance qui les sépare du nouveau texte et on accorde un grand poids aux documents similaires lors de la prise de décision. Cela se formule de la façon suivante :

Soit D $= \{(x_i, y_i)\}_{i=1}^{n}$ l'ensemble d'entrainement où $y_i \in Y = (0, 1, \ldots C-1)$ est l'identité de la classe cible de l'entrée x_i, C est le nombre de classes, d(.,.) est une fonction de distance , x est une entrée de test et $V(x, D, d(.,.), k)$ l'ensemble des k plus proches voisins de x parmi les entrées de n ainsi que leur cible associée.

La prédiction de classification par l'algorithme des k plus proches voisins est donc:

$$argmax_{(y \in Y)} \sum_{(x_i, y_i) \in V(x, D, d(.,.), k)} I_{y_i = y}$$

où $I_{y_i=y}$ est la fonction indicateur. La méthode de k-ppv utilise une mesure de similarité entre les documents pour déterminer les plus proches voisins. Celles fréquemment utilisée sont la distance euclidienne :

$$d(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

ou la similarité cosinus :

$$d(a, b) = 1 - \cos(a, b) = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2}}$$

Avec *n* le nombre de features.

La méthode des *k*-ppv a l'avantage d'être très simple à mettre en œuvre et d'utiliser directement l'ensemble d'apprentissage. Elle ne fait aucune hypothèse a priori sur les données et ne tente pas de créer un modèle pour représenter les classes. La qualité de la discrimination par cette méthode dépend du choix du nombre *k* de voisins considérés et de

la façon de calculer la distance entre les instances. Il est cependant souvent nécessaire de faire varier ce nombre *k* pour obtenir les meilleurs résultats possibles. Un autre problème important de la méthode des *k*-ppv est qu'elle nécessite un espace mémoire très important pour stocker les données et un temps assez long pour faire les différents calculs dans la phase de classification.

2.4.3 Machine à vecteurs de supports ou Séparateurs à Vastes Marges (SVM)

Cette technique fut introduite par Vapnik en 1995 [Vapnik 1995] et est une méthode de classification binaire au départ. Elle essaie de séparer les exemples positifs des exemples négatifs représentés par un vecteur de dimension *n* de façon linéaire. Cette méthode tente de trouver un hyperplan séparateur des exemples de l'ensemble de données avec une marge maximale (voir Figure 3).

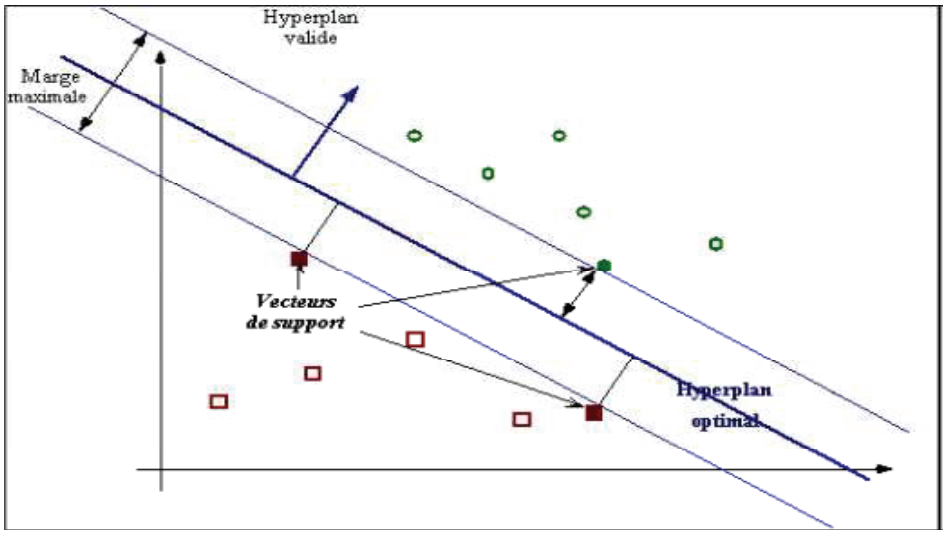

Figure 3 : Représentation schématique d'un SVM

Pour un ensemble d'échantillons D = $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ à deux classes $y_i = \{\pm 1\}$, le problème consiste à trouver un hyperplan tel que les données des étiquettes de classe +1 et -1 se trouvent de chaque côté de l'hyperplan et la distance des vecteurs les plus proches de l'hyperplan (pour chacune des deux classes) est maximale. Ces vecteurs les plus proches sont appelés vecteurs de support et la distance de ceux-ci par rapport à l'hyperplan constitue la marge.

D'une façon plus formelle, notre objectif est de trouver un hyperplan $wx + b = 0$, $w \in R \; et \; b \in R$, qui sépare les deux classes avec la plus grande marge. La recherche de la marge optimale permettant de déterminer les paramètres *w* et *b* de l'hyperplan conduit à un problème d'optimisation quadratique. On cherche un point qui minimise ou qui maximise une certaine fonction sujet à certaines contraintes.

2.4.3.1 SVM dans le cas de données linéairement séparables

Lorsque nous sommes en présence de données linéairement séparables (voir figure 4), c'est-à-dire qu'il existe un hyperplan $\in R^n$ qui sépare les exemples positifs et négatifs, l'hyperplan a pour équation $wx + b = 0$, et la distance d'un point à ce plan est calculée comme suit : $d(x) = \frac{|w\,x+b|}{\|w\|}$

L'hyperplan optimal étant celui pour lequel la distance aux marges (les points les plus proches) est maximale. Cela revient donc à minimiser $\|w\|$ sous certaine contrainte. La formulation dite primale des SVM s'exprime sous la forme suivante :

$$Min \; \frac{1}{2}\|w\|^2 \; sous \; la \; contrainte \; \forall \; i \; , y_i(wx_i + b) \geq 1$$

Cette équation peut être résolue entre autres par la méthode Lagrangienne d'optimisation quadratique avec contraintes pour maximiser la marge, on parle de formulation duale.

$$max \sum_{i-1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \, y_i y_j x_i x_j \; sous \; contrainte \; \sum_{i=1}^{n} \alpha_i y_i = 0 \; \forall \; i, \alpha_i \geq 0$$

Si on considère que α_i^* sont les solutions optimales de cette équation alors, l'hyperplan optimal est obtenu par l'équation suivante :

$$f(x) = \sum_{i=1}^{n} \alpha_i^* \, y_i (x . x_i) + b$$

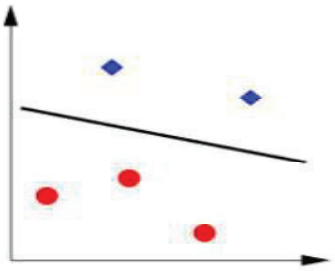

Figure 4 : SVM linéaire

2.4.3.2 SVM dans le cas de données non-linéairement séparables

Le problème des SVM tel qu'expliqué jusqu'à présent ne s'applique que lorsque l'on a des données linéairement séparables. Cependant il existe bien des cas où il est impossible de séparer entièrement les données avec un hyperplan. Lorsqu'il s'agit donc de données non-séparables linéairement (Figure 5), on projette le problème dans un espace de dimension plus grande dans lequel on espère trouver un séparateur linéaire.

On procède pour cela à une transformation des données d'un espace de départ vers un espace d'arrivée de manière qu'elle soit linéairement séparable dans l'espace d'arrivée.

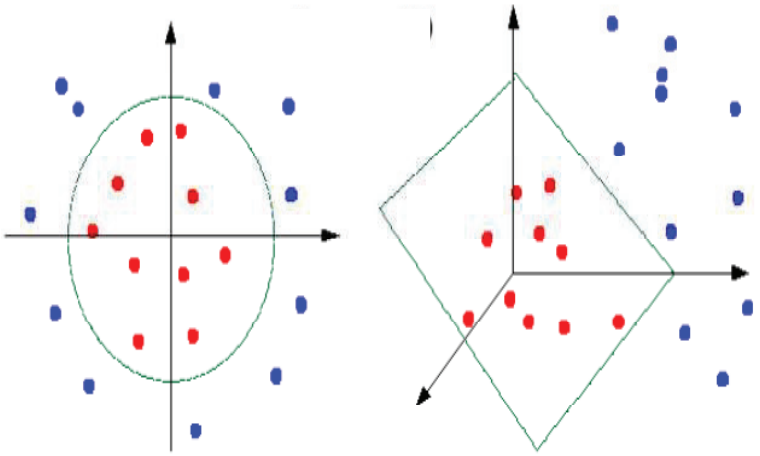

Figure 5 : SVM non linéaire

Soit la fonction de transformation φ et la transformation de la donnée *x*:

$$\varphi : \quad R^n \rightarrow R^r$$

$$x \rightarrow \varphi(x)$$

Pour trouver l'hyperplan séparateur, cela revient à résoudre le problème d'optimisation suivant :

$$max \sum_{i-1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \, y_i y_j \, \varphi(x_i)^T \varphi(x_j) \; sous \; contrainte \sum_{i=1}^{n} \alpha_i y_i = 0 \; \forall \; \alpha_i \geq 0$$

La résolution de cette équation nécessite le produit scalaire de vecteurs dans une grande dimension, ce qui n'est pas trivial. Le Kernel Trick consiste à utiliser une fonction noyau K tel que :

$$K(x_i, x_j) = \varphi(x_i)^T . \varphi(x_j)$$

D'où la formulation de l'hyperplan obtenu dans le cas non-linéaire est :

$$f(x) = \sum_{i=1}^{n} \alpha_i^* \, y_i K(x, x_i) + b$$

Les noyaux les plus couramment utilisés sont :

- le noyau linéaire : $K(x_i , x_j) = x_i . x_j$
- le noyau polynomial : $K(x_i , x_j) = (u x_i . x_j + v)^p$
- le noyau gaussien : $K(x_i , x_j) = \exp\left(- \frac{\|x_i - x_j\|^2}{2\sigma^2} \right)$

2.4.3.3 SVM pour la classification multi-classe

Dans le cas d'une classification multi-classes, on décompose le problème en séries de classifications binaires : *un-contre-un* ou *un-contre-le-reste*. Ces méthodes consistent respectivement à entraîner chacun des classificateurs pour séparer une classe du restant des classes ou à entraîner les SVMs afin d'obtenir toutes les frontières de décision séparant les classes une à une.

À l'heure actuelle, l'algorithme SMO (Sequential Minimal Optimization) [Platt 1998] est le plus utilisé pour appliquer les SVMs à des problèmes de grande taille. SMO est d'une part, un algorithme simple et rapide. D'autre part, en plus de ses performances en termes de temps de convergence, SMO n'exige pas un grand espace mémoire vu qu'il n'utilise pas des opérations sur la totalité de la matrice.

En plus des algorithmes décrits ici, il existe de nombreux autres algorithmes de classification qui ont également fait leur preuve dans le domaine. Citons entre autres des réseaux de neurones, des arbres de décision, ou des approches qui s'appuient sur une communauté de classificateurs comme le boosting et le bagging [Singer 1999]. Une

méthode envisagée pour classifier automatiquement des textes est donc d'interroger différents classificateurs et de combiner leurs décisions de classification, soit par un vote à majorité, soit en pondérant chaque classificateur selon sa performance testée sur des exemples de validation semblables aux documents en question.

Dès lors que les classificateurs ont été entrainés, il revient à en mesurer les performances et effectuer des comparaisons entre classificateur. La section qui suit décrira différentes techniques d'évaluation en classification.

2.5 Évaluation de performance d'un classificateur

Étant donné le nombre important de classificateurs qui existent, il est nécessaire de pouvoir les comparer et de savoir lequel répond le mieux à une tâche spécifique. C'est dans ce but que plusieurs techniques et critères ont été mis en œuvre. D'une façon générale, on divise l'ensemble en deux parties, une pour entraîner le système et une autre pour tester le modèle. L'on pourrait avoir parfois une troisième partie appelée ensemble de validation qui servira à optimiser les hyper-paramètres.

Pour mieux décrire les différents critères d'évaluation qui vont suivre, nous utilisons la matrice de contingence illustrée dans le Tableau 2, qui montre le nombre de cas de test dans chaque catégorie.

	Décision Positif	Décision négatif
Etiquette positif	a	b
Etiquette négatif	c	d

Tableau 2: Matrice de contingence pour les cas de test

On définit à partir des statistiques de cette table les mesures suivantes :

- Le taux de bonne classification ou l'exactitude (accuracy)

$$acc = \frac{a + d}{a + b + c + d}$$

- La Précision : pour une catégorie c'est le nombre de documents bien classés sur le nombre total de documents classés dans la catégorie

$$prec = \frac{a}{a + c}$$

- Le rappel : c'est le rapport entre le nombre de documents bien classé et le nombre de documents qui aurait dû être classé dans cette catégorie

$$rap = \frac{a}{a + b}$$

- F–mesure : La F-mesure correspond à une moyenne harmonique de la précision et du rappel

$$Mesure\ F = \frac{\left((1 + \beta^2) * Prec * rap\right)}{\left((\beta^2 * prec) + rap\right)}$$

Le paramètre β permet de pondérer la précision ou le rappel. Il est généralement égal à 1, La mesure devient alors:

$$Mesure\ F = \frac{2 * precision * rappel}{precision + rappel}$$

Dans le cas d'une classification multi-classe, ce qui est recherché, c'est un score global et non un score pour chaque classe. Il y a deux façons de faire une moyenne pour

obtenir un score global. La macro-moyenne calcule d'abord les scores pour chaque classe et fait ensuite une moyenne sur ces scores. La micro-moyenne regroupe d'abord les données de chaque classe dans une même table de contingence globale et calcule ensuite les scores à partir de celle-ci. La distinction à faire entre ces deux méthodes est que la macro-moyenne donne une importance égale à toutes les classes, tandis que la micro-moyenne donne une importance égale à tous les documents de test**.**

Notons que le critère de meilleur taux de classification n'est pas totalement approprié dans le cas de déséquilibre des classes. De ce fait la méthode d'analyse performante qui a été développée pour ce domaine de la classification est le 'Receiver Operator Characteristic' ROC [Metz, 1978] permettant à un classificateur d'être évalué vis-à-vis d'un ensemble de conditions possibles et la valeur scalaire communément extraite est l'aire sous la courbe ROC, AUC ("Area Under the ROC Curve"), [Hanley and McNeil, 1982] et [Bradley, 1997].

