Conservatoire National des Arts et Métiers

Année 2007

T H E S E

pour obtenir le grade de

docteur en informatique

Présentée et soutenue publiquement

par Didier NAKACHE

le 26 septembre 2007

Extraction automatique des diagnostics à partir des comptes rendus médicaux textuels

---

Préparée au sein du laboratoire CEDRIC – équipe ISID

sous la direction

d'Elisabeth Métais

Devant le jury composé de :

Président :	Pr Jacky Akoka	CNAM de Paris
Rapporteurs :	Pr Georges Gardarin	PRISM Université de Versailles Saint Quentin
	Pr Gilles Venturini	Ecole polytechnique de Tours
Examinateurs :	Pr Reind P. Van De Riet	Vrije Universitat (Pays-Bas)
	Pr Yacine Rezgui	University of Salford (Royaume Uni)
	Pr Jean François Timsit	CHU de Grenoble
Directeur de thèse :	Pr Elisabeth Métais	CNAM de Paris

SOMMAIRE

Page

REMERCIEMENTS

Je souhaite remercier très vivement Elisabeth Métais, non parce que c'est l'usage, mais pour m'avoir guidé dans mes travaux, conseillé avec professionnalisme et une très grande expertise, sans jamais compter son temps ni perdre sa bonne humeur. Ses apports majeurs me permettent aujourd'hui de vous présenter cette thèse.

Je suis très reconnaissant envers le Professeur Akoka de me faire l'honneur de présider le jury de cette soutenance et pour m'avoir accueilli dans l'équipe ISID. J'exprime toute ma gratitude au Professeur Gardarin et au Professeur Venturini pour accepter d'être mes rapporteurs de thèse. Je souhaite dire au Professeur Gardarin et son équipe du Prism et Oxymel combien j'ai apprécié de collaborer avec eux. Je remercie le Professeur Van De Riet et le Professeur Rezgui pour l'intérêt qu'ils portent à mon travail et en particulier pour avoir traversé les frontières pour assister à cette présentation. Je remercie également le Professeur Timsit pour son aide, soutien et forte implication dans ce projet.

Je tiens ensuite à exprimer ma gratitude à mon employeur, la CRAMIF, et à travers elle son Directeur Général, Mr Tonner, ainsi que Madame Bourez, Monsieur Berton, Mr Kebaili, Mr Martinez et leurs collaborateurs pour m'avoir soutenu et permis de réaliser cette thèse. Je remercie très vivement Jacques Chevalier, ainsi que l'ensemble de mes collègues de travail, les membres du laboratoire Cedric, l'association Outcome Rea, le Professeur Badran pour leur aide, collaboration et sympathique soutien. Merci à P. Bonnardel pour sa documentation sur Kappa. Mille mercis à Madame Bouali qui m'encourage depuis le début, ainsi qu'à Monsieur Aït Henani de l'Université de Lille et Messieurs Chachoua et Edeivotoh de l'EIVP. Enfin je remercie Sabine, Marie Luce, Eve, Michel, Gabrielle et Sandrine pour leur aide. Je souhaite exprimer ma sympathie à Martine Degrave, Monique Brossier, Patrice Venel et Martine Foucault pour leur soutien. Merci à tous mes sympathiques stagiaires pour leurs apports à ce projet : Annabelle Dierstein, Charles de Laguiche, Julien Le Goff, Eric Dutilleul, Grégoire Canes-Boussard, Loïc Crenn, Matthieu Ledarath, Roland Ngo, Arthur Dorival, Somruthay Thammachak, Matthieu Lucot, Frédéric Lallouet, Steve Beroard et Sabrina Dumont.

Je souhaite également remercier ma famille et plus particulièrement mon épouse Claudie, et mes enfants pour leurs encouragements constants.

CHAPITRE 1
INTRODUCTION

1.1 Présentation du sujet de la thèse

L'Assurance Maladie, ainsi que les établissements hospitaliers, possèdent de longue date de nombreuses bases de données d'informations médico-administratives. Ces bases constituent une richesse sous la forme d'un gisement de données à exploiter. Or, pour évaluer des actions visant à améliorer la santé publique, les analyses s'appuient majoritairement sur les diagnostics des patients hospitalisés. En effet, le facteur majeur de la qualification des données médicales, administratives, ou médico-administratives est l'identification du diagnostic. Si celui-ci ne pose pas de problème au praticien, son codage nécessite une précision particulière. Des diagnostics mal codés rendent les bases de données peu exploitables. Mais les informations saisies sont à finalité administrative ou comptable (codage PMSI [Programme de Médicalisation du Système d'Information] par exemple), ce qui exclut du codage les évènements non utiles, voire introduit des biais en vue d'optimiser le résultat. Ces observations induisent l'incapacité des bases de données analytiques à être en harmonie avec la réalité. Nous sommes confrontés à cette dualité dans laquelle les diagnostics sont codés à des fins de valorisation financière des séjours mais utilisés par les praticiens pour des analyses. La France, comme de nombreux pays, utilise la 10ème version de la Classification Internationale des Maladies (CIM-10). Cette dernière comporte plus de 52000 codes et est utilisée pour le codage du PMSI, mais également pour des études médico-administratives. Cette recherche de codes est une obligation légale pour chaque passage d'un patient à l'hôpital. Son rôle est triple :

- autorisation d'études épidémiologiques à échelle mondiale grâce à un codage systématique et standardisé des pathologies (c'est pour cet objectif que la CIM a été élaborée et mise à jour par l'OMS),
- base de calcul pour évaluer l'activité de l'hôpital,
- aide au calcul d'indicateurs de qualité des soins, ce qui fait que la qualité du codage a elle même été proposée comme indicateur de qualité [Foqual 2000].

Pour retrouver parmi les 52000 codes ceux qui ont motivé le séjour de son patient à l'hôpital, le praticien est confronté à un problème de temps et de subjectivité, en particulier eu égard aux domaines d'expertise du praticien. Le projet CIREA (**C**lassification **I**nformatique pour la **REA**nimation) fournit un outil informatique d'aide au codage CIM-10 qui permet aux

praticiens d'obtenir **un codage reproductible et de meilleure qualité** tout en y consacrant moins de temps. Après analyse, le logiciel propose automatiquement au praticien un ensemble de codes pertinents. CIREA est basé sur l'hypothèse que les comptes rendus d'hospitalisation (CRH) rédigés en langue naturelle sont riches d'informations et peuvent être exploités par des algorithmes d'analyse du texte (textmining). Les comptes rendus hospitaliers sont rédigés en langue naturelle à la fin de chaque séjour d'un patient à l'hôpital. L'idée est donc d'extraire automatiquement de ce document les codes CIM-10 correspondants aux pathologies du patient. Cet outil peut également être utilisé rétrospectivement. Dans ces comptes rendus, le praticien n'est limité ni par la méconnaissance des codes, ni par leur rigidité ou leurs lacunes, ni par le souci de valorisation, ni par le souci d'adhérer à la politique de codage en cours.

La finalité de nos travaux est de mettre en place un outil informatique d'aide au codage PMSI pour les services de réanimation. Mais les documents sont faits pour être lus par des humains, non pour être exploités par des systèmes automatisés, ce qui ne fait qu'accroître la complexité du problème. Pour y parvenir, nous avons dû résoudre plusieurs difficultés. En particulier, nous avons mis au point un algorithme de désuffixation, ainsi qu'un algorithme de classification des données textuelles en environnement bruité. Enfin, nous avons réfléchi à la problématique de l'évaluation et proposé de nouveaux indicateurs de mesure.

Le projet CIREA fait partie d'un ensemble plus vaste, le projet RHEA, qui vise à mettre en oeuvre des structures informatiques décisionnelles pour les services de réanimation.

Cette thèse s'articule autour de trois problématiques :

- implémentation de méthodes permettant d'extraire de la connaissance depuis les comptes rendus hospitaliers, en particulier sur la classification automatique par affectation des codes diagnostics, il s'agit du sous projet CIREA ;
- résolution de la problématique de l'évaluation ;
- conception d'une base de données orientées sujet et contenant des données médicales individualisées, il s'agit du projet RHÉA.

1.2 Le sous projet CIREA

1.2.1 Présentation générale de CIREA

Nous sommes partis de l'hypothèse que les textes rédigés en langage naturel étaient riches d'informations et contiennent de la connaissance non exploitée. Pour cela, nous avons collecté 40 000 comptes rendus hospitaliers avec les diagnostics codés. L'objectif de CIREA est de proposer des codes diagnostics sur un compte rendu. L'outil fonctionne en deux temps. Une première phase d'apprentissage permet d'identifier les relations entre le vocabulaire utilisé et les codes diagnostics inscrits par le praticien. Une seconde phase de restitution permet d'utiliser la banque de connaissances générées. Lorsque l'utilisateur présente un nouveau compte rendu, CIREA extrait les concepts médicaux en s'appuyant sur les ontologies, puis un algorithme de textmining propose une liste ordonnée de codes diagnostics selon la classification CIM 10 (Figure 1). L'outil explique également la justification des choix proposés en montrant la force de la relation entre les termes ou concepts utilisés et les diagnostics proposés.