Tout ce qui a été exposé jusqu'à là dans ce mémoire concerne la classification de façon générale. Il faut signaler que parfois les données auxquelles s'appliquent ces traitements présentent souvent des dis-proportionnalités importantes entre les différentes classes. Dans ce cas de figure la tâche devient plus complexe et les classificateurs ne répondent pas toujours comme ce à quoi on s'attend. La section qui va suivre exposera les difficultés rencontrées dans le cas de déséquilibre ainsi que les stratégies mise en place pour y résoudre.

2.6 Apprentissage dans le cas de données déséquilibrés

Le problème des classes disproportionnées a été un problème très étudié ces dernières années. Cela peut être illustré par un exemple simple : si 99% des données appartiennent à une seule classe, il sera difficile de faire mieux que le 1% d'erreur obtenu en classant tous les individus dans cette classe : selon les hypothèses que nous venons de citer, c'est même la meilleure chose à faire. Cet effet de déséquilibre sur les performances en apprentissage a été étudié sur des arbres de décision C4.5 par [Weiss et Provost]. La

Figure 6 suivante montre quatre courbes ROC, chacune provenant du même jeu de données mais avec une distribution différente.

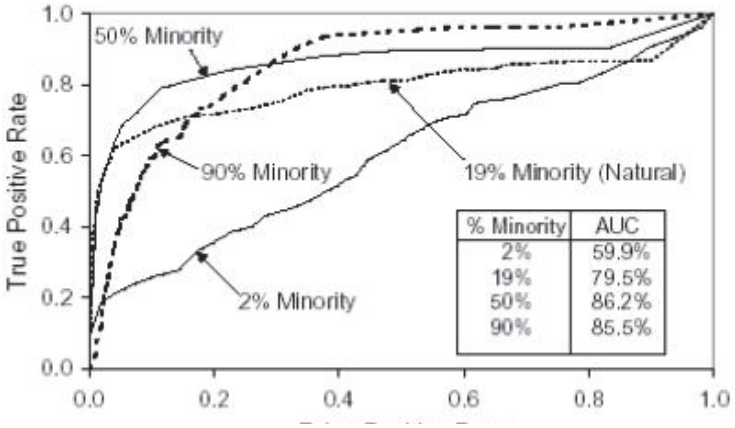

Figure 6: Courbe ROC montrant l'inefficacité dans des cas de données déséquilibrées [http://www.grappa.univ-lille3.fr/]

La courbe de la Figure 6 donne la performance des arbres de décisions sur l'apprentissage de différentes distributions d'une classe. La valeur de L'AUC varie en fonction de l'importance du déséquilibre. Les résultats montrent que les arbres de décision sont très sensibles aux déséquilibres dans les distributions des classes. La meilleure performance est obtenue lorsque les classes sont équilibrées (les classes minoritaires sont à 50% de la distribution). La performance devient la moins bonne lorsque les exemples des classes minoritaires ne représentent que 2% de la distribution. Deux stratégies a priori différentes ont été proposées par la communauté scientifique pour résoudre ce problème.

Afin de réduire au maximum l'effet de déséquilibre au sein des données de classification, il est recommandé de procéder à un rééquilibrage de la distribution des données. Les approches les plus communes utilisent le principe d'échantillonnage : soit l'on considère un sous-échantillonnage des objets de la classe majoritaire, soit l'on considère un sur-échantillonnage de la classe minoritaire. Les deux approches essaient de rééquilibrer la distribution des classes en termes d'instances incluses dans l'apprentissage. Si le sous-

échantillonnage implique des pertes d'informations potentiellement importantes contenues dans les objets, le sur-échantillonnage peut produire des effets de sur-apprentissage.

Dernièrement, différents chercheurs se sont intéressés à des approches de classification supervisée basées sur des règles pour le problème des classes disproportionnées. Par exemple, dans [Nguyen 2005], les auteurs proposent d'élaguer les corps de règles inductives issues d'un arbre de décisions. Certains chercheurs ont mis l'accent sur les défauts des approches de classification associative basées sur le cadre fréquence-confiance. Pour y remédier, [Chawla et Verhein 2007] utilisent des règles de classification dont les corps sont positivement corrélés avec une classe.

D'autres techniques ont été proposées, par exemple, la pondération des exemples d'apprentissage [Pazzani et *al* 1994], l'utilisation du Bootstrap [Catlett 1991] et [Sung et Poggio 1998], l'échantillonnage hétérogène [Lewis et Catlett 1994].

2.5.1 Les enjeux du déséquilibre pour les SVMs.

Pour les jeux de donnée moyennement asymétrique, les résultats empiriques montrent que, contrairement à d'autres techniques d'apprentissage machine, SVM peut produire une bonne hypothèse, en termes de la précision, sans aucune modification. Néanmoins, les performances diminuent lorsque le déséquilibre dans la distribution des données est important.

En termes de SVM, plusieurs tentatives ont été faites pour améliorer leur précision de la prédiction de classe [Akbani et *al*, 2004] et [Morik et *al*, 1999]. Des expériences montrent que SVM peut être en mesure de résoudre le problème des espaces vectoriels asymétriques sans introduire de bruit. Toutefois, les classificateurs qui en résultent peuvent sur-adapter aux données.

Pour traiter le problème de déséquilibre, plusieurs approches ont été données pour ajuster la frontière asymétrique. Nous en présentons quelques unes.

1. Une modification adaptative de la fonction noyau K basé sur la distribution des données d'apprentissage est une méthode efficace pour améliorer la SVM. Amari et Wu proposent une méthode de modification d'une fonction noyau pour améliorer la performance d'un classificateur SVM [Amari, Wu 1999]. Cette méthode est basée sur la structure de la géométrie Riemannienne induite par la fonction du noyau. L'idée est d'augmenter la séparabilité entre les classes en agrandissant l'espace autour de la surface frontière qui les sépare.

   Afin d'améliorer la méthode de Amari et Wu, Wu et Chang [Wu, Chang 2003] proposent un algorithme d'alignement de la frontière entre les classes, ce qui modifie également la matrice du noyau K basé sur la distribution des données d'apprentissage. Au lieu d'utiliser un espace d'entrée, ils procèdent à la transformation du noyau basé sur la distribution spatiale des vecteurs de support dans l'espace des caractéristiques.

   Les justifications théoriques et études empiriques montrent que la méthode de transformation du noyau est efficace sur la classification déséquilibrée, mais cette technique n'est pas suffisamment simple à mettre en œuvre efficacement.

2. Shawe-Taylor et Cristianini montrent que la distance d'un point de test de la frontière est liée à la probabilité d'erreurs de classification [Shawe-Taylor, Cristianini 1999]. Cette observation a motivé une technique connexe qui est utilisé. La technique est de fournir une pénalité plus sévère si une erreur est commise sur un exemple positif que si elle est faite sur un négatif par exemple. En utilisant les facteurs de coûts et en ajustant le coût de faux positifs et faux négatifs, ces sanctions peuvent être directement intégrées dans l'algorithme de SVM.

   En augmentant la marge sur le côté de la petite classe, cette méthode fournit un moyen d' induire une frontière de décision qui est beaucoup plus éloigné de la «critique» de classe qu' il est à l'autre. Mais dans ce modèle, l'équilibre entre

la sensibilité et la spécificité ne peuvent être contrôlés de manière adaptative résultant de sur-ajustement.

Synthèse

Ce chapitre a décrit un portrait global de la classification automatique de textes. Tout d'abord, il s'est attardé sur la nature du problème à résoudre, les difficultés rencontrées et les approches possibles pour le solutionner. Par la suite, il a été question de différentes stratégies de représentation des documents traités par un classificateur. Le choix judicieux d'un mode de représentation des données est nécessaire, comme pour toute application de l'apprentissage automatique. De même, des techniques de sélection et d'extraction de caractéristique/features ont été exposées. Celles-ci visant à réduire la taille du vocabulaire à traiter pour que les algorithmes évoluent dans un espace vectoriel de dimension raisonnable et diminuer le bruit au sein des caractéristiques. Il a été également question des méthodes d'apprentissage automatique mis à l'œuvre pour traiter ce problème surtout dans le cas de déséquilibre des données. On a pu constater la variété de techniques d'apprentissage pouvant amener une application informatique à classer des textes avec autonomie. En plus, on a fait la lumière sur les techniques d'amélioration des performances des classificateurs et leur processus d'évaluation.

Les techniques décrites sont génériques. Le chapitre suivant introduira un axe de recherche particulier dans lequel s'inscrivent nos travaux, à savoir l'application des techniques de classification pour faire la détection automatique de relations d'affaire dans les communiqués de presse. Pour cette tâche particulière, nous allons tester différentes approches, observer des problèmes qui surgissent et tester des méthodes pour les résoudre.

Chapitre 3

Méthodes pour la détection de liens d'affaire

Le présent chapitre décrit les différentes approches et les tests pour détecter des liens d'affaire entre différentes compagnies. Nous nous inspirons des techniques utilisées en classification de textes. Ce chapitre est organisé comme suit :

Nous décrivons d'abord le problème que nous traitons. Nous appliquons ensuite des approches de classification classiques pour le traiter. Des problèmes particuliers seront observés, notamment le problème de déséquilibre entre les classes et le problème des caractéristiques non-pertinentes. Des solutions à ces problèmes seront donc proposées, et nous effectuons des tests pour évaluer leur performance.

3.1 Description du problème traité et les données disponibles

Avant d'aborder une description détaillée de la tâche de détection de relation d'affaires, nous allons d'abord identifier les différents travaux similaires à cette tâche. Nous ferons ensuite une analyse des approches utilisées pour résoudre ce genre de problème et voir dans quelle mesure elles pourront s'adapter à notre cas. Nous tentons aussi de répondre aux questions suivantes : Pourquoi on ne peut pas se contenter uniquement de ces approches existantes pour régler le problème ? Quelle serait l'utilité d'une telle étude pour ce problème de détection de relations ?

3.1.1 La tâche de détection de relations d'affaire et son utilité

Plusieurs communiqués de presse sont émis chaque année dans le domaine de l'automobile et informent sur les différents échanges qu'il y a eu entre différents acteurs du domaine. Ils peuvent à cet effet renseigner sur les accords qui ont été signés, les clients de telle compagnie ou leur propriétaire. Ainsi à partir de ces communiqués de presse l'on serait capable d'avoir plus sur les différentes compagnies automobiles, à savoir : qui sont les leaders sur le marché? Quels sont ses clients ? Qui est le propriétaire de telle compagnie?...

Ces informations sont très précieuses pour l'analyse dans le domaine d'affaire. En effet, le projet dans lequel s'inscrit notre étude est proposé par des chercheurs de HEC-Montréal. Ce projet est axé globalement sur l'intelligence d'affaire qui tente d'analyser différents aspects des affaires : les joueurs importants dans un domaine, leurs liens d'affaires, la compétitivité des compagnies, etc. Différents documents sont utilisés pour ces fins : les communiqués de presses, les rapports annuels des compagnies, etc. Notre étude se concentre sur la détection des relations d'affaires à partir de ces communiqués de presses. Nous avons pris le domaine d'automobile comme domaine d'essai.

Notre tâche consiste à trouver le lien qui existe entre deux compagnies de façon automatique. Les types de lien sont prédéfinis par des experts en analyse d'affaire (voir des détails plus tard). Le lien pourrait être que l'une des compagnies est *propriétaire* de l'autre ou est *client* de l'autre. Voyons un exemple d'un extrait de communiqué de presse.

Exemple :

– lien d'affaire:

Magna has acquired the remaining 40% interest in MST from Kolbenschmidt and has reached an agreement in principle with Temic Telefunken Mikroelectronic ("Temic"), an affiliate of Daimler Benz, to sell Temic a 25% interest in MST.

– Pas de lien d'affaire:

The story reports reliable sources as saying Ford Motor Co. alone paid out more than $3.2 billion in warranty claims in 1992. The magazine estimates General Motors Corp. may have spent more than $5 billion and Chrysler Corp. in excess of $1 billion during the same period, based on their size in relation to Ford.

Le premier exemple montre qu'il existe un lien d'affaire entre l'entreprise *Magna* et *Daimler Benz*, tandis que dans le deuxième exemple, les deux acteurs *General Motors* et *Chrysler Corp* n'entretiennent pas de relation directe.