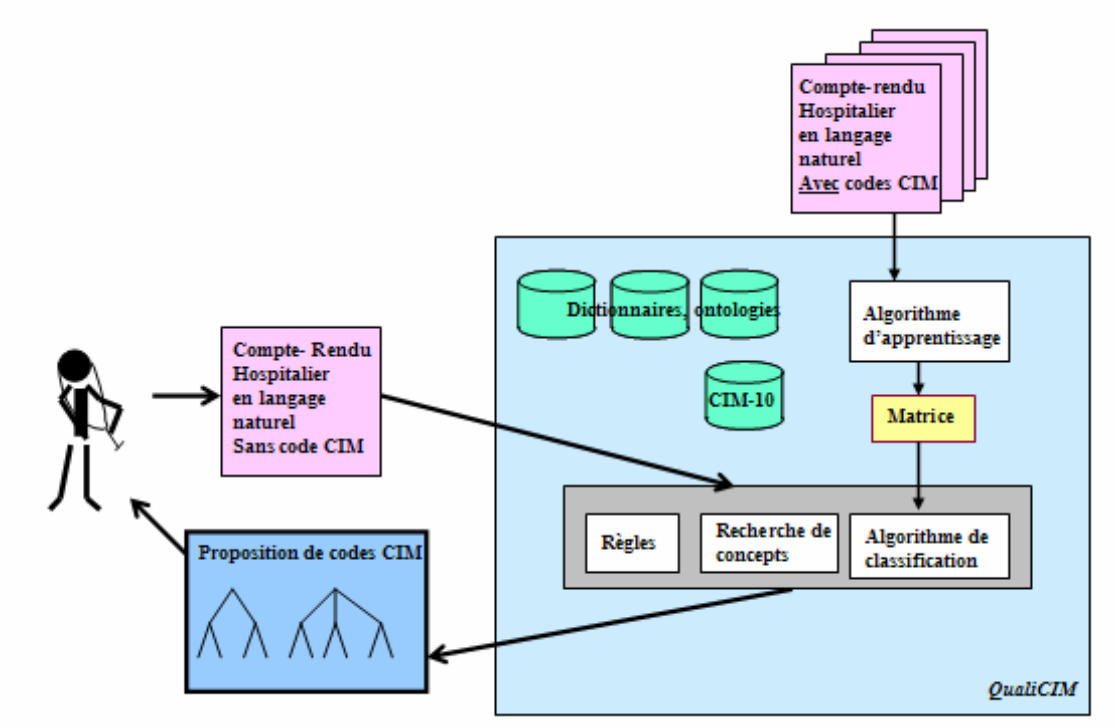


Figure 1. Sous projet CIREA

Nous avons choisi de développer le projet CIREA dans cette thèse.

1.2.2 Le contexte du PMSI

En France, la loi du 31 juillet 1991, oblige les établissements de santé à procéder à l'évaluation et à l'analyse de leur activité. Celle-ci repose sur le recueil systématique et le traitement automatisé d'une information médico-administrative minimale et standardisée recueillie au titre du PMSI (Programme de Médicalisation du Système d'Information). Il a été généralisé dans le secteur hospitalier public en 1994 et dans le secteur hospitalier privé en 1996. Présenté initialement comme outil épidémiologique (1980), puis de gestion interne (1984), il est devenu un outil d'allocation budgétaire. Il a également initié le codage des dossiers médicaux par le corps médical.

Lorsqu'un patient sort de l'établissement, un compte rendu de son hospitalisation est produit, c'est le résumé de sortie standardisé (RSS). Il est composé d'autant de résumés d'unité médicale (RUM) que le patient a fréquenté d'unités médicales pendant son séjour dans le secteur précité. La Classification Internationale des Maladies (CIM-10) est utilisée pour le codage des diagnostics des résumés de sortie. Cette codification, leur ventilation en diagnostic principal, diagnostic relié et diagnostics associés significatifs (DAS) permettent d'élaborer des Groupes Homogènes de Malades (GHM) et des Groupes Homogènes de Séjours (GHS). Les GHMs sont eux-mêmes regroupés en Catégories Majeures de Diagnostic (CMD) qui ne sont pas superposables aux grandes spécialités médicales. Chaque GHS fait l'objet d'un tarif national en euros depuis 2004. Il est ainsi possible d'établir le case mix des établissements et de mettre en œuvre la tarification à l'activité (T2A). Depuis 1998 (établissements publics et PSPH) et 2003 (secteur privé), le PMSI est étendu aux établissements de suite de soins de réadaptation, et depuis 2005 aux hospitalisations à domicile (tous deux basés sur des groupes homogènes et la CIM 10).

A cet effet, des guides de recommandations en matière de codage en réanimation ont été établis. L'intérêt de ces classifications est aujourd'hui étendu car d'une part, le dossier médical patient (DMP) utilise cette codification et d'autre part, le principe du case mix[1] s'internationalise : la très récente "Conference on European casemix-based hospital

[1] Le case mix des hôpitaux est l'étude de l'ensemble des cas d'hospitalisation (motifs, durées, age, …) à des fins d'analyse et de valorisation de séjours.

prospective payment systems"[2] organisée en France à l'initiative du ministère montre bien la convergence des différents pays vers ce système de codage.

1.2.3 Les enjeux socio économiques

A chaque sortie de patient, un diagnostic principal, un ou plusieurs diagnostics secondaires et des actes thérapeutiques (par exemple une intervention chirurgicale) doivent être sélectionnés par le praticien pour décrire la motivation d'un séjour hospitalier. Le diagnostic principal a un poids prépondérant dans l'algorithme qui oriente le séjour du patient vers un groupe homogène de malades (GHM). Ce GHM sera ensuite valorisé en points ISA (Indice Synthétique d'Activité) qui, après traitement, contribueront à déterminer le budget alloué pour l'année suivante. Du fait :

- de la diversité des patients de réanimation,
- du poids important des défaillances dans les dépenses (exemple ventilation mécanique pour insuffisance respiratoire aiguë) alors que les diagnostics qui orientent dans les groupes les mieux valorisés sont plus souvent des diagnostics « étiologiques »,
- de l'association fréquente de défaillances multiples dont le poids respectif dans les dépenses est impossible à discerner,

le choix du diagnostic principal et des diagnostics associés est par nature peu reproductible alors même qu'il est capital dans l'algorithme actuel.

Outre l'aspect budgétaire, dans un outil décisionnel tel que celui proposé dans CIREA, des études sur les codes diagnostics peuvent apporter des renseignements intéressants d'un point de vue médical. Ils peuvent également servir de technique de classification pour étudier un groupe de malades (en utilisant les GHM).

CIREA permet de ne pas pénaliser la valorisation financière d'un séjour, mais aussi de mettre un code médicalement juste, et donc scientifiquement intéressant.

[2] http://www.sante.gouv.fr/htm/dossiers/t2a/colloque/diapos/diapos.htm

1.2.4 Objectifs et justification de la recherche

Les patients qui arrivent dans les services de réanimation sont souvent porteurs de pathologies multiples et complexes, ce qui complique le codage du PMSI.

La saisie des codes diagnostic (CIM 10), voire des actes (CCAM), est compliquée car elle nécessite une recherche laborieuse sur catalogue comportant plus de 50 000 items. Sur la base de données Cubréa, colligeant les informations destinées aux PMSI et regroupant 30 services de réanimation en Ile de France, nous savons que le nombre moyen de diagnostics saisis varie entre 2.8 et 6.7, et leur nombre maximum entre 6 et 30 selon les services. Cette grande disparité ne s'explique pas par des recrutements spécialisés [Cubréa 1998]. Il y aurait donc lieu de penser que c'est un facteur humain qui explique ces écarts. Le nombre de codes diagnostics, le diagnostic principal, les comorbidités associées sont très différents d'un médecin codeur à l'autre, ce qui explique en partie la mauvaise représentation de la réanimation dans le PMSI [Misset 2000]. CIREA permet :

- d'améliorer la qualité du codage ;
- d'améliorer la reproductibilité du codage qui est à ce jour très faible ([Alberti Timsit 2000] et [Santoli 2002]) ;
- de libérer du temps pour les praticiens : le temps nécessaire au codage diagnostic est très largement influencé par celui nécessaire à la découverte des codes dans les catalogues existants. Une automatisation de la procédure diminue sensiblement le temps nécessaire au codage ;
- de rendre plus juste la valorisation financière du séjour. La saisie du diagnostic contribue fortement à la valorisation financière du séjour hospitalier ;
- de mettre en œuvre un système d'aide au codage reproductible en parfaite adéquation avec la réalité et les différentes réglementations.

Les travaux réalisés pour ce projet sont transposables à d'autres secteurs fonctionnels.

1.3 Présentation du plan

Nous avons choisi d'orienter le développement de cette thèse sur l'extraction des diagnostics à partir des comptes rendus médicaux rédigés en langage naturel. Aussi, le chapitre 2 expose l'état de l'art des méthodes de classification des données textuelles. Nous présentons les différentes méthodes de fouille de textes (textmining), y compris dans les phases de préparation des données, permettant la classification ainsi que les corpus habituellement utilisés pour évaluer ces algorithmes. Nous avons également réalisé un travail de comparaison de ces différentes méthodes en vue de déterminer lesquelles sont les plus efficientes et avons mis en évidence les différentes variantes des corpus utilisées dans les publications.

Le chapitre 3 expose en détails les différentes étapes que nous avons dû accomplir et les prolématiques que nous avons résolues pour mener à bien ce travail, de la constitution de la base de données, son nettoyage, les phases d'apprentissage et les résultats. Nous avons rencontré deux difficultés majeures : l'ajout d'une troisième dimension et l'environnement bruité, en particulier par le nombre de diagnostics. En effet, les algorithmes classiques permettent d'attribuer une étiquette à un document texte. Dans le cas des comptes rendus hospitaliers, un document contient des concepts et possède plusieurs codes diagnostics. Nous sommes dans un problème d'étiquetage multi labels. Il nous faut donc travailler avec ces trois dimensions : concepts, diagnostics, documents. Certains concepts contribuent pour tout ou partie à l'identification des diagnostics, mais parfois aucunement. Il fallait donc trouver des méthodologies permettant d'appréhender le bruitage induit par ces incertitudes.

La résolution des différents problèmes nous a amené à développer deux nouveaux algorithmes : pour la préparation des données à travers la désuffixation, et pour la classification automatique. Ces algorithmes sont comparés avec ceux de l'état de l'art.

La dernière problématique majeure à laquelle nous avons été confrontés concerne l'évaluation. En effet, nous avons mis en évidence que le codage initial utilisé en phase d'apprentissage n'était pas reproductible. En demandant à deux experts de recoder en aveugle les comptes rendus déjà codés, nous avons obtenus un taux de reproductibilité intégrale de 18%. Ce point particulier soulève deux problèmes :

- comment faire un apprentissage correct avec des données incertaines, c'est à dire comment réaliser un apprentissage optimal en environnement flou ?
- comment évaluer les résultats compte tenu de cette observation ? En effet, un code diagnostic proposé par l'application peut être correct alors qu'il ne correspond pas au codage initial.

Ce dernier point pose la problématique de l'évaluation. Aussi, nous consacrons le chapitre 4 à cette problématique. Nous avons développé des mesures adaptées à ces contraintes d'incertitude.

Enfin, nous présentons le projet Rhéa, système décisionnel pour les évènement iatrogènes et nosocomiaux en réanimation dont CIREA n'est qu'un sous projet. Avant de conclure, nous présentons un bilan des résultats obtenus.

CHAPITRE 2
ETAT DE L'ART DES CLASSIFICATIONS DE DOCUMENTS TEXTUELS REDIGES EN LANGAGE NATUREL

2.1 Présentation générale de l'état de l'art

A l'instar de [De Rosnay 95] nous avons choisi d'utiliser une démarche rétroprospective : présenter l'historique des méthodes et algorithmes de classification pour mieux appréhender le présent. Cette approche nous est apparue pertinente car nous pensons qu'elle permet de mieux comprendre la situation actuelle, où nous sommes à une phase de convergence de différentes approches. Aucune méthode n'est universelle mais chacune étant pertinente, nous observons l'émergence de la mise en œuvre de méthodes hybrides, combinant différentes approches.

Nous avons d'abord souhaité définir quelques termes utilisés car cela est impactant sur les phases d'évaluation. Nous présentons ensuite les différentes méthodes utilisées en prenant comme fil conducteur l'historique. Enfin, nous présentons l'époque contemporaine combinant ces différents éléments.

2.2 Définition de la classification

Nous abordons cette présentation en définissant quelques termes qui sont utilisés dans ce document. En particulier, nous proposons de définir les termes 'classification', 'catégorisation', et 'segmentation'.

2.2.1 Fondements

Selon [Wiki 2006] nos moyens limités d'entendement nous obligent à effectuer des classifications des objets que nous devons traiter en catégories. Celles-ci ont été considérées par la philosophie :

- au départ comme préexistantes à l'observation : c'est la démarche dite platonicienne où l'on considère que les catégories en question préexistent à l'entendement humain, qui ne fait que les découvrir plus ou moins imparfaitement. Cette démarche subsista

approximativement jusqu'à la fin du Moyen Âge, où elle était curieusement désignée sous le nom de réalisme.

- par la suite estimées comme des regroupements ad hoc et ne visant qu'à la commodité d'usage : il n'existerait pas en soi de « champignons comestibles » et de « champignons vénéneux », mais l'effet observé des champignons nous aurait conduits à les classer fonctionnellement en comestibles et en vénéneux. Cette démarche opposée au réalisme du Moyen Âge fut nommée nominalisme. Bertrand Russell [Russell 1921] fait remarquer que si l'on devait les nommer aujourd'hui, on permuterait les deux appellations.

La classification automatique vise à créer ces catégories à partir de procédés ne faisant intervenir que les données, sans la subjectivité de l'expérimentateur. Bien que les premières bases de l'approche algorithmique de la classification automatique soient relativement anciennes, ce n'est qu'avec le développement de l'informatique qu'elles sont devenues possibles à mettre en œuvre sur de grands échantillons de données.

2.2.2 Etymologie et histoire des définitions

Les termes 'classification' et 'catégorisation' ont des histoires et des origines très différentes. A ce jour, il semble persister une confusion entre ces termes [Rehel 2005]. Aucune définition scientifique n'a pu être trouvée, hormis le Webster qui donne deux sens au mot classification : celui de 'taxonomie' (ou 'taxinomie') et celui de … catégorie.

La première définition de la classification date de 1787 [Féraud 1787]. Le terme apparaît pour la première fois dans la cinquième édition du dictionnaire de l'Académie Française en 1798 [ACA 1798] sous la définition : « *distribution en classes et suivant un certain ordre* ». Le terme 'catégorisation' n'existe pas dans le dictionnaire de l'Académie Française, contrairement au mot 'catégorie'. On pourrait néanmoins le définir comme étant l'action de créer des catégories ou le résultat de cette action. Ce mot vient du grec katêgoria (ou katègoriai : « qualité attribuée à un objet ») et du bas latin. Aristote définit les catégories comme étant les « espèces les plus générales de ce qui est signifié par un mot simple ». Il regroupe dans un même ensemble des éléments proches et dénombre dix catégories. D'un point de vue philosophique, elles ne sont pas sans analogie avec les 'Principes', opposés deux à deux, dont certains Pythagoriciens voulaient que tout fût formé : fini et infini, pair et impair,

repos et mouvement, etc… D'autres mouvements philosophiques (les stoïciens, les logiciens de Port Royal, ou Kant par exemple) discuteront du nombre ou du bien fondé des catégories.

2.2.3 Classification, catégorisation et segmentation

On trouve de nombreuses définitions, souvent complémentaires, parfois contradictoires. [Jalam 2003] définit la catégorisation de textes comme étant la recherche d'une relation bijective qui consiste à "*chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (étiquettes, classes)*". [Sebastiani 2004] ajoute la notion de classes cibles prédéfinies et voit la catégorisation de textes comme étant une tâche de tri. Enfin, [Turenne 2001] définit la classification en ajoutant la notion de hiérarchie de classes à travers certaines propriétés communes. Mais finalement, [Rehel 2005] rappelle que dans la littérature scientifique, les termes classification et catégorisation sont indifféremment utilisés. Pourtant, il apparaît important de définir ces termes car l'impact est important pour l'évaluation [Nakache 2006].

Compte tenu de l'historique de ces deux termes et de leur acception actuelle, **nous proposons de définir la classification** comme étant l'action d'organiser un ensemble en structures ordonnées ou hiérarchisées. **Nous définissons la catégorisation** comme étant l'action d'affecter des éléments, qui possèdent des caractéristiques communes, à des catégories pré-établies, sans relation d'ordre. Cette définition se rapproche de la pensée aristotélicienne de par l'absence d'ordre ou hiérarchie et l'analogie avec les Pythagoriciens est conservée à travers le mode de pensée booléen (un élément appartient ou non à une catégorie). Par contre, dans une classification, il sera possible de mesurer l'écart entre la proposition et la valeur attendue. Enfin, **nous définissons la segmentation** ou le clustering comme étant une catégorisation dont les classes ne sont pas obligatoirement connues à priori.

	Classification	Catégorisation	Segmentation
Résultat	Structure avec relation d'ordre (arbre ou graphe)	Structure préexistante au traitement, sans relation d'ordre	Groupes avec des propriétés communes
Structure résultante avec relation d'ordre	Oui	Non	Non
Structure résultante préexistante	Oui	Oui	Non
Ajout de nouvelles classes	Rare	Rare	Fréquent, en fonction du besoin
Exemple	Classification des emplois	Catégories socio professionnelles, Spam ou non, …	Segmentation de clientèle
Evaluation : mesure en cas d'erreur entre classe prédite et classe réelle	[0;1] car la structure hiérarchique permet de valoriser l'écart, par exemple selon le niveau du plus grand groupe commun	{0;1} car un individu appartient ou non à une classe (raisonnement booléen)	{0;1} car un individu appartient ou non à une classe (raisonnement booléen)
Exemples d'algorithmes	CAH (classification ascendante hiérarchique) et algorithmes par apprentissage	Algorithmes par apprentissage	Knn, K means, cartes auto organisatrices de Kohonen

Figure 2. Tableau comparatif entre classification catégorisation et segmentation

Pour chacune de ces tâches, nous avons en entrée des individus statistiques (pour notre cas, des documents) $\{d_1, d_2,\ldots d_i\}$, décrits par un ensemble d'attributs $\{a_1, a_2,\ldots a_j\}$ que nous souhaitons organiser en classes $\{c_1, c_2,\ldots c_k\}$. L'exemple des emplois est caractéristique de notre proposition de définitions. Pour une classification des emplois, la notion d'ordre et de hiérarchie est présente alors que la catégorie socio professionnelle représente l'appartenance ou non à un groupe, sans relation d'ordre.

2.2.4 Quelques autres définitions

Nous ferons référence et utiliserons quelques notions que nous souhaitons préciser, pour lever toute ambigüité. [Dubois 1994] définit la notion de 'terme' comme représentant une unité signifiante constituée d'un mot (terme simple) ou de plusieurs mots (terme complexe), et qui désigne une notion de façon univoque à l'intérieur d'un domaine. S'appuyant sur cette définition, [Meilland 2003] oppose le terme au(x) mot(s) par sa référence inhérente à un domaine. [Seydoux 2005] utilise la notion de "concept hyperonyme" pour désigner un nœud non feuille d'une ontologie. Nous utiliserons cette dénomination dans ce contexte.

2.3 Les comptes rendus hospitaliers

[Bachimont 1998] et [Dechilly 1999] définissent le compte rendu hospitalier comme étant un document de production destiné à être lu par un autre praticien. Il s'agit de documents très peu structurés avec quelques rubriques communes dont : motif, histoire de la maladie, évolution, et conclusions.