Les chercheurs de HEC-Montréal réalisaient cette tâche en se basant sur des règles préétablies manuellement, ce qui n'est pas une chose aisée. Voici quelques exemples de règles utilisées:

1- *In the car industry, when the title is "Ward's report" it is automatically "no link"*

2- *When the names of the 2 companies are in the title, there is a very good chance that there is a link*

3- *When the names of the 2 companies are in the same paragraph, there is a chance that there is a link*

4- *Links are found in the first 2 or 3 paragraphs or in the 2 or 3 last paragraphs more often than in the rest of the pressrelease.*

5- *When the pressrelease is a report from one of the 2 companies, there is usually a link, for example: if the title is "Quarterly results of Pfizer" and we are looking for a link between Pfizer and another companies, there is a very good chance that there is a link.*»

Cette approche voit ses limitations dans le fait qu'elle nécessitait la lecture intégrale des communiqués de presse pour identifier ces règles avant de décider s'il y a relation ou pas. Ce qui demandait énormément de temps pour ces chercheurs étant donné la masse importante de documents. Il est donc souhaitable d'automatiser cette tâche. C'est donc dans ce cadre que notre étude s'inscrit. Nous avons pour objectif d'étudier plus précisément les performances de la classification: est-ce qu'une approche de classification automatique pourrait donner une bonne performance pour la détection des relations d'affaire? Quel classificateur utilisé? Et quelles stratégies de sélection de caractéristiques adoptées pour obtenir de meilleurs résultats? Comment paramétrer la classification? …

Nous allons décrire les données à notre disposition dans la section suivante.

3.1.2 Les données disponibles pour la tâche

Nous disposons du corpus HEC-DATA fournit par des chercheurs en Analyse d'affaire de HEC-Montréal. Ce corpus contient un ensemble de 44160 communiqués de presse de 50 compagnies automobiles américaines émises au cours des années 1994 à 2005 en langue anglaise. Ces communiqués de presse traitent entre autres des annonces concernant des compagnies automobiles ainsi que des diverses relations qu'entretiennent ces dernières. Précisons que ce corpus avait déjà été préalablement annoté et chaque document est relié à sa (ses) classe(s). Ce corpus comporte 18 623 428 mots. Ces communiqués de presse ne vont pas droit au but en indiquant seulement l'information pertinente à propos des relations d'affaires. Ils présentent également d'autres informations superflues pour notre tâche. Par exemple, la plupart des communiqués de presse contiennent une description des compagnies comme : «…a leading company in…». Mais pour la tâche d'identification des relations d'affaires, ce genre de description n'est pas utile. Cela indique que des prétraitements sont nécessaires avant toute autre chose afin de cibler la partie utile. Cette idée a été confirmée par des expériences préliminaires. L'exemple

suivant (Figure 7) illustre cela et montre un extrait de communiqué de presse avec des données superflue.

1/7/78 (Item 78 from file: 813)
0668101 DE002

TITLE: GOVERNOR ENGLER INTRODUCES MICHIGAN QUALITY COUNCIL FOUNDING MEMBERS, STEERING COMMITTEE

DATE: January 21, 1994 10:59 EST WORD COUNT: 740

ROCHESTER, Mich., Jan. 21 /PRNewswire/ -- Gov. **John Engler today emphasized the importance of total quality practices to Michigan public and private sector organizations, when he introduced members of the Michigan Quality Council and unveiled the organization's logo.**

The Michigan Quality Council was created by Gov. Engler in November 1993 to promote Total Quality Management (TQM) practices in business.
The council also will establish and administer an annual Michigan Quality Leadership

MICHIGAN QUALITY COUNCIL FOUNDING MEMBERS AND STEERING COMMITTEE
Founding Members
Mr. M. Lawrence Parker, Director - Quality Programs, Chrysler Corp.
(A) Mr. Dan Whelan, Manager - Corporate Quality, Ford Motor Co.
Mr. Joseph Bransky, Director of North American Operations and Quality and Reliability, General Motors Corp. (A) Mr. Jim Warren, Vice President - Total Quality Management, Rockwell Automotive
Ms. Jeanne Heller, Executive Vice President, Manpower of Detroit, Inc.
Mr. Ray Eisbrenner, President, Eisbrenner Public Relations
Mr. Gerald Pine, Dean of the School of Education and Human Services, Oakland University
Ms. Sharon Rothwell, Director of the Office of the State Employer - Management and Budget - Special Advisor on Quality Management
Ms. Pat Wightman, TQM Coordinator - Department of Commerce, State of Michigan
Mr. Peter Behrens, Partner, Grant Thornton Accountants and Management Consultants
Mr. William Vance, Manager of Quality Systems, Haworth Inc.
Mr. Michael Wild, Vice President - Quality, AAA Mr. John Gruizenga, Senior Vice President Quality and Technical Services, Steelcase North America
Mr. David Whitwam, Chief Executive Officer, Whirlpool Corp.
Representative to be named later, Cold Heading Co.
Steering Committee
Mr. Jack Bourget, President and CEO, Manpower of Detroit, Inc.
(A) -- Both Founding Members and Steering Committee Members.
CONTACT: John Truscott of the Governor's Office, 517-335-6397; or Carole Davies of Eisbrenner Public Relations, 810-641-1446, for Michigan Quality Council

Figure 7 : Extrait d'un communiqué de presse avec des données superflues

(Tout le texte en gris représente des données superflues)

Nous disposons également de 7 différents liens d'affaire pouvant exister entre ces compagnies et qui ont été également identifiés par les chercheurs de HEC-Montréal. Nous présentons dans le tableau ci-après (Tableau 3) ces relations ainsi que la répartition des documents classées manuellement dans les classes.

classes	Nom de la Classe	Nombres de documents pertinents	Description
1	RnD	640	Relation entre deux compagnies visant le développement
2	Division	93	Relation d'affiliation entre deux compagnies
3	Association	1236	Une compagnie est associée de l'autre
4	Client	3477	Une compagnie est client d'une autre
5	Co-owner	436	Dans le cas d'une relation de co-propriétaire
6	Owner	1081	L'une est propriétaire de l'autre compagnie
7	Contract	151	Les deux compagnies sont liées par un contrat
8	No-link	34046	Il n'existe aucun lien entre les compagnies concernées

Tableau 3 : Description des différentes classes de liens

La figure ci-après (Figure 8) montre une autre vue graphique de la répartition des documents dans différentes classes.

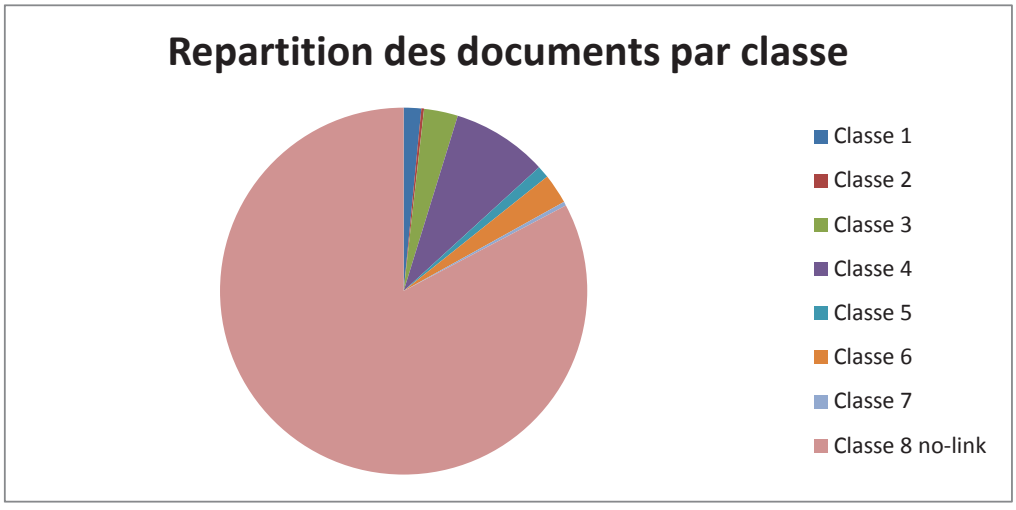

Figure 8 : Représentation graphique de la répartition des documents par classe

Il faut noter le grand déséquilibre entre les classes. Notamment, il existe beaucoup de documents pour la classe «no-link» (un communiqué de presse qui ne représente aucune des relations identifiées manuellement) par rapport aux autres classes.

On constate que les documents *no-link* représentent environ les 82% de l'ensemble des documents tandis que les documents *link* en représentent juste 18%. Cela montre l'importance d'un déséquilibre entre les différentes classes.

Nous disposons aussi d'une liste de noms des différentes compagnies. Ce sont ces compagnies que les chercheurs de HEC-Montréal traitent dans leurs analyses. Nous supposons donc que cette liste couvre complètement les compagnies d'intérêt, et que nous n'avons pas à recourir à une approche de reconnaissances automatique des entités nommées. Notons que les variations de ces noms de compagnies sont aussi manuellement identifiées. Par exemple, pour *Ford*, nous avons *Ford, Ford corp., et Ford Motors corp,* etc. Le tableau en annexe (annexe 1) présente l'ensemble des compagnies traitées.

3.2 Difficultés du projet

Nous considérons le problème de détection de liens d'affaire comme un problème d'apprentissage supervisé à partir de données textuelles, dont l'événement à prédire est représenté par 8 classes. Comme montré par les statistiques des données de notre corpus, notre corpus présente un déséquilibre prononcé entre les différentes classes (Figure 8).

Dans la plupart des problèmes de classification supervisée, le corpus est déséquilibré et les différentes classes ne sont pas représentées de manière équitable dans l'ensemble d'apprentissage. Un déséquilibre trop important affecte généralement négativement la précision des algorithmes d'apprentissage qui tentent de favoriser la classe majoritaire. Cependant, dans notre cas le fait de favoriser à classer des textes dans la classe

no-link n'aiderait pas les analyses d'affaire subséquente. Il serait nécessaire de trouver des stratégies pour régler ce problème.

Un autre problème important concerne la sélection des caractéristiques, qui affecte beaucoup la performance. Il n'est pas évident d'identifier facilement les mots qui indiquent s'il y a relation d'affaire ou pas car les communiqués de presse n'abordent pas directement ces liens d'affaire et contiennent beaucoup de données superflues.

L'autre aspect de difficulté réside au niveau de la nomenclature des différentes catégories de liens d'affaire. Certaines de ces classes peuvent paraître similaires, par exemple les classes 'association' et 'contract' ont une certaine similarité. Pour une entreprise qui a signé un contrat avec une autre, celui-ci pourrait être identifié comme étant son associé. Or dans notre cas, ces deux classes sont différentes. Ainsi, la frontière entre les classes n'est pas nette.

La section suivante fera le point sur la manière dont le problème est cerné. Nous présentons l'environnement d'expérimentation mis en place, les différentes méthodes testées et exposons notre méthodologie pour accomplir cette tâche.

3.3 Prétraitements et Environnement d'expérimentation

D'abord, mentionnons que dans tous les cas, les documents sont représentés selon l'approche vectorielle : chaque caractéristique représentant un mot du vocabulaire. Nous avons également appliqué la technique de réduction du vocabulaire tout en supprimant une liste de mots vides de sens (*«stop words»*). Quant à la technique de pondération des caractéristiques, c'est l'approche TFIDF qui a été retenue. Pour réduire la taille des vecteurs tout en éliminant les mots inutiles, une sélection de caractéristique est ensuite appliquée.

Une première phase de sélection consiste tout simplement à mettre de côté les mots dont la fréquence d'apparition parmi les documents ne dépasse pas un certain seuil

minimal, fixé empiriquement à 4 d'après des expérimentations. Autrement dit, si un terme n'apparaît pas dans au moins 4 documents différents, il est éliminé du vocabulaire. Cela aiderait à réduire une bonne partie des mots rares du vocabulaire.

Afin de réduire encore davantage la taille du lexique et d'accroître les résultats obtenus lors des opérations subséquentes de classification, nous avons procédé à une deuxième phase de sélection basée sur une autre mesure de sélection particulière, en l'occurrence le gain d'information. Il s'agit d'une méthode très utilisée en classification pour la sélection des caractéristiques informatives. Encore une fois, un seuil empirique a été fixé, et ce à 0.01. Donc, les mots dont la valeur du gain d'information par rapport aux classes est sous ce seuil sont éliminés du vocabulaire.

Le filtrage par TF-IDF dans notre cas consiste à retenir les mots dont la mesure tf-idf a une grande valeur.

La table suivante (Tableau 4) montre la taille du vocabulaire obtenu à chaque phase de traitement.

	Initial	Filtrage TF-IDF	Filtrage fréquence	Filtrage gain d'information
Taille vocabulaire	18 226	17 940	14 902	10 968

Tableau 4 : la taille du vocabulaire selon le traitement

Après avoir procédé à l'extraction et au filtrage du vocabulaire, nous avons représenté chaque document par un vecteur pondéré, et le corpus total en une matrice. À partir de cette matrice, nous avons mené nos expériences de classification.

Avant de pouvoir effectuer nos expérimentations, nous avons divisé aléatoirement la matrice en trois sous ensembles afin de générer les ensembles d'apprentissage et de test nécessaires à l'opération de classification. Le contenu de chacun des ensembles à été déterminé aléatoirement, en ne tenant pas compte de la distribution des classes attribuées manuellement par les éditeurs du corpus. L'ensemble d'apprentissage a été constitué de 2/3

du corpus (27440 documents), alors que l'ensemble de test a été constitué du 1/3 du corpus (13720 documents). Comme en témoigne la littérature sur le problème de la classification des données, il s'agit d'un ratio classique pour ce genre de tâche. Dans certaines de nos expérimentations nous utilisons également une validation croisée.

3.4 Les différentes approches de classification testées et résultats d'expérimentations

Nous avons testé plusieurs algorithmes de classifications existantes : arbre de décision, SVM, Naïve Bayes. Nous avons utilisé le package Weka pour certains algorithmes et SVMperf [Joachims 1998] pour SVM. Nous décrivons dans cette section nos tests avec les algorithmes de base.

3.4.1 Techniques de base pour la classification

L'approche de base utilise tous les mots du corpus et considère un texte comme «sacs de mots ». C'est une méthode de base souvent utilisée dans d'autres études. Nous voulons commencer par cette approche afin de nous comparer plus facilement à l'état de l'art.