Sur le plan sémantique, nous constatons que les comptes rendus sont tous rédigés dans un style narratif, avec peu de conjugaisons autres que le présent, et très peu de formes flexionnelles. Ce point est important à souligner car les algorithmes de désuffixation autorisent de nombreuses formes possibles inutiles dans ce contexte.

2.4 Convergence des méthodes

L'explosion des ressources textuelles non structurées, en particulier le développement exponentiel des pages web, a suscité dès les années 1980 un regain d'intérêt pour les techniques de traitement automatique de documents. [Turenne 2001] montre la multiplicité des besoins fonctionnels qui en résultent : catégorisation de documents, retour d'expérience, constitution de ressources lexicales, système d'alerte, création de bases de connaissances…. Nous pouvons y ajouter la recherche d'information, le résumé automatique de documents, l'expansion de requêtes, le filtrage d'e-mails ou de pages web, … Parmi ces différentes problématiques, nous nous intéressons plus particulièrement à la classification automatique de documents textuels (TC).

Nous pouvons représenter le problème de classification assez simplement sous un angle purement mathématique ([Jalam 2003], [Jaillet 2004]). Dans cette approche, l'objectif d'une classification est d'identifier une fonction discriminante permettant d'affecter les bonnes classes pour chaque document. Mathématiquement, cela se traduit par la recherche d'une fonction Φ classifiante :

$$\Phi : D \times C \rightarrow \{VRAI, FAUX\}$$

Où D représente l'ensemble des documents, C l'ensemble des catégories. Cette fonction retourne Vrai si le document D_i appartient à la classe C_j.

Pour résoudre ce problème, de nombreuses approches ont été expérimentées. Nous allons dans un premier temps en faire un inventaire rapide.

2.4.1 Les premières approches

Les premières approches visèrent à utiliser des techniques existantes et à les adapter au besoin fonctionnel. Après quelques observations, dénombrements et statistiques ([Sparck Jones 1972], [Zipf 1949]), l'approche la plus naturelle consiste à décrire chaque document par un ensemble de mots clés identifiés par les experts du domaine [Anton 1988]. Mais cette approche atteint très rapidement ses limites (méthode WORD). [Rocchio 1971] et [Salton 1975] ont implémenté des formules basées sur les occurrences de mots avec succès. Cette approche s'est poursuivie par la mise en œuvre de bases de règles ([Serradura 2002]), et de systèmes experts ([Hardt 1988], [Vleduts 1987]). Le résultat le plus célèbre est celui obtenu avec le projet CONSTRUE, développé par le Carnegie Group, dont les résultats obtenaient environ 90% de précision et de rappel sur le corpus Reuters [Hayes 1990], mais uniquement sur un sous ensemble de ce corpus (environ 3% du corpus initial). Ces dernières méthodes posaient cependant deux problèmes, le temps de génération des règles et leur incapacité à produire de nouvelles classes.

2.4.2 Induction, machine learning et régressions

Dans le prolongement, nous avons vu se développer l'adaptation de systèmes à induction pour résoudre les problématiques de classification automatique de documents. Les systèmes WASP-1, RIPPER et CHARADE ([Apte 1994], [Moulinier 1996], [Cohen 1996]) sont des systèmes à induction de règles DNF (Disjunctive Normal Form). Bien qu'existant depuis les années 1960 [Maron 1961], les concepts de "machine learning" ont trouvé ici un domaine d'application efficace. Ainsi, nombre d'algorithmes à base d'apprentissage ont été appliqués avec succès. Les arbres de décision utilisent une sélection de mots informatifs en se basant sur le gain d'information. [Mitchell 1996] a implémenté avec succès ID3 [Quinlan 1986] dans le projet DTREE, [Moulinier 1997] a implémenté l'algorithme C4.5. [Lewis 1994] a pleinement exploité Naïve Bayes pour la classification automatique de documents, en se basant sur le corpus Reuters. Peu d'études comparatives existent entre ces approches et elles

sont contradictoires, souvent utilisées en introduction d'une autre approche [Apté 1994]. Enfin, les réseaux de neurones [Wiener 1995] et [Ng 1997] ont été adaptés pour faire de la classification automatique de documents, ainsi que KNN [Yang 1994]. Citons enfin une dernière catégorie d'approches plus originale, avec LLSF [Yang et Chute 1992] basée sur une régression multiple mais surtout SVM ([Vapnik 1995] et [Joachims 1998]).

2.4.3 Approche TALN

Après avoir expérimenté les problèmes de classification par des méthodes mathématiques et statistiques, certains chercheurs ont tenté une approche en se basant sur le TALN (traitement automatique du langage naturel), par exemple [Deerwester et al. 1990] avec LSI (Latent Semantic Indexing), système basé sur les réductions de dimensions en utilisant les notions de synonymie et polysémie. Cette approche a posé de nouvelles problématiques, en particulier de désambiguïsation [Audibert 2003], de représentation du langage [Zweigenbaum 1999], …

2.4.4 Mais où en sommes-nous aujourd'hui ?

La méthode idéale et universelle n'existe pas. Chacune des méthodes testées présente des avantages et des inconvénients. Nous observons aujourd'hui un usage combiné des différentes méthodes. Généralement, les techniques de TALN sont utilisées dans les phases amont jusqu'à une représentation vectorielle des documents, puis une ou plusieurs techniques sont mises en œuvre.

2.4.5 Les projets de classification textes médicaux

De nombreuses recherches existent déjà sur la classification automatique à partir de données textuelles, mais avec des différences significatives par rapport a CIREA. Ces études utilisent en général la nomenclature UMLS et ont fait depuis longtemps l'objet de publications : RIME [Berrut 1989], NLPAD [Zweigenbaum 1992], ARISTOTE [Ledoray 1992], RECIT [Baud 1992], MEDLEE [Friedman 1993] [MedLEE 95], LSP-MLP[Sager 1995], MENELAS [Zweigenbaum 1995].

Le projet [MAOUSSC 96] effectue une analyse des CRH sur une base de 80 000 dossiers, mais ne porte que sur les actes, pas sur les diagnostics. De plus, ceux-ci sont codés en utilisant le CDAM (Catalogue Des Actes Médicaux), très différents de la CIM (Classification Internationale des Maladies). D'autres projets de traitement de données médicales textuelles ont permis l'élaboration d'ontologies spécialisées. Le projet [MedLEE 95], pionnier historique dans ce domaine en France s'appuie sur le projet [Menelas 94] : Menelas prend en charge la problématique de la représentation et du codage des connaissances médicales tandis que MedLEE est un logiciel d'analyse de compte rendus, mais limité à la radiologie (et mammographie), puis étendu à la cardiologie (1997). Le projet [PERTOMed 2003], en cours de développement se fixe pour objectif la création d'une ontologie en vue de l'aide au codage. [Blanquet 1999] a créé une méthode d'extraction automatisée de diagnostics codés en CIM 10 à partir de comptes rendus (résumés de sorties anonymes). Cette approche a l'avantage d'être simple, efficace (98.2 % de bonnes affectations sur une moyenne de 4.7 codes présentés) mais elle s'appuie sur un thésaurus spécifique et ne concerne que le secteur de l'hématologie. De plus, elle fait appel à un balisage XML et demande à l'utilisateur de naviguer dans les différentes possibilités pour valider le bon codage. Enfin, elle prend peu en compte la spécificité des services de réanimation où les patients sont souvent porteurs de pathologies multiples, et le système devra, outre l'identification des codes possibles, mettre en œuvre des règles de gestion pour les classer. Cette recherche est la plus proche du présent projet.

Les recherches menées à ce jour ont permis de mettre en œuvre des classifications automatiques à partir du langage naturel mais dans des domaines très circonscrits (hématologie, radiologie, et cardiologie), s'appuyant souvent sur des thésaurii spécifiques.

Sur le plan algorithmique, les principales méthodes existantes traitant de la classification automatique de documents textuels sont les suivantes : SVM, KNN, LLFS, les réseaux Bayésiens.

Rialle [Rialle 2001] et Claire Grover [Grover 2002] se sont également intéressés au traitement du langage naturel en médecine à l'aide du balisage XML. L'idée maîtresse de ces projets consiste à annoter les comptes rendus avec des balises XML par extraction de mots clés, à structurer et ajouter des éléments de classifications. L'approche est intéressante mais les moyens et finalités sont différents du présent projet. D'une part, ces travaux s'appuient sur des bases de données très importantes (38 708 745 mots utilisant Medline composée de 348 566 références dans le cadre de ces études, mais sachant qu'à ce jour elle en compte plus de 480 000). D'autre part, le balisage est surtout efficace pour retrouver les comptes rendus à partir de mots clés avec les outils XML, plutôt que de les coder.

D'autres études se sont penchées sur l'utilisation du résultat : par exemple, sur l'automatisation du codage pour l'évaluation de la performance de chaque service de soins [Teres D et al 1998, Fallon et al 1997], et de la qualité des soins [Randolph A et al 1998] [Timsit 2001].

Très peu de projets traitent du PMSI, hormis les travaux de M. Bensadon et J.W. Thomas [Bensadon et al. 1997], qui ont réalisé des analyses statistiques de régression, en particulier pour estimer la durée du séjour à partir des CMD (catégories majeures de diagnostics) issus des RSA (résumés de sortie anonymes), Lombrail qui a étudié, dans le contexte français les conditions nécessaires permettant d'aboutir à un codage de qualité [Lombrail et al. 1991], et B. Misset [Misset 1997] qui s'est plus spécifiquement penché sur le problème de classification des actes de réanimation dans le cadre du PMSI. Mais ces différentes approches par rapport au PMSI ne traitent pas d'analyse automatique à partir du langage naturel.

L'état de la recherche montre également le développement de modules de recherche et d'analyse sémantique du langage médical à d'autres fins, en particulier pour la constitution de bases de connaissances comme le projet Nomindex [Pouliquen 2002], ou l'utilisation du langage médical pour faire des requêtes ou extraction de connaissances dans des bases de données médicales [Zweigenbaum 2001].

D'autres recherches se sont spécialisées dans des secteurs spécifiques, à la périphérie du projet : Bodenreider [Bodenreider 2000] s'est penché sur l'identification des noms propres dans les terminologies médicales ; Pierre Zweigenbaum [Zweigenbaum 1997] a recherché les différentes méthodes de représentation sémantiques par des graphes ; Jacquemin [Jacquemin 2000] a cherché a extraire des informations et des données à partir de données textuelles.

Les recherches sur l'extraction des données ont également donné lieu à de nombreuses publications. On citera ClinQuery [Safran 1989], un langage de requête spécialisé dans l'extraction de concepts médicaux, développé au Boston's Beith Israel Hospital. Mais cet outil semble souffrir de quelques imperfections et imprécisions (difficultés à restituer certaines informations importantes et mauvaise gestion du temps) qui ont amené d'autres équipes à développer en particulier DXTractor, outil basé sur un langage spécifique d'extraction : MQL (Medical Query Language), opérationnel sur la base OSCAR, datawarehouse médical [Costar 1999]. Parmi les plus anciens (mais en évolution permanente), le projet LSP/MLP (Linguistic String Project / Medical Language Processor) développe depuis les années 60 un analyseur de la langue anglaise (étendu au français) visant à analyser des textes narratifs pour en faciliter le requêtage. A ce jour, une expérimentation est en cours avec le codage en SNOMED, mais ce type d'outil est également concurrencé par les balisages XML. Ces différentes technologies semblent cependant converger.

Pour les systèmes basés sur XML, les outils habituels liés à cette norme fonctionnent parfaitement. Par exemple, les DomDocuments, les DTD, les formats de fichiers XSD (description des structures) et XSL (description des restitutions).

2.5 Méthodologie pour classifier les documents textuels

La synthèse des différents documents (par exemple [Sebastiani 1999], [Réhel 2005], [Pouliquen 2002], [Stricker 2000], [Sebastiani 2002]) permet de constater un large consensus sur la méthodologie globale à adopter pour la classification automatique des documents. Celle-ci comprend trois phases principales : une phase d'apprentissage, une phase de tests, une phase de restitution. Ces phases peuvent êtres couplées ou précédées d'une phase de traitement du langage naturel.

2.5.1 La phase linguistique

La linguistique peut intervenir à diverses étapes du processus. En général, elle intervient dans la préparation des données et est utilisée pour représenter les tokens sans ambiguïté.

2.5.2 La phase d'apprentissage

La phase d'apprentissage est la plus importante. Au cours de cette phase initiale, nous devons effectuer trois choix :

- choix des documents (filtering),
- choix de la représentation des documents (descripteurs),
- choix du classifieur.

2.5.2.1 Choix des documents

La première étape consiste à constituer une base de données de documents textuels. Pour cela, on peut soit utiliser un corpus existant, soit en constituer un ad hoc. Les documents doivent être étiquetés.

2.5.2.2 Choix de la représentation des documents

Il faut choisir les descripteurs du document, c'est-à-dire comment les représenter pour un système d'apprentissage. Les descripteurs représentent l'information atomique d'un document. Ils sont censés indiquer de quoi parle le document [Laporte 2000]. On parle aussi d'unités élémentaires ou de tokens [Jacquemin et Zweigenbaum 2000]. Le but étant de les choisir de manière à perdre le moins d'information sémantique possible.

2.5.2.3 Choix du classifieur

La dernière étape consiste à identifier l'algorithme (le classifieur) qui fera la meilleure estimation de la classification d'un document.

2.5.3 La phase de tests

Lorsque les documents sont vectorisés, le corpus sera découpé en deux groupes de données. Le premier groupe permettra d'identifier le meilleur classifieur et les paramètres qui lui sont associés. Mais l'évaluation portera sur le second groupe, inutilisé en phase d'apprentissage.

2.5.4 La phase de restitution

Lorsque le système est opérationnel, il suffit de présenter un nouveau document au système pour obtenir une prédiction sur la classification.

2.6 L'approche linguistique

Déterminer un champ sémantique, c'est dégager la structure d'un domaine de significations. Traditionnellement cette structure est symbolisée par des relations d'ensembles hiérarchiques ou non hiérarchiques bien particulières. Il existe 4 types de relations étudiées en linguistique :

- l'existence d'un sens proche (synonymie),
- l'existence d'un sens opposé (antonymie),
- la relation de spécificité (hyponymie / hyperonymie),
- la relation partie-tout (méronymie / holonymie) :

Les relations hyperonymiques traduisent une relation de généricité ("couleur" par rapport à "bleu") (inversement on a la relation hyponymique) alors que les relations méronymiques traduisent un phénomène d'inclusion ("bleu azur" est inclus dans le "bleu") (inversement on a la relation holonymique).

Pour aborder la problématique de classification automatique de comptes rendus hospitaliers, il est possible d'utiliser de nombreuses méthodes, souvent complémentaires. Nous allons d'abord aborder les aspects du traitement automatique du langage.

Les premières recherches concernant le Traitement Automatique du Langage (T.A.L.) visaient à créer un modèle unique, capable de formaliser la compréhension du langage dans sa globalité. Cette approche, initiée par Chomsky, s'est vite avérée difficile à mettre en œuvre.

A l'heure actuelle, les systèmes de T.A.L. utilisent une approche à base de corpus, certes moins ambitieuse mais beaucoup plus pragmatique. Celle-ci consiste à utiliser un corpus de documents représentatifs du domaine à modéliser que le système va utiliser pour obtenir une couverture du langage suffisante pour que le système de T.A.L. donne des résultats de bonnes qualités.

Pour pouvoir présenter d'une manière claire les techniques de T.A.L., nous allons adopter une vision très théorique qui décompose l'analyse d'un texte en quatre étapes successives d'analyse :

- l'analyse morpho-lexicale, qui se préoccupe de la structure des mots,
- l'analyse syntaxique dédiée à l'analyse de la structure des phrases,
- l'analyse sémantique qui s'intéresse au sens des phrases considérées individuellement,
- l'analyse pragmatique qui s'attache à remettre un contexte autour des phrases.

Dans la pratique, un texte n'est pas soumis à l'analyse successive de différents modules, chacun traitant une caractéristique spécifique du langage. Ces différents niveaux d'analyse interagissent dynamiquement au cours du traitement d'un document écrit.

De plus, certains outils se consacrent à l'analyse de la structure des phrases alors que d'autres tentent de comprendre les textes à partir du sens des mots. De tels systèmes, basés sur l'analyse sémantique, s'intéressent à la signification des mots plutôt qu'à leur rôle syntaxique.

2.6.1 Analyse morpho-lexicale

L'analyse morpho-lexicale a pour objectif d'identifier les mots du texte (simples, composés, noms propres, abréviations) et leurs traits (genre et nombre). L'analyse morpho-lexicale se décompose en trois étapes :

- la segmentation, dont le but est de découper le texte en phrases puis en mots distincts,
- la lemmatisation qui s'attache à déterminer la forme canonique des mots isolés dans l'étape précédente,
- l'étiquetage dont l'objectif est d'identifier la bonne catégorie morphosyntaxique (verbe, nom…) des mots selon le contexte.

Cette dernière étape est très importante car elle conditionne l'interprétation du texte. Elle est également très délicate. En effet, il est parfois difficile d'attribuer une catégorie à un mot.

2.6.2 Analyse syntaxique

Une fois les mots identifiés, on s'intéresse à la structure syntaxique des phrases. On s'attache donc à fournir les analyses plausibles des phrases, en corrélation avec une grammaire qui spécifie les structures autorisées du langage.

Il existe de nombreux modèles grammaticaux complexes et élaborés, mais certaines formes grammaticales comme les ellipses ou encore les anaphores sont difficilement reconnues par les grammaires actuelles.

2.6.3 Analyse sémantique

Il faut maintenant donner un sens aux phrases du texte. Cette étape s'appelle l'analyse sémantique. Celle-ci va considérer les phrases de manière isolées pour tenter d'identifier leur sens hors contexte. Pour déterminer le sens d'une phrase, une première étape va se préoccuper du sens de chacun des mots constituant la phrase. Ensuite, à l'aide des informations fournies par l'analyse syntaxique, le sens complet de la phrase pourra être déduit grâce à la connaissance des relations existantes entre les mots.

2.6.4 Analyse pragmatique

A ce stade, il convient d'intégrer le sens littéral de chaque phrase dans son contexte, c'est l'étape traitant de la pragmatique. Pour la bonne compréhension d'un texte, un lecteur a besoin de connaître un certain nombre d'éléments qui ne sont pas exprimés explicitement dans le texte : connaissances relatives à la culture générale, au sujet abordé, etc. C'est la part implicite du langage.

L'enchaînement des étapes précédentes conduit parfois à des ambiguïtés qu'il est possible de supprimer en utilisant l'analyse pragmatique de manière dynamique.

2.6.5 Conclusion

A ce jour, les outils de T.A.L. traitent correctement les textes jusqu'au niveau sémantique. Pour cela, les projets de recherche s'appuient sur des ontologies de plus en plus performantes. Initialement très sectorielles (par exemple les ontologies médicales), nous disposons aujourd'hui de ressources orientées sur le sens commun. Ainsi, une phrase comme "l'avocat mange" n'est pas ambiguë grâce à des ontologies comme CYC. Il est important de souligner que ces traitements interviennent dans la préparation des documents en vue de leur classification et améliorent sensiblement les résultats.