Ainsi notre approche de base consiste à considérer tous les mots du corpus comme étant des caractéristiques de base et faire une classification multi-classe en un seul bloc, c'est-à-dire de traiter les 8 classes de la même manière.

Ensuite nous utilisons une mesure de sélection de caractéristique appliquée de la manière suivante. Nous avons appliqués une mesure de pondération basée sur le TF-IDF et nous avons retenu juste les 1000 premières caractéristiques qui ont les plus grandes valeurs. Nous avons retenue les 1000 premiers parce qu'un nombre important de caractéristique

n'améliorait pas les résultats, et cela a été vérifié dans nos expérimentations préliminaires avant de faire ce choix.

Les techniques de base que nous avons testées sont les suivantes : k plus proches voisins, SVM, Arbre de décision, et Naïve Bayes.

Dans le Tableau 5, nous montrons les performances de ces algorithmes. Notons qu'aucune sélection de caractéristiques n'a été utilisée ici, à part le filtrage de mots outils.

Une sélection de caractéristique avec le gain d'information (voir Tableau 6) a également été testée pour pouvoir mener une étude comparative avec notre approche.

	k-ppv		**Smo**		**Tree**		**Nb**	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Clas1	0.314	0.611	**0.5**	0.867	0.238	0.71	0.225	**0.886**
Clas2	0	0.5	0	**0.693**	0	0.591	**0.048**	0.679
Clas3	0.466	0.68	**0.55**	**0.867**	0.436	0.819	0.257	0.831
Clas4	0.586	0.831	0.579	**0.904**	0.548	0.853	0.54	0.89
Clas5	0.185	0.57	0.291	**0.929**	**0.343**	0.621	0.172	0.79
Clas6	0.431	0.703	**0.502**	**0.906**	0.424	0.842	0.301	0.822
Clas7	**0.087**	0.531	0.082	**0.8**	0.035	0.338	0.044	0.715
Clas8 (no-link)	**0.916**	0.758	0.914	0.775	0.903	0.808	0.814	**0.824**
Moy-Macro de F1	0.853	0.758	**0.856**	0.793	0.837	0.808	0.755	**0.83**
Moy-Micro de F1	**0.852**		0.850		0.829		0.701	

Tableau 5 : Résultats comparatifs des différents classificateurs

Légende :

k-ppv est le classificateur des k plus proches voisins, dans notre cas nous retenons k=10

Smo est un classificateur basé sur les SVM dans le cas multi-classe

Tree représente les arbres de décisions

Nb est le classificateur Naïve baye

Moy-Macro représente les valeurs Macro-moyenne

Moy-Micro représente les valeurs Micro-moyenne

Les figures suivantes (Figure 9 et Figure 10) nous donnent une comparaison du point de vue AUC de ces classificateurs ainsi que les résultats en termes de moyenne.

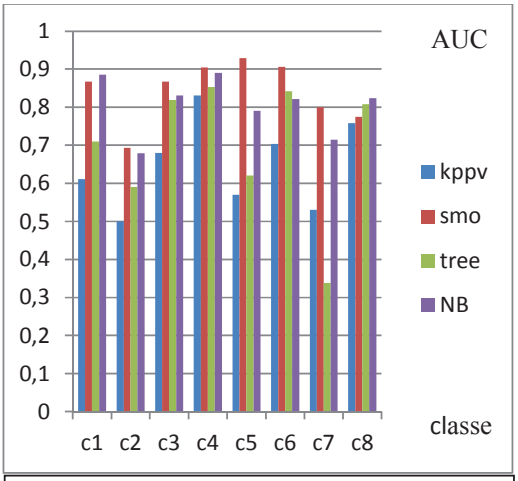

Figure 10 : Histogramme présentant la performance AUC des différents classificateurs

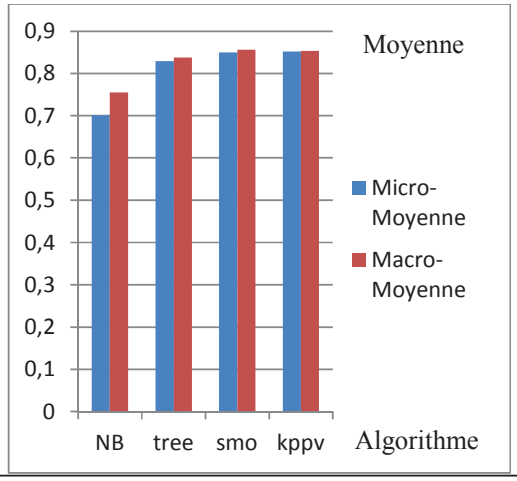

Figure 9: Histogramme présentant la moyenne Micro et Macro F1 des différents classificateurs

Observons qu'aucun algorithme ne fonctionne toujours mieux que tous les autres algorithmes avec toutes les mesures utilisées. Il est donc difficile de choisir un algorithme le plus adapté avec certitude. Cependant, dans le Tableau 5, nous pouvons observer que pour la plupart des classes SMO performe le mieux : cet algorithme a donnée les meilleures

performances dans 6 classes sur 8. C'est également SMO qui a donné la meilleure valeur de moyenne macro-F1. Notons que dans notre cas, avec un déséquilibre important, la moyenne micro donne une très grande importance à la classe 8 (*no-link*), ce qui est moins approprié pour nous, car cette classe ne nous renseigne pas sur les relations possibles entre les compagnies. La mesure de moyenne macro-F1 nous semble plus appropriée.

Ceci nous incite à choisir SVM comme notre méthode de base pour fin de comparaison. Ce choix est aussi consistant avec les résultats généraux décrits dans la littérature. Ainsi, nous allons effectuer nos expérimentations en utilisant SVM dans la suite du travail.

3.4.2 Méthodes de base pour la sélection des caractéristiques

Les premiers résultats avec les méthodes de base sont certainement améliorables en utilisant les méthodes de sélection de caractéristiques. Dans cette section, nous allons tester les méthodes de sélection fréquemment utilisées en catégorisation. Pour chaque stratégie de sélection, nous avons retenus un certain nombre de caractéristique. Le tableau 6 nous présente un extrait des caractéristiques retenues dans chaque cas. Il faut noter que le filtrage initial est juste composé des mots du corpus initial après suppression des mots outils vides de sens. La deuxième stratégie de sélection basée sur les fréquences consistait à retenir les mots apparaissant au minimum dans quatre documents du corpus. Les deux dernières stratégies sont basées successivement sur le TF-IDF et le gain d'information.

Caractéristiques retenues avec diverses stratégies de filtrage			
Filtrage Initial	Filtrage Fréquence (freq≥4)	Filtrage Tf-idf	Filtrage gain D'information (GI≥0.01)
18 226 caractéristiques	14 902 caractéristiques	17 940 caractéristiques	10 968 caractéristiques
hacienda	zone	income	General
inheritance	cradle	million	Venture
limiter	testament	quarter	Mercury
slew	acclaimed	share	Brand
Fleetwood	accumulator	increase	Facility
dazzling	organized	class	Joint
vivendi	Stanley	sale	Ford
differentiate	procedure	cash	Manufacture
amplification	ten	**subordinate**	Automobile
reconstruct	signify	voting	**Supplier**
kernel	carbon	total	Employee
franchisee	cleaner	production	Actual
tabletop	airport	dollar	Materially
decentralized	separation	radio	Differ
up	revolution	month	**dealership**
basic	Lynn	earning	statement
Newman	stakeholder	asset	powertrain
biomass	Bianchi	end	motor
liaison	sinter	period	litigation
Devill	attendance	expense	highly

Tableau 6 : Extrait des caractéristiques retenues dans chaque cas de filtrage

Dans ce cas les vocabulaires obtenus pour la représentation des documents n'étaient pas très descripteurs des relations, et nous pouvons constater qu'à part les mots *dealership*, *supplier*, et *subordinate* qui pourraient être descripteurs de relation, les autres ne le sont pas. Ces exemples nous montrent que les méthodes de sélection de caractéristiques traditionnelles ne sont pas suffisantes. Ainsi, nous allons développer notre propre méthode de sélection plus tard.

3.4.2.1 Résultats d'expérimentation avec différentes méthodes de sélection de caractéristiques existantes

Les résultats de classification obtenus avec différentes stratégies de filtrage se présentent comme l'indiquent les tableaux ci-après. Dans ces tableaux, nous indiquons seulement les moyennes macros.

Classes	Précision	Rappel	F-Mesure	AUC
Clas 1	0.128	0.224	0.163	0.768
Clas 2	0	0	0	0.648
Clas 3	0.181	0.177	0.179	0.622
Clas 4	0.279	0.125	0.173	0.719
Clas 5	0	0	0	0.613
Clas 6	0.197	0.494	0.281	0.77
Clas 7	0.017	0.085	0.029	0.665
Clas 8	0.866	0.83	0.848	0.62
Moyenne	0.77	0.731	0.746	0.633

Tableau 8 : Résultat avec filtrage Initial

Classes	Précision	Rappel	F-Mesure	AUC
Clas 1	0.314	0.164	0.216	0.695
Clas 2	0	0	0	0.714
Clas 3	0.34	0.241	0.282	0.655
Clas 4	0.388	0.18	0.245	0.657
Clas 5	0	0	0	0.518
Clas 6	0.346	0.407	0.374	0.755
Clas 7	0	0	0	0.598
Clas 8	0.874	0.94	0.906	0.584
Moyenne	0.793	0.828	0.806	0.596

Tableau 7 : Résultat avec filtrage par fréquence

Classe	Précision	Rappel	F-Mesure	AUC
Clas 1	0.479	0.522	0.5	0.867
Clas 2	0	0	0	0.693
Clas 3	0.574	0.527	0.55	0.867
Clas 4	0.474	0.743	0.579	0.904
Clas 5	0.615	0.19	0.291	0.929
Clas 6	0.539	0.469	0.502	0.906
Clas 7	1	0.043	0.082	0.8
Clas 8	0.937	0.892	0.914	0.775
Moyenne	0.873	0.851	0.856	0.793

Tableau 9 : Résultat avec tf-idf

Classe	Précision	Rappel	F-Mesure	AUC
Clas 1	0.67	0.0625	0.647	0.831
Clas 2	0.233	0.051	0.107	0.791
Clas 3	0.75	0.653	0.599	0.851
Clas 4	0.464	0.563	0.509	0.818
Clas 5	0.763	0.664	0.71	0.884
Clas 6	0.701	0.658	0.679	0.863
Clas 7	0.567	0.205	0.318	0.792
Clas 8	0.842	0.837	0.839	0.733
Moyenne	0.798	0.794	0.795	0.75

Tableau 10 : Résultat avec GI

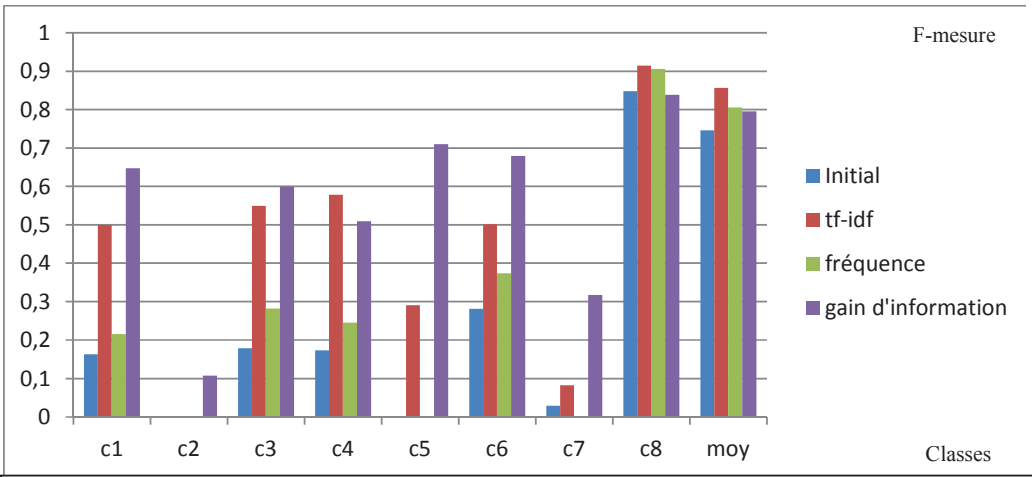

Figure 11: histogramme comparatif des différentes stratégies de filtrage des caractéristiques

3.4.2.2 Interprétations des résultats obtenus pour les méthodes de sélection de caractéristiques

La figure 11 montre que les caractéristiques retenues pour la classification agissent énormément sur les résultats. Nous pouvons constater que lorsque la sélection de ces caractéristiques est basée sur une bonne stratégie, nous pouvons obtenir de meilleures performances de nos classificateurs.

En somme les résultats obtenus avec le filtrage initial de base ne sont pas très bons. On constate dans la plupart des cas que les documents des classes moins représentées n'ont pas pu être bien classés et ont parfois un score F1 égale à 0. Cela s'explique par le fait que les caractéristiques retenues pour l'apprentissage sont décidées fortement par les classes majoritaires qui ont beaucoup plus de documents (la classe no-link). Les plus petites classes n'ont pas pu être bien représentées. En ce qui concerne le gain d'information, les résultats sont en général améliorés. Néanmoins, comme nous avons observée dans le tableau 6, beaucoup de caractéristiques retenues ne sont pas informatives. Il est donc possible d'améliorer ce filtrage, ce que nous allons faire dans la section suivante.

En utilisant SMO et en faisant une classification multi-classe en un seul bloc, les résultats globaux ne sont pas mauvais. Néanmoins si l'on examine les résultats individuels pour chaque classe, on constate qu'à part la classe 8 qui donne un bon score moyen, les autres classes ne sont pas bien classées et ont un score F-mesure très faible. Cela pourrait s'expliquer par le fait que la classe 8 présente plus d'exemples que toutes les autre classes de façon considérable et dans ce cas les classificateurs ont tendance à la favoriser des autres classes, ce qui n'est pas désirable. Nous allons traiter ce problème dans la suite de ce document. Pour l'instant, attardons nous sur le problème de sélection de caractéristiques.

Les méthodes de sélection de caractéristiques examinées considèrent tous les mots d'un document de la même façon. Or, intuitivement, les mots plus proches des noms de compagnies ont plus de chance de décrire une relation que les mots plus éloignés. Nous allons exploiter cette idée dans la section suivante.

3.4.3 Sélection des parties de document selon la position

Les mots dans un communiqué de presse n'ont pas une chance uniforme de décrire une relation. Examinons l'exemple ci-après qui décrit une relation de client entre Magna et Volvo :

NYSE:MGA) announced today that it has reached an agreement to acquire substantially all of the automotive components operations and assets of Pebra GmbH Paul Braun ("Pebra") located in Germany for approximately Cdn $31 million, subject to certain specified price adjustments.
This agreement is subject to, among other things, the completion of due diligence, the receipt of the necessary regulatory and other approvals and the negotiation and execution of a definitive purchase agreement.
The transaction is expected to be completed on or about June 30, 1996.

Magna also announced that its Decoma Exterior Systems Group ("Decoma") has formed a joint venture with a North American OEM.
The joint venture, which will be 70% owned and managed by Decoma, will operate a U.S. based fascia manufacturing and paint facility under the name "Orion Paint & Plastics L.L.C."
to supply the OEM under a long-term supply agreement.
OEM **customers** located in Europe, including Mercedes Benz, Audi, BMW, **Volvo**, Rover and Ford.

Cet exemple illustre que la compagnie Magna et Volvo sont en relation d'affaire et le mot '*customers'* est le mot le plus descriptif de cette relation de 'client'. Dans cet exemple, nous pouvons observer que ce mot apparaît entre les noms 'Magna' et 'Volvo'. Nous appelons ces noms de compagnies les mots pivots. La première intuition que nous allons utilisé est de favoriser les mots qui se trouvent entre les mots pivots dans notre représentation du document. Ceci consiste à filtrer les parties du document de telle sorte que seulement la partie entre les noms de compagnies est gardée, comme illustrée par la figure suivante :

NYSE:MGA) announced today that it has reached an agreement to acquire substantially all of the automotive components operations and assets of Pebra GmbH Paul Braun ("Pebra") located in Germany for approximately Cdn $31 million, subject to certain specified price adjustments.
This agreement is subject to, among other things, the completion of due diligence, the receipt of the necessary regulatory and other approvals and the negotiation and execution of a definitive purchase agreement.
The transaction is expected to be completed on or about June 30, 1996.

Magna **also announced that its Decoma Exterior Systems Group ("Decoma") has formed a joint venture with a North American OEM.**
The joint venture, which will be 70% owned and managed by Decoma, will operate a U.S. based fascia manufacturing and paint facility under the name "Orion Paint & Plastics L.L.C."
to supply the OEM under a long-term supply agreement.
OEM customers located in Europe, including Mercedes Benz, Audi, BMW, Volvo, Rover and Ford.

Nous espérons que ce filtrage nous permet de nous concentrer sur les caractéristiques les plus utiles pour les relations.

Cependant, cette approche peut aussi poser le problème suivant : nous avons observé que les noms de deux compagnies en relation peuvent être juste séparés par une virgule. Dans ce cas nous risquons de ne pas avoir de caractéristiques. C'est le cas dans l'exemple ci-après qui cherche à trouver la relation entre les compagnies Crysler et Dana:

Today that six companies have signed an agreement to study the Stickney Avenue Landfill and Tyler Street Dump, Toledo, Ohio.
The companies are Allied Signal Inc.; **Chrysler Corp**.; **Dana Corp**.;
The agreement, called a consent order, requires the companies to do an engineering evaluation/cost analysis (EE/CA) study to assess human- health and environmental threats posed by the two sites.

The EE/CA will also evaluate possible cleanup methods.
These methods will be detailed in an EE/CA document to be completed in Spring 1995.

E.I. du Pont Nemours and Co.; GenCorp Inc.; and Centerior Service Co.

Dans cet exemple Chrysler Corp et Dana Corp sont belle et bien en relation d'affaire. Cela nous amène à modifier cette stratégie en incluant non seulement les mots entre les noms des compagnies mais également les mots autour dans une certaine fenêtre.

La taille de cette fenêtre demeure un paramètre à déterminer lors de nos expériences. En utilisant cette stratégie, nous espérons pouvoir inclure le mot important « agreement » dans l'exemple précédent.

3.4.3.1 Résultats d'expérimentations avec l'utilisation des mots pivots

Dans la suite de nos expériences nous nous sommes focalisés sur l'utilisation des positions des mots pour représenter les documents. Ainsi un vecteur de contexte est défini autour des différents noms d'entreprises (mots pivots) afin de repérer des caractéristiques pouvant indiquer un lien d'affaire. Dans ce cas nous utilisons seulement la portion du texte définie par les mots pivots pour déterminer les fréquences.

Dans une première expérience, nous gardons seulement les mots entre les noms des compagnies comme caractéristiques. Les résultats obtenus sont décrits dans le tableau 11.

Dans une deuxième expérience, nous gardons les mots autour des noms des compagnies dans un voisinage donné. Dans cette expérience, nous fixons la taille de la fenêtre à 3, c'est-à-dire les 3 mots qui précèdent et qui suivent les noms des compagnies. Les résultats sont décrits dans le tableau 12.

Classe	Précision	Rappel	F-Mesure	AUC
Clas 1	0.748	0.663	0.703	0.918
Clas 2	0.167	0.022	0.038	0.847
Clas 3	0.856	0.714	0.779	0.953
Clas 4	0.563	0.639	0.599	0.906
Clas 5	0.869	0.672	0.758	0.979
Clas 6	0.83	0.639	0.722	0.936
Clas 7	0.705	0.205	0.318	0.88
Clas 8	0.931	0.94	0.935	0.803
Moyenne	0.889	0.888	0.887	0.824

Tableau 11: utilisation des mots 'entre'

Classe	Précision	Rappel	F-Mesure	AUC
Clas 1	0.747	0.632	0.685	0.922
Clas 2	0.667	0.086	0.152	0.855
Clas 3	0.851	0.714	0.776	0.947
Clas 4	0.569	0.653	0.608	0.917
Clas 5	0.875	0.688	0.77	0.982
Clas 6	0.828	0.628	0.714	0.932
Clas 7	0.7	0.185	0.293	0.887
Clas 8	0.931	0.94	0.936	0.805
Moyenne	0.89	0.889	0.888	0.826

Tableau 12: utilisation des mots 'autour'

La figure 13 ci-après montre une comparaison graphique entre ces deux méthodes et deux autres méthodes de sélection de caractéristiques classiques.

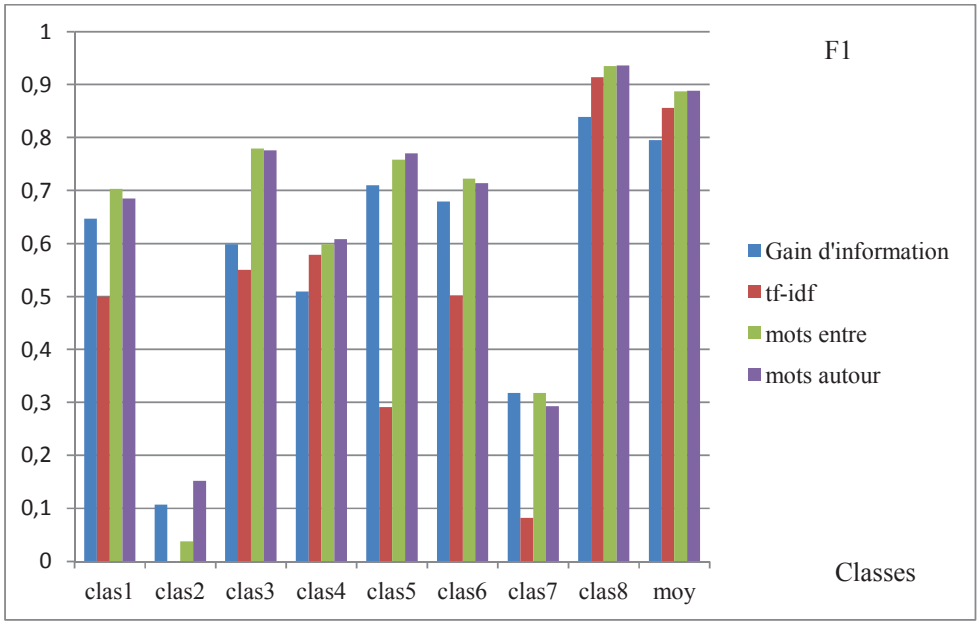

Figure 13: Histogramme des résultats comparatifs de différentes stratégies de sélection des caractéristiques

3.4.3.2 Interprétations des résultats obtenus avec les mots pivots

D'après l'histogramme de la Figure 13, on note une meilleure performance pour toutes les classes avec l'utilisation des mots pivots. Avec l'approche basée sur l'utilisation des mots pivots nous obtenons un score F1 moyen autour de 88%, tandis que c'est 79% avec le gain d'information et 85% avec la méthode basée sur tf-idf. Cette comparaison montre que les méthodes de sélection basées sur les mots pivots sont plus adaptées à notre problème que les méthodes de sélection générales.

Afin d'avoir une idée sur les caractéristiques choisies par les méthodes basées sur les mots pivots, nous illustrons dans le tableau 13, une liste des caractéristiques retenues.

En comparaison avec le Tableau 6, on remarque que plus de mots utiles pour décrire les relations sont retenus. Par exemple, les mots *contract, affiliate,* etc. sont très indicateurs d'une relation d'affaire. Cela pourrait bien expliquer les résultats obtenus précédemment dans l'utilisation des mots pivots.

Extrait des caractéristiques obtenues avec filtrage par position (toutes les 8780 caractéristiques ont été retenues)			
Company	president	facility	account
contract	source	model	**developpment**
car	global	business	complete
production	light	canadian	**associate**
system	corporation	radio	output
industry	**supplier**	engineering	part
affiliate	build	administrative	honda
content	news	powertrain	word
sale	vice	**customer**	european
unite	operation	engine	engineer
leader	product	information	lead
manufacturer	**patnership**	market	service

Tableau 13 : Les caractéristiques sélectionnées par la méthode des mots pivots

Malgré ces résultats encourageants, il reste néanmoins à signaler que lorsque nous analysons les scores F1 pour chaque classe pris séparément, on se rend compte que ces valeurs sont encore faibles pour certaines classes. Il s'agit des classes sous-représentées, c'est-à-dire les classes pour lesquelles le nombre de documents est insuffisant. Les

caractéristiques étant sélectionnées globalement pour toutes les classes, c'est la classe dominante c'est-à-dire '*no-link*', qui influence le plus cette sélection. Ainsi, la performance pour la classe '*no-link*' est tout à fait raisonnable, supérieur à 90%. Cependant, un système qui peut seulement bien classer dans cette classe n'est pas utile.

Notons qu'il est aussi possible de combiner plusieurs méthodes de sélections. Nous avons utilisé tout le document pour créer une représentation de base (un vecteur de base) avec une sélection par le GI. Ensuite on utilise les 2 méthodes de sélection basées sur les mots pivots et ça donne 2 autres vecteurs. Nous faisons une combinaison linéaire de ces vecteurs pour obtenir un vecteur final. Les résultats obtenus par une telle combinaison n'ont pas pu être très bénéfique.

Nous porterons dans la section suivante nos expérimentations sur l'effet de la taille de la fenêtre des mots pivots afin d'identifier une taille plus adaptée.

3.4.4 Expérience sur l'effet de la taille de la fenêtre des mots pivots

Les résultats obtenus jusqu'à présent avec l'utilisation des mots pivots sont très encourageants. Néanmoins il faut analyser ces résultats classe par classe puisque les différentes tailles de ces dernières ne sont pas égales. On constate en effet que pour les plus petites classes, les classificateurs performent moins bien. Il faut donc trouver un moyen pour améliorer les résultats, prendre une fenêtre de taille des documents en est une. C'est dans cette vision que nous avons mené les expériences suivantes afin d'améliorer les scores des petites classes. La stratégie utilisée consiste à définir une taille de fenêtre de sélection des caractéristiques plus grande pour les classes minoritaires qui ont obtenus de faible score de classification. Il s'agit des classes 2, 4, 7. Les résultats obtenus en variant la taille de fenêtre pour ces classes se présentent dans les tableaux 14, 15, et 16 ci-après.