2.7 Préparation des données par la desuffixation

Plusieurs algorithmes de désuffixation existent, les plus célèbres d'entre eux étant [Porter 1980], [Lovins 1968] et [Paice 1996]. L'algorithme de Porter consiste en sept phases successives et une cinquantaine de règles applicables. Le suffixe le plus long est toujours prioritaire. Malheureusement, il s'agit d'algorithmes de désuffixation pour la langue anglaise, dont les dérivés morphologiques se prêtent facilement à ce type d'adaptation. Peu d'algorithmes équivalents sont disponibles pour le Français. Le plus important d'entre eux étant [Carry 2002], algorithme de désuffixation pour le français, qui est en fait une transposition de l'algorithme de Porter à la langue française. Tout comme l'algorithme de Porter, Carry se décompose en phases successives. C'est également le suffixe le plus long qui détermine la règle à appliquer. Ainsi, le mot 'chienne' deviendra 'chienn' par suppression du 'e' final, puis 'chien' par suppression de la double consonne finale. Certaines règles sont soumises à conditions, par exemple pour éviter que le mot 'tissaient' ne se transforme en 't' par application de la règle qui supprime le 'issaient' final. L'algorithme de Porter a été adapté à de très nombreuses langues.

D'autres approches plus riches consistent à utiliser un étiquetage morpho syntaxique des mots présents dans le document [Namer 2000], ou à des analyseurs morphologiques comme Nooj [Silberztein 2004] ou Mmorph [Petitpierre 1996]. Mais ces techniques font appel à des traitements lourds, des bases de données volumineuses et nécessitent des réactualisations régulières. Ceci est d'autant plus vrai dans le langage médical où de nouveaux termes apparaissent régulièrement.

Pour mesurer le niveau de modifications apporté par la désuffixation, plusieurs techniques existent comme le nombre moyen de mots, le niveau de compression obtenu, le nombre moyen de caractères supprimés, ou la distance de Hamming. Mais [Frakes 2003] compare ces différentes mesures sur quatre algorithmes et montre que chacun obtient le meilleur résultat selon la mesure d'évaluation retenue.

2.8 La vectorisation des documents

La majorité des outils de classification se base sur une représentation vectorielle des documents. Celle-ci consiste à exprimer synthétiquement chaque document comme étant un ensemble de descripteurs significatifs (représentation en sac de mots par exemple). Cette représentation se réalise dans un tableau lexical qui permet de ranger les décomptes de chacune des formes dans chacune des parties du corpus :

	Document i
Concept j **Ou classe j**	**Effectif ij**

Figure 3. Représentation sous forme de tableau lexical

2.8.1 Techniques de préparation des données

2.8.1.1 Simplification de documents

Pour améliorer l'indexation quelques processus supplémentaires sont souvent mis en œuvre (le but étant de simplifier les documents) : l'élimination des mots-outils, la conversion de caractères, les corrections orthographiques, la reconnaissance de mots composés, et l'étiquetage lexical.

L'élimination des mots-outils (mots stop, stopword) consiste à éliminer les mots du langage courant qui ne contiennent pas beaucoup d'information sémantique. (Exemple : "a", "le", "la", "de", …). Certains indexeurs retirent systématiquement tous les petits mots (d'un, deux voire trois caractères).

La conversion de caractères couvre deux aspects : la casse (majuscules ou minuscules) et la conversion des caractères diacritiques. Beaucoup de moteurs d'indexation sont insensibles à la casse et traitent de la même manière les mots en majuscules ou en minuscules. Une fonction transformera par exemple tous les mots en majuscules. Par rapport aux caractères diacritiques, il existe encore d'anciens textes écrits en ascii sur 7 bits (c'est-à-dire sans accents), qu'il faut comparer avec des textes accentués. Encore aujourd'hui, certains titres

de documents écrits en majuscules ne sont pas accentués. Une solution consiste donc à convertir chaque caractère diacritique en ascii sur 7 bits dans le processus d'indexation.
Lorsque l'on veut indexer des textes multilingues, on convertit les caractères dans un format commun (ex: iso-latin [ISO-latin 1987], ou unicode [Unicode 1997]).

Les fautes d'orthographe ou fautes de frappe peuvent être corrigées automatiquement avant la phase d'indexation grâce à un processus de correction orthographique. Une correction automatique est très souvent associée au moteur d'indexation afin de ne pas créer deux descripteurs différents pour un mot mal orthographié. [Laporte 2000, p. 39] consacre un chapitre de son livre à ce propos ("détection et correction d'erreurs"). Le danger étant de "corriger" abusivement certains mots corrects mais inconnus du lexique (notamment les noms propres). [Bodenreider et Zweigenbaum 2000] ont réalisé une étude spécifique sur la détection de noms propres dans l'UMLS.

La reconnaissance de mots composés est une source de difficultés. Certains indexeurs utilisent une table de mots composés du langage pour les identifier comme ne formant qu'un seul mot ([Gross 1986] ou [Habert et Jacquemin 1993]). En effet, il paraît indispensable de considérer "bec de lièvre" comme un mot à part entière (et ne pas l'indexer par "bec" et "lièvre").

Enfin, l'étiquetage lexical est un outil qui associe à chaque mot (ou à plusieurs) des informations d'ordre morphologique, grammatical, syntaxique voire sémantique (exemple : "les poules du couvent couvent"). Certains systèmes d'étiquetage lexical intègrent une reconnaissance des mots composés (qui seront traités comme un mot simple par la suite).

Notons qu'il existe des outils d'accentuation automatique, qui peuvent utiliser un étiquetage lexical des mots, ce qui permet, dans la plupart des cas de lever les ambiguïtés [Simard 1996]. Certains outils utilisent un corpus d'apprentissage (comme [El-Beze 1995]) ce qui permet de lever les ambiguïtés selon l'usage le plus fréquent. Exemple : la phrase "FRACTURE DE COTE" peut tout à fait être accentuée sous les deux formes : "Fracture de côté" ou "Fracture de côte", mais dans le domaine médical la seconde forme sera, de loin, la plus probable. Pierre Zweigenbaum et Natalia Grabar ont présenté dans TALN 2002 [Zweigenbaum 2002] une étude comparative des différentes méthodes d'accentuation des

mots inconnus appliquées au thésaurus biomédical MeSH. Ces méthodes obtiennent une précision de 86 à 96 % (± 4 %) pour un rappel allant de 72 à 86 %.

2.8.1.2 La loi de Zipf

Dans les années 1930, un scientifique de l'université de Harvard, G.K. Zipf, [Zipf 1949] a constaté que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc... En classant les mots d'un texte par fréquence décroissante, on observe que la fréquence d'utilisation d'un mot est inversement proportionnelle à son rang. Cette loi peut s'exprimer de la manière suivante :

Fréquence d'un mot de rang N = (Fréquence du mot de rang 1) / N

À partir de cette loi, on constate qu'en traçant pour chaque mot le couple rang / effectif dans un repère logarithmique, alors le nuage de points paraît linéaire. Il est d'ailleurs possible de calculer l'équation de la droite de régression linéaire sur ce nuage. La répartition des points autour de cette droite montre que la linéarité n'est qu'approximative mais reste toutefois fortement significative.

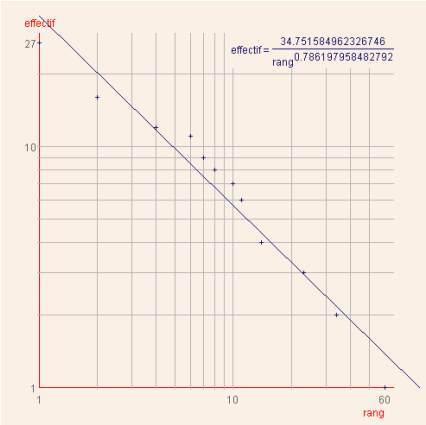


Figure 4. La loi de Zipf

En étudiant la liste des fréquences, on peut noter que les mots les plus fréquents sont les mots grammaticaux, et que leur ordre d'apparition dans la liste est stable d'un texte à l'autre à partir du moment où le texte est de longueur significative. Les mots lexicaux apparaissent ensuite, en commençant d'abord par les mots thématisant le document.

Cette loi, pertinente dans le domaine littéraire, laisserait supposer que les termes significatifs d'un document se trouvent juste après écrêtage des articles et mots grammaticaux. Lors de la préparation des données, il est fréquent de supprimer les termes fréquents, car considérés comme peu porteurs de sens, ainsi que les termes trop rares (hapax,

termes dont le nombre d'occurrences est inférieur à un seuil, généralement de 3 à 5, mais historiquement le hapax est un terme présent une seule fois dans un corpus), car considérés comme étant probablement des erreurs, par exemple typographiques.

2.8.1.3 Segmentation de textes

Un programme d'indexation, s'il travaille sur de grands documents, pourra tenir compte des différentes unités d'indexation (unité linguistique) que sont la phrase, le paragraphe, ou le document dans son ensemble. Ce qui implique l'utilisation d'un programme de segmentation. Le plus simple (mais le moins performant) étant de reconnaître une phrase comme étant une suite de mots suivie d'un point. Lorsque l'on travaille sur des textes ayant un format défini, on peut parfois extraire la notion de phrase ou de paragraphe en analysant le format (SGML, XML, dans une moindre mesure HTML). Dans le cas des comptes rendus hospitaliers, les rubriques sont connues à l'avance et facilement identifiables.

Une autre solution consiste à utiliser des étiqueteurs syntaxiques qui permettent, notamment, d'identifier les paragraphes et les phrases d'un texte. L'indexation de textes est, le plus souvent, une étape préalable d'un système plus complet comme la recherche d'informations, l'attribution de mots-clés, la similarité de documents ou la synthèse automatique.