Classe	Précision	Rappel	F-Mesure
Clas 1	0.673	0.807	0.734
Clas 2	0.526	0.357	0.426
Clas 3	0.836	0.761	0.797
Clas 4	0.577	0.640	0.607
Clas 5	0.769	0.786	0.777
Clas 6	0.809	0.792	0.801
Clas 7	0.441	0.520	0.477
Clas 8	0.945	0.934	0.939
Moyenne	0.697	0.699	0.694

Tableau 15 : Mot pivots avec taille de fenêtre égale à 3

Classe	Précision	Rappel	F-Mesure
Clas 1	0.672	0.817	0.737
Clas 2	0.627	0.551	0.586
Clas 3	0.827	0.795	0.810
Clas 4	0.597	0.648	0.621
Clas 5	0.799	0.796	0.797
Clas 6	0.820	0.800	0.809
Clas 7	0.561	0.592	0.552
Clas 8	0.945	0.934	0.939
Moyenne	0.731	0.741	0.730

Tableau 14 : Mot pivots avec taille de fenêtre égale à 5

Classe	Précision	Rappel	F-Mesure
Clas 1	0.732	0.803	0.766
Clas 2	0.333	0.333	0.333
Clas 3	0.869	0.746	0.803
Clas 4	0.577	0.647	0.610
Clas 5	0.823	0.806	0.814
Clas 6	0.796	0.679	0.697
Clas 7	0.388	0.564	0.459
Clas 8	0.945	0.934	0.939
Moyenne	0.682	0.670	0.677

Tableau 16 : Mot pivots avec taille de fenêtre égale à 7

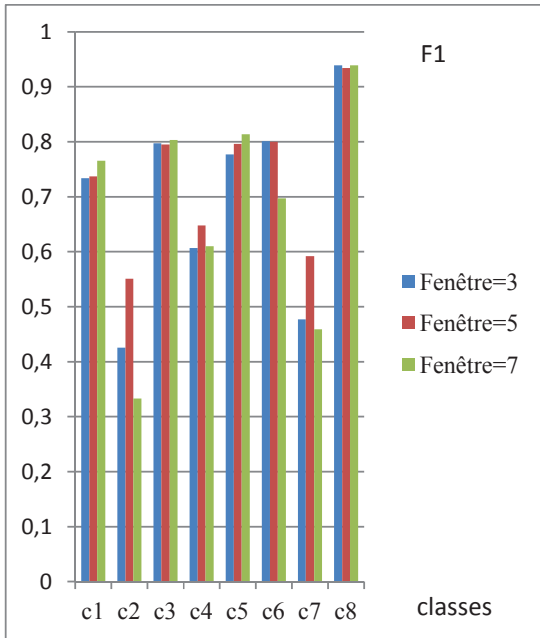

Figure 14 : Histogramme représentatif des résultats avec des tailles de fenêtre variable

D'après la figure 14, les résultats montrent que la taille de la fenêtre est un paramètre important et déterminant pour la classification et les classes sous représentées nécessitent une taille de fenêtre plus grande que les autres classes.

Ainsi le choix de fenêtre peut toujours poser des problèmes. Une fenêtre trop petite ne couvre pas les mots utiles qui sont loin, mais une fenêtre trop grande va inclure des mots bruits. Néanmoins nous porterons dans la section suivante nos expérimentations sur le problème de déséquilibre afin de trouver une solution adéquate et retenons ainsi une taille de fenêtre égale à 3 pour la suite.

3.4.5 Traitement du déséquilibre des classes

3.4.5.1. Technique de ré-échantillonnage de notre base

Le déséquilibre entre les classes a un grand impact négatif sur la performance de classification, notamment en favorisant les classes majoritaires. L'équilibrage des données d'entrainement est une solution généralement utilisée pour traiter ce problème. L'équilibrage de la base d'entrainement a pour effet de ré-calibrer le nombre d'exemples utilisés pour chaque classe, de telle manière que chaque classe possède un nombre d'exemples équivalents.

Deux méthodes de base de rééquilibrage sont utilisées : le sur-échantillonnage des classes minoritaires et le sous-échantillonnage des classes majoritaires. Le sur-échantillonnage consiste à sélectionner plus d'échantillons pour constituer une base d'entrainement. Dans le cas d'une petite classe, ceci signifie que certains exemples sont sélectionnés plus d'une fois (au hasard). Le sous-échantillonnage signifie qu'on ne sélectionne qu'une partie des exemples pour une grande classe. Dans notre expérimentation, nous allons combiner ces deux approches : sous-échantillonner la classe *no-link* et sur échantillonner les autres classes pour arriver à un nombre d'exemples d'entrainement équivalent pour chaque classe. Cette approche est implantée comme suit :

Soit m le nombre moyen d'exemples de toutes les classes. Nous considérons m comme le nombre d'échantillons cible. Ainsi, pour chaque classe, nous effectuons m tirage avec remise au hasard.

3.4.5.2. Résultats d'expérimentation obtenus pour le traitement du déséquilibre

Notre base d'apprentissage était constituée de 27440 documents dont 21952 documents pour la classe majoritaire *no-link* et 56 documents dans la plus petite classe. La moyenne de toutes les classes est 3430 exemples/classe. Ainsi, le ré-échantillonnage va tirer ce nombre de documents pour chaque classe.

Les résultats avec l'ensemble d'entrainement non-équilibré (Tableau 18) et équilibré (Tableau 17) sont présentés dans les tables suivantes. La figure 15 montre une comparaison de l'AUC pour les différentes classes.

Classe	Précision	Rappel	F-Mesure	AUC
Clas 1	0.747	0.632	0.685	0.922
Clas 2	0.667	0.086	0.152	0.855
Clas 3	0.851	0.714	0.776	0.947
Clas 4	0.569	0.653	0.608	0.917
Clas 5	0.875	0.688	0.77	0.982
Clas 6	0.828	0.628	0.714	0.932
Clas 7	0.7	0.185	0.293	0.887
Clas 8	0.931	0.94	0.936	0.805
Moyenne	0.89	0.889	0.888	0.826

Tableau 18 : Résultat SMO avec mot pivot sur ensemble non équilibré

Classe	Précision	Rappel	F-Mesure	AUC
Clas 1	0.199	0.556	0.293	0.828
Clas 2	0.51	0.2	0.019	0.682
Clas 3	0.421	0.682	0.521	0.864
Clas 4	0.415	0.784	0.543	0.917
Clas 5	0.166	0.878	0.28	0.946
Clas 6	0.258	0.657	0.372	0.877
Clas 7	0.5	0.1	0.05	0.664
Clas 8	0.873	0.695	0.811	0.796
Moyenne	0.875	0.799	0.760	0.811

Tableau 17 : Résultat SMO avec mot pivot sur ensemble équilibré

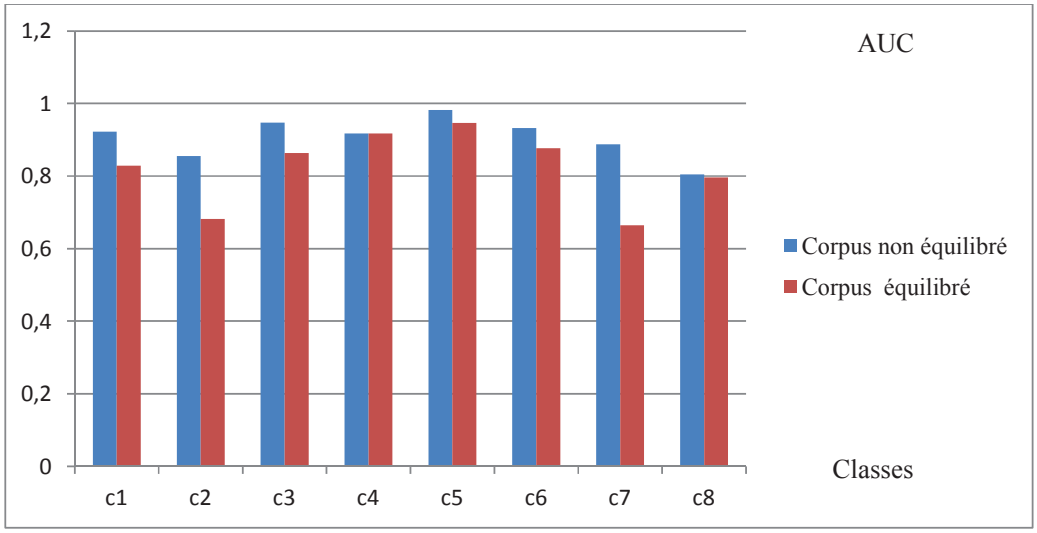

Figure 15 : Histogramme comparatif des résultats pour un corpus équilibré et un corpus non équilibré.

3.4.5.3. Interprétations des résultats sur le déséquilibre

Les résultats obtenus après un équilibrage de la base n'ont pas pu être améliorés : Le rééquilibrage a baissé la performance de presque toutes les classes. Ce résultat montre que le rééquilibrage par ré-échantillonnage du corpus ne suffit pas pour notre application.

Une particularité que nous observons dans notre application est qu'une des classes (*no-link*) est beaucoup plus grande que les autres. Cette classe se distingue des autres classes parce qu'elle ne représente aucune relation d'affaire identifiée. Ainsi, il est raisonnable de penser à une approche à deux niveaux : d'abord séparer la classe *no-link* des autres classes, ensuite séparer les différentes classes représentant une relation. Nous allons décrire cette approche dans la prochaine section.

3.4.6 Classification à deux niveaux

L'idée intuitive derrière cette approche est la suivante : quand il y a plusieurs classes de relation et une classe sans relation, il est difficile de séparer une des classes de relation des autres classes de relation et de la classe sans relation en même temps, ce que SMO tente de faire. Ceci parce qu'il est possible d'avoir des caractéristiques importantes similaires entre les classes de relation. Par exemple, un mot important pour la relation *client* pourrait aussi être important pour la relation *partenariat*. Or, les mots décrivant une relation sont plus susceptibles d'exister seulement dans les classes de relation. Ainsi, il est plus facile de différentier les classes de relation et la classe sans relation. Une fois cette séparation est faite, nous pouvons nous concentrer sur la distinction entre les classes de relation.

Cette approche à deux niveaux a un autre avantage important: les caractéristiques utilisées aux deux niveaux peuvent être différentes, et ceci semble plus intuitif. En effet, pour le premier niveau, toute caractéristique qui peut décrire une relation est importante, mais elle ne sera pas nécessairement assez discriminative au deuxième niveau. Ainsi, ses utilisations seront différentes. Autrement dit, il serait mieux de faire procéder à deux sélections de caractéristiques différentes. La figure 16 illustre ce processus à deux étapes.

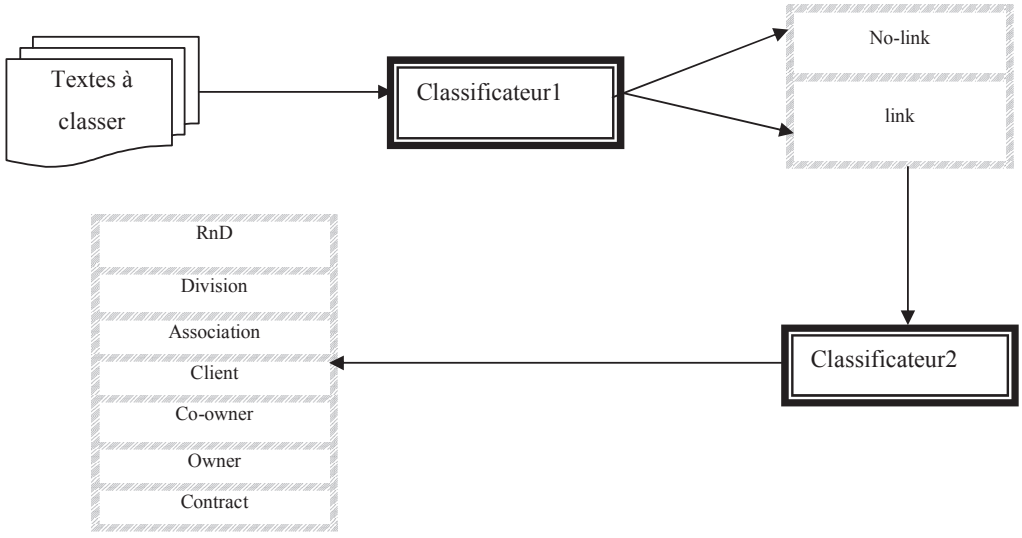

Figure 16 : Méthodologie générale pour la classification

3.4.6.1 Résultats obtenus avec une classification à deux niveaux

A ce niveau nous avons divisé la tâche en deux parties qui consiste à effectuer une première classification qui va nous permettre de séparer les documents *no-link* des documents *link* et ensuite on fera une classification multi-classe pour les documents *link*. Les résultats obtenus se présentent comme l'indique les tableaux ci-après. Les conditions de cette expérimentation sont les suivantes : une sélection de caractéristiques basée sur les mots pivots a été faite et l'apprentissage a été fait par svm.

Les résultats présentés par le tableau 19 sont issus d'une classification binaire du 1er classificateur avec comme classe *link* et *no-link*.

Accuracy	**Précision**	**Rappel**	**F1**	**AUC**
90.12	70.58	74.35	72.42	94.04

Tableau 19 : Résultat global du 1er classificateur

La figure 17 ci-après montre la F-mesure de la classe no-link et les classes link

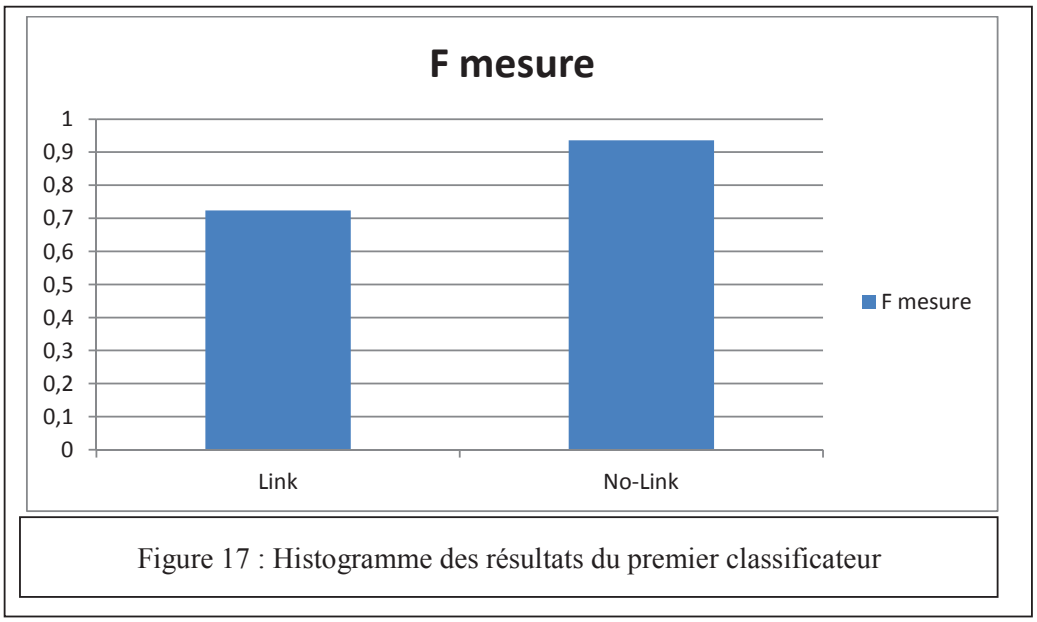

Figure 17 : Histogramme des résultats du premier classificateur

Une fois que les documents no-link sont séparés des autres documents *link*, nous procédons à une nouvelle classification multi-classe dont les 7 classes sont (Rnd, division, association, client, co-owner, owner, contract). Les résultats obtenus sont présentés dans le tableau 20.

Classe	**Précision**	**Rappel**	**F-Mesure**
Clas 1	0.673	0.965	0.793
Clas 2	0.526	0.714	0.606
Clas 3	0.836	0.904	0869
Clas 4	0.577	0.940	0.715
Clas 5	0.769	0.991	0.866
Clas 6	0.809	0.980	0.887
Clas 7	0.441	0.963	0.605
Moyenne	0.662	0.923	0.763

Tableau 20 : Résultats du 2ième classificateur sans prendre en compte les erreurs du 1er classificateur avec SVM

Afin de comparer avec les résultats précédents, nous regroupons les deux étapes en une seule table en prenant en compte les erreurs du premier classificateur. Les résultats issus des deux étapes de catégorisation se présentent dans le tableau 21.

Classe	Précision	Rappel	F-Mesure
Clas 1	0.673	0.807	0.734
Clas 2	0.526	0.357	0.426
Clas 3	0.836	0.761	0.797
Clas 4	0.577	0.640	0.607
Clas 5	0.769	0.786	0.777
Clas 6	0.809	0.792	0.801
Clas 7	0.441	0.520	0.477
Clas 8	0.945	0.934	0.939
Moy-Macro	0.697	0.699	0.694

Tableau 21 : Résultats issus des deux classificateurs

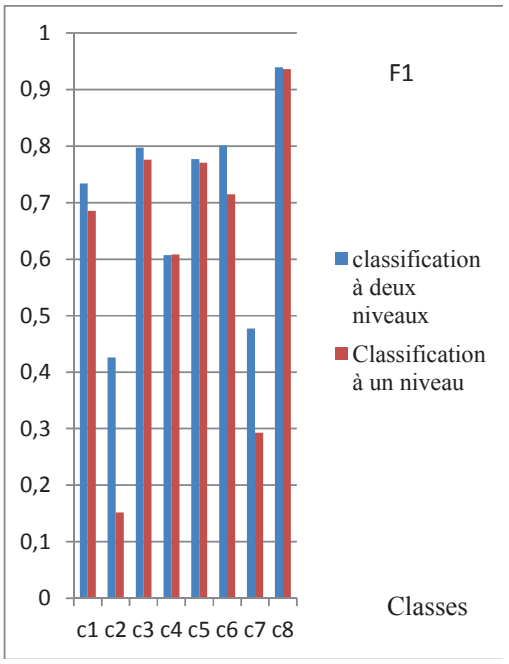

Figure 18 : Représentation graphique des résultats globaux des deux classificateurs

3.4.6.2 Interprétations sur les résultats d'une classification à deux niveaux

Une analyse de ces résultats (figure 18) montre une amélioration considérable des résultats pour chacune des classes, ce qui indique que le problème de déséquilibre a pu être mieux pallié et la performance pour la plupart des classes a pu être augmentée. Ces résultats confirment notre intuition que la différenciation entre la classe sans relation et les classes de relations est une tâche différente que de différencier entre les classes de relations. En effectuant une classification à deux niveaux, la classe sans relation est traitée séparément, ce qui permet de mieux cerner cette différence.

Comme nous avons indiqué précédemment, les caractéristiques utiles pour les deux niveaux peuvent être différentes. Ainsi, il est raisonnable d'effectuer des sélections des caractéristiques séparées en deux étapes, ce que nous allons tester dans la section suivante.

3.4.7 Sélection des parties de document en deux étapes.

Une bonne représentation des documents agit énormément sur les résultats de classification. Nous avons pu constater que différentes classes ont des mots caractéristiques différents et les mêmes mots ne caractérisent pas forcément les mêmes relations d'affaire. Par exemple pour la catégorie 'client' les mots déterminants de cette classe ne seront certainement pas les mêmes que pour la catégorie 'contract' comme l'indique le tableau 22.

Mots caractéristiques différents pour différentes classes	
Catégorie 'Client'	Catégorie ' Contract'
customer	declaration
client	Sign up
buyer	undertake
purchaser	contract

Tableau 22: Caractéristiques déterminantes d'une classe

Ainsi intuitivement, nous devons tenter de séparer une classe en se fiant plus sur les caractéristiques spécifiques pour la classe. Concrètement, nous devons idéalement effectuer une sélection des caractéristiques pour séparer chaque classe. Mais cette approche demande plusieurs sélections de caractéristiques, ce qui peut être une procédure coûteuse. Une simplification est de procéder à une sélection par niveau de classification : une première sélection pour séparer la classe *no-link* des autres classes, et une deuxième sélection pour séparer entre les classes de relation.

Dans cette section, nous faisons deux expérimentations :

- Une première expérience en adoptant une représentation commune des documents pour toutes les classes, c'est-à-dire que les mêmes caractéristiques sélectionnées sont utilisées pour représenter les documents au niveau du premier classificateur et au niveau du deuxième également.
- Une deuxième expérience en adaptant une représentation spécialisée à chaque niveau de classification. Dans ce cas nous sélectionnons les caractéristiques en deux étapes.

D'abord nous faisons une première sélection des caractéristiques pour le premier niveau en utilisant tous les documents du corpus, celle-ci servira pour représenter tous les documents. Une fois que les documents *no-link* sont séparés. Nous procédons ensuite à une deuxième sélection de caractéristique qui, cette fois-ci est basée juste sur les documents de la classe link. Dans les deux cas, la sélection se fait avec les mots autour des mots pivots.

3.4.7.1 Résultats d'expérimentations sur la sélection des parties de document à deux étapes

Les résultats obtenus avec cette stratégie de sélection des parties de document en deux étapes se présentent comme suit (Tableau 23):

Résultats avec Représentation commune des documents	Classes	Précision	Rappel	F-Mesure
	1	0.673	0.807	0.734
	2	0.526	0.357	0.426
	3	0.836	0.761	0.797
	4	0.577	0.640	0.607
	5	0.769	0.786	0.777
	6	0.809	0.792	0.801
	7	0.441	0.520	0.477
	8	0,945	0,934	0,939
Moyenne		0.697	0.699	0.694

Résultats avec Représentation Spécialisée des documents	Classes	Précision	Rappel	F-Mesure
	1	0.706	0.778	0.740
	2	0.524	0.383	0.449
	3	0.812	0.778	0.795
	4	0.574	0.639	0.605
	5	0.762	0.779	0.770
	6	0.829	0.792	0.800
	7	0.429	0.530	0.478
	8	0,930	0,911	0,924
Moyenne		0.695	0.659	0.69

Tableau 23 : Résultats pour l'utilisation d'une chaine de représentation commune et spécialisée des documents

La figure 19 ci-après montre une comparaison graphique des résultats d'une sélection à deux étapes avec celle classique.

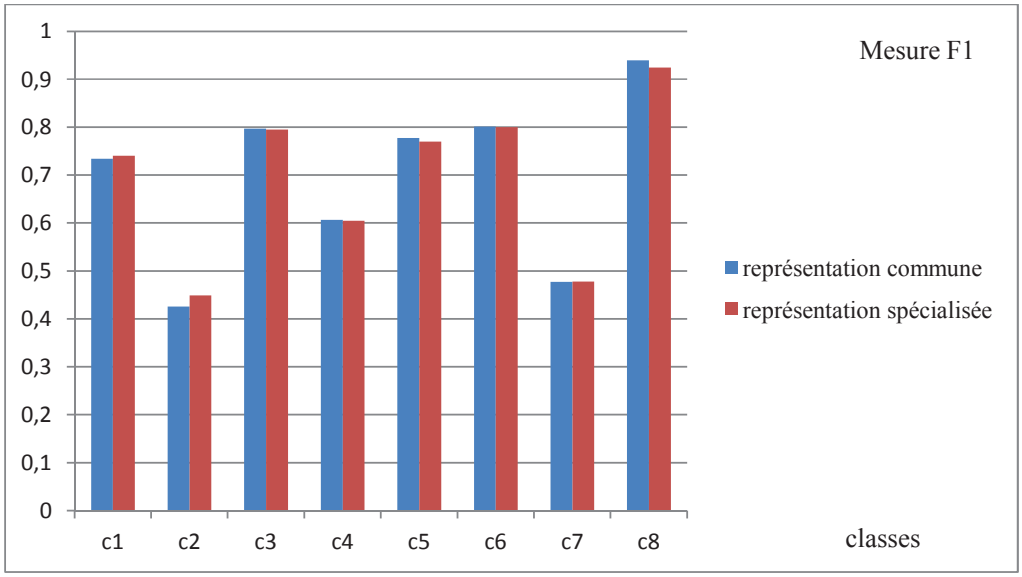

Figure 19 : Histogramme comparatif de la représentation commune et spécialisée des documents

3.4.7.2 Interprétations des résultats d'une sélection à deux étapes

La figure 19 nous indique qu'une représentation spécialisée peut apporter de légères améliorations dans certaines classes (classes 1 et 2), mais n'améliore pas la performance générale. Cela s'explique par le fait que l'on retrouve à peu près les mêmes descripteurs des documents link dans les deux cas de figure et donc les résultats de classification demeurent presque identiques. Une raison possible est que nous essayons de sélectionner les features importants pour toutes les classes au deuxième niveau. Comme plusieurs classes sont

considérées ensemble, les features sélectionnés ne sont pas nécessairement plus spécifiques à une classe particulière, par rapport à une sélection en une seule étape.

Une possibilité d'améliorer le processus de sélection des features est de la faire pour chaque SVM qui tente de séparer une classe des autres. Dans ce cas, il est possible que les features sélectionnés deviennent plus spécifiques à la classe. Nous laissons ceci à un autre travail futur.

Synthèse

Dans ce chapitre, nous avons décrit une série de méthodes de classification pour la détection de relations d'affaire. Nous avons d'abord utilisé les techniques classiques sur ce problème, y compris les algorithmes et les méthodes de sélection de caractéristiques génériques. Ces méthodes constituent les méthodes de base que nous tentons d'améliorer.

Les améliorations que nous avons présentées dans ce chapitre sont sur les deux aspects suivants :

- Une meilleure sélection de caractéristiques importantes en utilisant la position des mots par rapport aux mots pivots ;
- Une classification à deux niveaux pour traiter le problème de déséquilibre de classes.

Nos expérimentations ont montré que ces deux approches apportent des améliorations notables : En ciblant certaines parties de documents plutôt que le document entier, nous réussissons à garder les parties plus susceptibles de décrire une relation. En utilisant une classification à deux niveaux, nous pouvons mieux séparer la classe sans relation des autres classes de relation, ainsi pour mieux pallier au problème de déséquilibre entre les classes. Nous avons trouvé que cette méthode est plus adaptée que la méthode générique de ré-échantillonnage de données.

Dans nos expérimentations, nous avons aussi tenté de déterminer des caractéristiques importantes spécifiques à une classe. Mais cette tentative n'a pas obtenu d'effet positif escompté. Cela peut être dû au fait que nos tests sont encore trop simples. D'autres tests dans le futur peuvent être nécessaires.

Les méthodes utilisées dans notre travail ne sont pas exhaustives. D'autres méthodes existent. Le but de notre travail est de tester la faisabilité d'une méthode de classification automatique pour la détection des relations d'affaire. Pour cela, nous pouvons conclure que la faisabilité est démontrée, et que le but est atteint. Mais des améliorations sont possibles en exploitant les particularités des données. Nous laissons ceci à des travaux futurs.

Chapitre 4

Conclusion et Perspectives

4.1 Bilan

Dans le cadre de ce mémoire, nous nous sommes intéressés à la détection automatique de relations d'affaire. Il s'agit d'une tâche importante pour des analyses d'affaire. Les résultats d'une telle tâche permettent aux spécialistes de ce domaine de s'informer sur la nature de la relation qu'entretiennent deux compagnies automobiles.

Les données que nous avons étudiées représentent un ensemble de communiqués de presse des compagnies automobiles. L'un des aspects problématiques mis en évidence est que dans bon nombre de problèmes réels d'apprentissage supervisé les classes sont déséquilibrées, c'est-à-dire que certaines classes d'intérêt ont beaucoup d'exemples mais d'autres en ont beaucoup moins. C'est le cas du corpus sur lequel nos travaux se sont portés.

Le but de ce travail est de tester la faisabilité d'une approche basée sur la classification, et ceci comme une approche alternative à une approche manuelle. Dans cette étude, nous avons d'abord testé les approches classiques comme : Naïve Bayes, arbre de décision, k plus proches voisins et SVM. Il s'avère qu'aucun algorithme n'est toujours meilleur que les autres, mais l'algorithme SVM semble donner des résultats plus stables sur les différentes classes. Ainsi, cet algorithme est utilisé comme la méthode de base par la suite.