2.8.1.4 L'étiquetage

On distingue trois types d'étiquetage :

- L'indexation manuelle, pour laquelle une personne a préalablement désigné les termes d'indexation : les descripteurs associés à chaque texte. Le projet CIREA entre dans ce cadre puisque nous disposons des données étiquetées manuellement ;
- L'indexation semi-automatique (ou supervisée) : un programme détecte, pour chaque document, des descripteurs qui sont proposés à un utilisateur (qui peut valider, supprimer ou, parfois, ajouter des descripteurs) ;
- L'indexation automatique (appelée parfois "non supervisée" par opposition) : le programme fonctionne sans intervention humaine.

2.8.2 Choix des descripteurs

Il faut maintenant retenir une stratégie de représentation des documents pour les vectoriser. Habituellement les descripteurs sont :

- les mots du document : toute chaîne de caractères compris entre deux séparateurs (espace, virgule,…).
- Les racines des documents. Dans ce cas, on parlera de racinisation. Ce procédé consiste à ne conserver que les 'n' premiers caractères de chaque terme. Dans [Grabar 1999], Pierre Zweigenbaum a démontré que pour la terminologie médicale, les premiers caractères d'un terme étaient suffisamment significatifs pour en permettre la racinisation, moyennant l'application de quelques contraintes. Son expérimentation a permis d'obtenir 95% de bons résultats. Cette hypothèse permet de regrouper des mots exprimés dans des déclinaisons différentes.
- Les documents désuffixés. Le principe consiste a retirer les suffixes. Il existe plusieurs algorithmes de désuffixation, les plus célèbres d'entre eux étant [Porter 1980], voire [Lovins 1968] et [Paice 1996]. L'algorithme de Porter consiste en sept phases successives et une cinquantaine de règles applicables. Le suffixe le plus long est toujours prioritaire. Malheureusement, il s'agit d'algorithmes de désuffixation pour la langue anglaise, dont les dérivés morphologiques se prêtent facilement à ce type d'adaptation. Peu d'algorithmes équivalents sont disponibles pour le Français. Le plus important d'entre eux étant [Carry 2002], algorithme de désuffixation pour le français, qui est en fait une adaptation de l'algorithme de Porter à la langue française. Tout comme l'algorithme de Porter, celui de Carry se décompose en phases successives. C'est également le suffixe le plus long qui détermine la règle à appliquer. Ainsi, le mot 'chienne' deviendra 'chienn' par suppression du 'e' final, puis 'chien' par suppression de la double consonne finale. Certaines règles sont soumises à conditions, par exemple pour éviter que le mot 'tissaient' ne se transforme en 't' par application de la règle qui supprime le 'issaient' final. L'algorithme de Porter a été adapté à de très nombreux langages.
- Les lemmes : un processus appelé lemmatisation consiste à retrouver le mot à l'origine de celui qui est cherché. Ces approches plus riches consistent à utiliser un étiquetage morpho-syntaxique des mots présents dans le document [Namer 2000], ou à des

analyseurs morphologiques comme Nooj [Silberztein 2004] ou Mmorph [Petitpierre 1996]. Mais ces techniques font appel à des traitements lourds, des bases de données volumineuses et nécessitent des réactualisations régulières. Par exemple, le dictionnaire ABU de la langue française contient plus de 300 000 termes [Abu 2005]. Elles sont cependant très performantes.

- Les concepts, termes ou mots-clés : il s'agit d'expressions (pouvant contenir un ou plusieurs mots). Dans le domaine médical, on peut utiliser les ontologies. Cette liste de concepts sera le plus souvent décrite dans un thésaurus (dans le cas des termes, on parlera de terminologie).
- Plus rarement : Les N-grammes : Il s'agit d'une représentation originale d'un texte en séquences de N caractères consécutifs. On trouve des utilisations de bigrammes et trigrammes dans la recherche documentaire (ils permettent de reconnaître des mots de manière approximative et ainsi de corriger des flexions de mots ou même des fautes de frappe ou d'orthographe). Ils sont aussi fréquemment utilisés dans la reconnaissance de la langue d'un texte (par exemple: [Harbeck et Ohler 1999], [Dunning 1994]).
- Les contextes : dans le cas du "Latent Semantic Indexing" ([Deerwester et al. 1990]) les documents et leurs mots sont représentés sur d'autres dimensions où les mots apparaissant dans un même contexte sont proches. Cette indexation est le résultat d'une analyse des co-occurrences des mots dans un corpus (tout comme l'Analyse Factorielle des Correspondances).

Voici quelques exemples de résultats d'indexation sur un document ne contenant que la phrase "Les accidents vasculaires cérébraux" :

Par mots	"Les", "accidents", "vasculaires", "cérébraux"
Par lemmes	"le", "accident", "vasculaire", "cérébral"
Par racines	"accident", "vascul", "cérébr"
Par concepts	"A.V.C." ou "accident vasculaire cérébral"
Par bigrammes	"_l", "le", "es", "s_", "_a", "ac", "cc", "ci", "id", "de", "en", "nt", "ts", "s_", "_v", "va", "as" … "au", "ux", "x_"

Termes, lemmes ou concepts ?

Une des premières difficultés en analyse de données textuelles est de déterminer s'il est plus judicieux de raisonner sur des termes, leurs lemmes, ou sur les concepts. [Besançon 2001] introduit également le terme de 'concept simple', qui consiste à prendre pour chaque terme le premier concept possible, sans désambiguïsation ni analyse. Les résultats montrent que le raisonnement en 'concepts simple' est similaire à la classification des termes. Mais les meilleurs résultats sont obtenus par extraction de concepts, sans toutefois paraître significatifs. Cette analyse montre des écarts faibles entre ces diverses représentations, surtout lorsque l'on améliore le rappel. [Pouliquen 2002] a également utilisé avec succès les concepts dans le domaine médical.

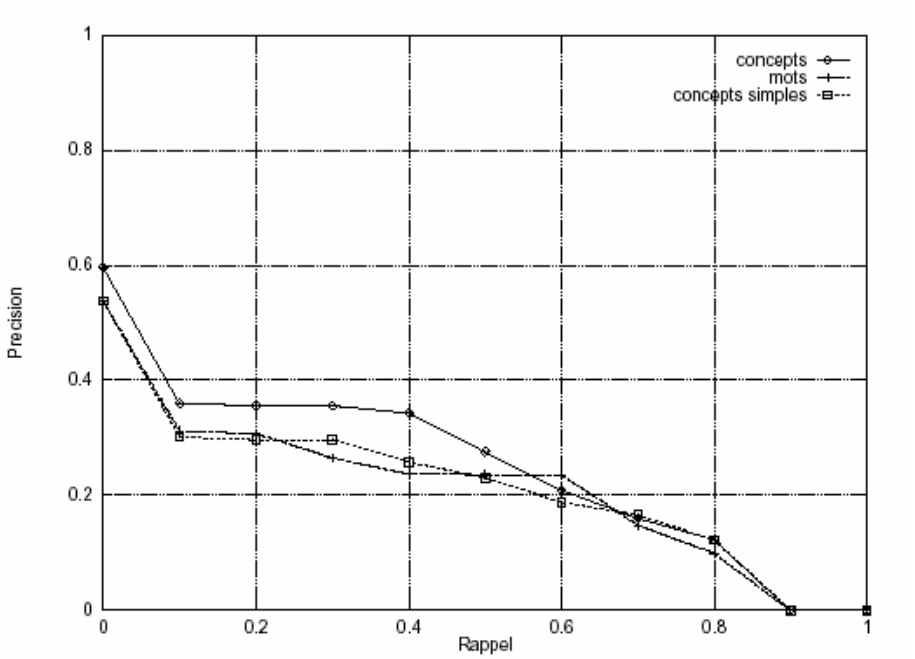


Figure 5. Comparaison d'une classification sur mots ou concepts

Cependant, [Frakes 1992] démontre que l'emploi de lemmes en lieu et place des termes sous leurs formes fléchies n'apporte aucune amélioration significative. [Brun 2003] a d'ailleurs utilisé ces conclusions avec succès dans la mise en œuvre d'un système de reconnaissance automatique de la parole. Enfin, [Lewis 1992b] a montré que l'utilisation du mot comme unité de représentation d'un document semble être adaptée pour des tâches de classification.

La littérature anglophone fait le plus souvent l'éloge de l'indexation par mots, une évaluation a même montré que la racinisation ("Stemming", la plus simple des méthodes linguistiques) n'améliorait pas de manière significative les performances [Frakes 1992, p. 141]. Ce qui a même conduit des auteurs à déclarer: "*S'il est difficile de démontrer l'utilité d'un processus aussi simple que la racinisation, comment pouvons-nous justifier notre intérêt dans des procédés plus ambitieux du traitement du langage naturel*" [Church 1995, p. 310]. Le danger serait de ne se référer qu'à la littérature anglophone, car de multiples autres études ont démontré que pour les autres langues, les processus supplémentaires augmentaient de manière significative les performances des systèmes (pour la racinisation par exemple citons [Sheridan et Ballerini 1996], [Choueka et Zampoli 1992]).
Le choix efficace dépend donc du problème à traiter et pour le traitement des comptes rendus écrits en français, une exploration par concepts est une piste à explorer.

2.8.3 Les méthodes de sélection de descripteurs

Quel que soit le modèle d'apprentissage utilisé, la problématique de sélection de descripteurs se pose, car, avec la représentation en sac de mots, chacun des mots d'un corpus est un descripteur potentiel. Or pour un corpus de taille raisonnable, ce nombre peut être de plusieurs centaines de milliers. En général, il est admis que les mots les plus fréquents peuvent être supprimés : ils n'apportent pas d'information sur le sens d'un texte puisqu'ils sont présents sur l'ensemble des textes. Les mots très rares, qui n'apparaissent qu'une ou deux fois sur un corpus, sont également supprimés, car il n'est pas possible de construire de statistiques fiables à partir d'une ou deux occurrences.