Nous avons aussi testé les méthodes de sélection de caractéristiques afin de choisir un ensemble de caractéristiques décrivant le mieux les relations. Cependant, nos

expérimentations ont montré que ces méthodes de sélection de caractéristiques détériorent la performance.

Cette première série d'expérimentations montre que les techniques existantes peuvent être utilisées dans une certaine mesure pour notre tâche, mais les résultats ne sont pas toujours consistants avec ceux décrit dans la littérature, notamment sur la sélection de caractéristiques. Nous pensons donc que des approches mieux adaptées à nos données doivent être développées.

Ces approches spécifiques à cette application concernent deux aspects : la sélection des parties de document pour représenter le contenu (plutôt que le document entier) et la classification à deux niveaux. La première approche vise à éliminer les parties non pertinentes du document pour la fin de détection de relation, et la deuxième approche vise à mieux traiter le problème de déséquilibre. Nos expérimentations montrent que ces approches apportent des améliorations notables sur les approches de base.

La conclusion globale de cette étude est que le problème de détection de relations d'affaire peut être traité comme un problème de classification, mais qu'il est nécessaire d'ajuster ces techniques de base aux données traitées.

4.2 Perspectives

Nos travaux comportent évidemment certaines limites ouvrant la voie à d'autres avenues de recherche. Plusieurs aspects peuvent être considérés ultérieurement.

1. Nous avons traité un seul corpus de communiqués de presse dans le domaine de l'automobile. Il serait intéressant de tester notre approche sur d'autres corpus de textes dans des domaines différents, et examiner d'autres types de documents dans le domaine d'affaire. En effet les chercheurs de HEC-Montréal ont créé un autre corpus du domaine pharmaceutique. Il serait intéressant de voir comment les différentes méthodes testées dans le domaine de l'automobile fonctionnent sur ce corpus.

2. En plus, nous pensons qu'il serait également intéressant de voir notre tâche un peu comme de la fouille d'opinion. Dans ce cas l'on prendra la présence de liens d'affaire (link) comme des opinions positives et absence de liens d'affaire comme les opinions négatives (no-link). L'approche serait de trouver des adjectifs, verbes ou mots porteurs ou non de liens d'affaires. Dans le domaine de fouille d'opinions, il est important de déterminer les mots porteurs d'opinion. Nous pensons qu'il est de même dans notre application pour les relations. Dans ce cas, on s'intéresserait plus particulièrement à l'étape d'acquisition du vocabulaire caractérisant une opinion positive ou négative d'un document. En plus d'une sélection de caractéristique de base, nous pouvons aussi penser à des approches qui utilisent des ressources linguistiques comme dictionnaire ou thésaurus, et éventuellement une acquisition automatique de ces mots.

3. Il serait aussi possible de s'appuyer sur le résultat d'une analyse syntaxique de la langue naturelle : les mots ayant une relation syntaxique avec les mots pivots (les noms de compagnies) peuvent être plus susceptibles de décrire une relation d'affaire. Ainsi, on pourrait utiliser ceci pour sélectionner ou pondérer les mots caractéristiques. Cependant, cette approche demanderait une analyse de la langue naturelle qui a une certaine complexité en opération. Il reste à voir si le gain possible justifie ce traitement assez complexe.

Malgré la portée encore limitée de cette étude, nous avons démontré que le problème important de détection supervisée de relations d'affaire peut être traité automatiquement. Ceci ouvre la porte pour d'autres études plus approfondies.

Annexe 1

Liste des compagnies automobiles		
1- Chrysler Corp	18- Hayes Lemmerz	35- Federal Mogul
2- Daimler Benz	19- Cooper Tire Rubber	36- Autoliv
3- General Motors Corp	20- Metaldyne	37- Tomkins
4- Ford Motor	21- Linamar	38- Harley Davidson
5- Volvo	22- Modine Manufacturing	39- Tenneco Automotive
6- Magna International	23- Monaco Coach	40- American Axle Manufacturing
7- Visteon	24- Federal Signal	41- BorgWarner
8- Goodyear Tire Rubber	25- Winnebago Industries	42- Nacco Industries
9- TRW Automotive	26- Wabash National	43- Dura Automotive
10- Paccar	27- United Components	44- Ducati Motor
11- Eaton Corp	28- Superior Industries	45- Gentex
12- Navistar International	29- Coachmen Industries	46- ThyssenKrupp Budd
13- Genuine Parts	30- Bandag	47- Accuride
14- Dana Corp	31- Gentek	48- National R V Holdings
15- Arvin Industries	32- Uni Select	49- Wescast
16- Meritor Automotive	33- EaglePicher	50- Aftermarket Technology
17- ArvinMeritor	34- Keystone Automotive	51- Cascade Corp

Annexe 1 : Liste des différentes compagnies automobiles traitées

Bibliographie

[Aiolli 2003] F. Aiolli et A. Sperduti, « *Multi-prototype support vector machine* », *IJCAI*, pages 541-546, Mexico 2003.

[Akbani et *al* 2004] R. Akbani, S. Kwek, and N. Japkowicz, (2004), « Applying Support Vector Machines to Imbalanced Datasets » ,in the Proceedings of the 2004 European Conference on Machine Learning (ECML'2004).

[Amari et Wu 1999] Amari et Wu, « Improving support vector machine classifiers by modifying kernel functions ». Neural Networks,12 pages 783-789. 1999

[Bellot, 2003] P. Bellot, « *Méthodes de classifications et de classifications de textes* », Université d'Avignon et des Pays de Vaucluse, Mai 2003.

[Catlett 1991] Catlett, MegaInduction: « *Machine learning on very large databases* ». PhD. Thesis, School of computer Science, University of Technology, Sydney, Australia.

[Chawla et Verhein 2007] F.Verhein, S.Chawla. «*Using Significant, Positively Associated and Relatively Class Correlated Rules For Associative Classification of Imbalanced Datasets*». The 2007 IEEE International Conference on Data Mining (ICDM'07). Pages 28-31, Omaha NE, USA. October 2007.

[Dash et *al* 1997] M. Dash, H. Liu, « *Feature Selection Methods for Classification* », Intelligent Data Analysis: An International Journal 1, pages 131-156, 1997.

[Deerwester et *al*, 1990] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman. « *Indexing by Latent Semantic Indexing »*. Journal of the American Society for Information Science, 41(6), pages 391-407, 1990.

[Dumais 1998] S. Dumais, J. Platt, D. Heckerman, M. Sahami. « *Inductive Learning Algorithms and Representations for Text Categorization* ». Proceedings of the seventh International Conference on Information and Knowledge Management (CIKM' 98), pages 148-155, 1998.

[Fix et Hodges 1951] E. Fix, J.L. Hodges, « *Discriminatory analysis, nonparametric discrimination consistency properties* », Technical Report 4, United States Air Force, Randolph Field, TX

[Ibekwe-SanJuan 2007] F. Ibekwe-SanJuan, 2007. « *Fouille de textes »*. Paris, France : Hermès-Lavoisier.

[Jensen 2005] R. Jensen, « *Combining rough and fuzzy sets for feature selection* ». PhD Thesis, University of Edinburgh, 2005.

[Joachims 1998] T. Joachims. « *Text Categorization with Support Vector Machines: Learning with Many Relevant Features »*. Proceedings of the Tenth European Conference on Machine Learning (ECML'98), Springer Verlag, pages 137-142, 1998.

[Lewis 1992] D. Lewis. « *An evaluation of phrasal and clustered representations on a text categorization task* ». In A. PRESS, Ed., *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval*, pages 37-50, New York, US, 1992.

[Lewis et Catlett 1994] Lewis, Catlett. « *Heterogenous uncertainty sampling for supervised learning* ». Proceedings of 11th ICML, pages 148-156. San Francisco, CA: Morgan Kaufmann. 1994.

[Liu et *al* 2005] H. Liu, L. Yu, « *Toward Integrating Feature Selection Algorithms for Classification and Clustering* », IEEE Trans. on Knowledge and Data Engineering, 17(4), pages 491-502, 2005.

[Maron 1961] M. Maron. « *Automatic Indexing: An Experimental Inquiry* ». Journal of the Association for Computing Machinery, 8, pages. 404-417, 1961.

[Memmi 2000] D. Memmi, « *Le modèle vectoriel pour le traitement de documents* ». Cahiers Leibniz pages 14, *INPG,* 2000.

[Mendez et *al* 2007] J.R. Mendez, B. Corzo, D. Glez-Peña, F. Fdez-Riverola, F. Díaz, «*Analyzing the Performance of Spam Filtering Methods When Dimensionality of Input Vector Changes*», in P. Perner (Ed.): MLDM, *LNAI 4571*, pages 364-378, 2007.

[Metz 1978] C. Metz, « *Basic principles of roc analysis* ». In Seminars in Nuclear Medicine, volume 3, 1978.

[Moulinier 1997] I. Moulinier. « *Apprentissage et Acquisition de Connaissances* ». Thèse de l'université Paris VI, 1997.

[Pazzani et *al*. 1994] Pazzani et al. « *Reducing Misclassification Costs* ». Proceedings of the 11th International Conference on Machine Learning, ICML 94, pages 217-225 1994.

[Pisetta et al. 2006] V. Pisetta, H. Hacid, F. Bellal, et G. Ritschard. « *Traitement automatique de textes juridiques* ». In R. Lehn, M. Harzallah, N. Aussenac-Gilles, et J. Charlet (Eds.), Semaine de la Connaissance (SdC 06), Nantes 2006.

[Platt 1998] J. C. Platt. «*Sequential minimal optimization: A fast algorithm for training support vector machines*». Technical Report MSR–TR–98–14, Microsoft Research, 1998. Available at http://www.research.microsoft.com/˜ jplatt/smo.html

[Pouliquen et *al*. 2002] B.Pouliquen, D. Delamarre, et P. L. Beux. « *Indexation de textes médicaux par extraction de concepts, et ses utilisations* ». In : A. Morin and P. Sébillot (eds.): 6th International Conference on the Statistical Analysis of Textual Data, JADT'2002, Volume 2, St. Malo, pages 617–627. France, 2002.

[Sahami, 1998] M. Sahami. « *Using Machine Learning to Improve Information Access ».* Ph.D. Dissertation, Stanford, 1998.

[Salton 1971] G. Salton. « *The SMART Retrieval System - Experiments in Automatic Document Processing* ». Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.

[Salton 1973] G. Salton, C.Yang « *On the specification of term values in automatic indexing* », journal of documentation, 29, pages 351-372, 1973

[Salton et al. 1975] G. Salton, A. Wong, et C. S. Yang. «*A vector space model for automatic indexing Commun* ».*ACM* 18(11), pages 613-620, 1975.

[Salton 1983] G.Salton, M. J. Mc gill, « *Introduction to modern information retrieval* », 1983.

[Salton 1988] G. Salton, C. Buckley. «*Term-weighting approaches in automatic text retrieval*». Inf. Process. Manage. 24(5), pages 513–523, 1988.

[Schütze et *al* , 1995] H. Schütze, D. A. Hull, J. O. Pedersen. «*A Comparison of Classifiers and Document Representations for the Routing Problem*». Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR'95), pages 229-238, 1995.

[Scott 1999] S. Scott, S. Matwin. «*Feature Engineering for Text Classification*». Proceedings of ICML-99, 16th International Conference on Machine Learning , pages 379-388, Morgan Kaufmann, San Francisco, US, 1999.

[Sebastiani 1999] F. Sebastiani. «*A Tutorial on Automated Text Categorisation»*. Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence , pages 7-35, 1999.

[Shawe-Taylor, Cristianini 1999] Shawe-Taylor, J. and Cristianini, N. (1999) « Further results on the margin distribution ». In Proceedings of the 12th Conference on Computational Learning Theory.

[Singer,1999] Y.Singer , R. Schapire «*Improved boosting algorithme using confidence-rated predictions*». Machine Learning, volume 37, n°3, pages 297-337, 1999

[Sjöblom 2002] M. Sjöblom, «*Le choix de la lemmatisation, différentes méthodes appliquées à un même corpus*». In JADT : 6es Journées internationales d'Analyse statistique des Données Textuelles, 2002.

[Stricker 2000] M. Stricker. «*Réseaux de neurones pour le traitement automatique du langage*» : conception et réalisation de filtres d'information. Thèse de Doctorat, Electronique, ESPCI, 2000.

[Sung et Poggio 1998] Sung, Poggio «*Example-based learning for view-based face detection*». IEEE Trans., PAMI 20, pages 39-51, 1998.

[Vinot et *al* 2003] R.Vinot, N. Grabar, et M. Valette, « *Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet*». In actes du colloque TALN 2003, 11-14 juin 2003, Batz sur Mer, pages 257-284.

[Wiener et *al* ,1995] E. D. Wiener, J. O. Perdersen, A. S. Weigend. «*A Neural Network Approach for Topic Spotting*». Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), pages 317-332, 1995.

[Wiener 1993] E. D. Wiener. «*A Neural Network Approach to Topic Spotting in Text*». Ph.D. Dissertation, Stanford, 1993.

[Wu, Chang 2003] Wu, G., & Chang, E. (2003). «Adaptive feature-space conformal transformation for imbalanced data learning.» Proceedings of the 20th International Conference on Machine Learning.

[Yang et Depersen, 1997] Y. Yang, J. O. Pedersen. «*A Comparative Study on Feature Selection in Text Categorization*». Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pages 412-420, 1997.

[Zarogoza, 1999] H. Zarogoza. «*Modèles dynamiques d'apprentissage numérique pour l'accès à l'information textuelle*». Thèse de l'université Paris VI, 1999.