Cependant, même après la suppression de ces deux catégories de mots, le nombre de candidats reste encore très élevé, et il est nécessaire d'utiliser une méthode statistique pour déterminer les mots utiles pour la discrimination entre documents pertinents et documents non pertinents.

Parmi les méthodes les plus souvent utilisées figurent le calcul de l'information mutuelle [Lewis 1992] [Mouliner 1997] [Dumais et al. 1998], la méthode du khi-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]; d'autres méthodes sont également possibles [Sahami 1998] [Zaragoza 1999].

Une comparaison de l'information mutuelle et de la méthode du khi-2 avec d'autres méthodes est effectuée dans [Yang et Pedersen 1997] ; il semble en résulter que l'information mutuelle est légèrement supérieure aux autres.

Une autre approche, appelée latent semantic indexing (LSI) proposée par [Deerwester et al. 1990], consiste à effectuer une décomposition en valeurs singulières de la matrice dont chaque colonne représente un document grâce à un vecteur des occurrences des termes qui le composent. Cette matrice est projetée dans un espace de dimension plus faible où les descripteurs considérés ne sont plus de simples termes. Avec cette méthode, les termes apparaissant ensemble sont projetés sur la même dimension. Cette représentation est censée résoudre partiellement le problème des synonymes et des termes polysémiques. Initialement, cette approche a été utilisée pour effectuer de la recherche d'informations et permet théoriquement de trouver des documents pertinents pour un document candidat, même s'ils ne partage aucun mot. Cette méthode de réduction des dimensions a ensuite été utilisée en entrée des modèles d'apprentissage numérique.

La méthode LSI a été utilisée pour la sélection de descripteurs dans [Wiener et al. 1995] pour sélectionner les entrées d'un réseau de neurones ; la comparaison avec une méthode plus simple de sélection de termes montre très peu de différences, bien que la méthode LSI proposée dans l'article soit légèrement améliorée et implique l'utilisation d'un plus grand nombre de descripteurs. [Schütze et al. 1995] ont également utilisé la méthode LSI pour sélectionner les descripteurs, et ont comparé les résultats obtenus avec une sélection de descripteurs effectuée avec la méthode du khi-2 : la sélection avec la méthode LSI n'améliore pas les performances.

2.8.4 Le nombre de descripteurs retenus

Les méthodes de sélection de descripteurs fournissent, en général, une liste de descripteurs ordonnés du plus important au moins important (la notion d'importance dépend de la méthode de classement considérée) ; il reste ensuite à déterminer combien de descripteurs sont à conserver dans cette liste. Ce nombre dépend souvent du modèle, puisque, par exemple, les machines à vecteurs supports sont capables de manipuler des vecteurs de

grandes dimensions alors que pour les réseaux de neurones, il est préférable de limiter la dimension des vecteurs d'entrées.

Pour choisir le bon nombre de descripteurs, il faut déterminer si l'information apportée par les descripteurs en fin de liste est utile, ou si elle est redondante avec l'information apportée par les descripteurs en début de liste.

Dans son utilisation des machines à vecteurs supports, Joachims [Joachims 1998] considère l'ensemble des termes du corpus Reuters, après suppression des mots les plus fréquents et l'utilisation de racines lexicales (les stems). Il reste alors 9962 termes distincts qui sont utilisés pour représenter les textes en entrée de son modèle. Il considère que l'ensemble de ces termes apporte de l'information, et qu'il est indispensable de les inclure tous. Cependant [Dumais et al. 1998] utilisent également les machines à vecteurs supports mais ils ne considèrent que 300 descripteurs pour représenter les textes. Ils obtiennent néanmoins de meilleurs résultats que Joachims sur le même corpus ; ce qui laisse à penser que tous les termes utilisés par Joachims n'étaient pas nécessaires. Dans leur étude sur la sélection de descripteurs, [Yang et Pedersen 1997] critiquent [Koller et Sahami 1996] qui étudient l'impact de la dimension de l'espace des descripteurs en considérant des représentations allant de 6 à 180 descripteurs. Pour Yang et Pedersen, une telle étude n'est pas pertinente, car l'espace des descripteurs doit être de plus grande dimension ("an analysis on this scale is distant from the realities of text categorization").

À l'opposé, d'autres auteurs considèrent qu'un très petit nombre de descripteurs pertinents suffisent pour construire un modèle performant. Par exemple, [Wiener et al. 1995] ne retiennent que les vingt premiers descripteurs en entrée de leurs réseaux de neurones. Plus récemment, [Stoica et Evans 2000] ont proposé une méthode de sélection de descripteurs pour leur système CLARIT [Evans et Lefferts 1995] et montrent que, pour obtenir des performances optimales avec leur système, 30 termes suffisent en moyenne sur le corpus Reuters.

Entre ces deux ordres de grandeurs, d'autres auteurs choisissent de conserver une centaine de mots en entrée de leur modèle [Lewis 1992] [Ng et al. 2000]. Finalement, il n'est pas prouvé qu'un très grand nombre de descripteurs soit nécessaire pour obtenir de bonnes

performances, puisque, même avec des modèles comme les machines à vecteurs supports qui sont, en principe, adaptées aux vecteurs de grandes dimensions, les résultats sont contradictoires.

2.8.5 Méthodes pour le poids des mots

Il existe plusieurs possibilités pour constituer le vocabulaire et ainsi définir le poids des mots.

Fréquence de mots : A partir de l'ensemble des termes de tous les documents, on comptabilise la fréquence d'apparition de chaque terme.

Fréquence de document des mots : Cette approche diffère de la précédente car on ne comptabilise pas la fréquence des termes, mais on compte la fréquence d'apparition dans les documents (exemple : un terme apparaît dans 40 % des documents).

2.9 Les mesures de distances et de similarités

Les approches vectorielles sont utilisées pour faire de la classification de documents mais servent également comme technique de base pour formaliser la représentation des documents afin que ceux-ci soient exploités par d'autres techniques. Nombre de ces méthodes utilisent des mesures de distances ou de similarité.

Les principales mesures de distances entre deux vecteurs sont les formules de Salton, la fonction cosinus, ou la formule de Jaccard. Ces trois fonctions se basent sur le nombre de propriétés communes ou non. [Rajman 1998] et [Morin 1999] présentent en détail différentes mesures ainsi que leurs propriétés.

Il est important de souligner que nous utilisons le terme de 'distance' par abus de langage car une distance vérifie trois propriétés :

- $d(x,y)=d(y,x)>=0$

- d(x,x)=0
- d(x,y)<=d(x,z)+d(z,y)

Cette troisième propriété, appelée inégalité triangulaire, est rarement vérifiée dans les fonctions de « distance » mises en œuvre.

Si l'on considère Di et Dj, deux documents dont les propriétés sont représentées par des poids. Ces deux documents sont donc représentés par un vecteur de mots pris dans l'ensemble des mots possibles :

Di={poids i1, poids i2, …Poids it} et Dj= {poids j1, poids j2, …Poids jt}

Dans les systèmes binaires, la valeur d'un poids Poids_jk est réduite à 0 ou 1 (terme absent ou présent dans le document). Nous désignons par :

A=le nombre de propriétés partagées par les objets Di et Dj
B=le nombre de propriétés de l'objet Di non partagées par Dj
C=le nombre de propriétés de l'objet Dj non partagées par Di

2.9.1 Mesure de Salton

La mesure de similarité de Salton est la suivante :

$$sim(D_i, D_j) = \frac{\sum_{k=1}^{t} \min(weight_{ik}, weight_{jk})}{\sum_{k=1}^{t} weight_{ik}}$$

Soit, dans un système booléen : sim (Di,Dj)=A/B

2.9.2 Mesure du Cosinus

Dans le cadre du modèle vectoriel classique, les approches utilisant des métriques à base de cosinus sont les plus fréquentes [Salton and Buckley 1990]. Différentes variations de

cette approche ont été implémentées dans le système SMART, bien connu dans le domaine de la recherche documentaire [Salton and Buckley 1988], en particulier par l'approche de la mesure de la dissimilarité.

La mesure de similarité par le cosinus s'exprime ainsi :

$$sim(D_i, D_j) = \frac{\sum_{k=1}^{t} weight_{ik}.weight_{jk}}{\sqrt{\sum_{k=1}^{t} weight_{ik}^{2} weight_{jk}^{2}}}$$

Pour un système binaire, nous avons alors : $sim(Di, Dj) = \frac{a}{\sqrt{bc}}$

Cette mesure a été utilisée efficacement dans [Pouliquen 2002].
La mesure de dissimilarité SMART est :

$$D_i = (w_{i,k})_k$$

avec $w_{i,k} = 0.5.(1 + p_{i,k} / \max_l(p_{i,l})).\log(N / n_k)$ si $p_{i,k} \neq 0$

sinon, $w_{i,k} = 0$

où $w_{i,k}$ est le poids du terme T_k dans le document D_i, p_{ik} est la fréquence relative de T_k dans D_i, N représente le nombre total de documents dans la base documentaire et n_k le nombre de documents contenant le terme T_k.
et atn(Di,Dj)=Di•Dj (dissimilarité SMART atn), où • représente le produit scalaire
et atc(Di,Dj)=cos(Di,Dj) (dissimilarité SMART atc).

2.9.3 La distance de Kullback-Leibler

Dans le domaine du Text Mining, les distributions de cooccurrences de mots-clés peuvent être utilisées comme support pour l'exploration de grandes collections de documents [Feldman1995]. Pour ce type de tâche, une mesure est nécessaire pour quantifier le degré d'intérêt d'une distribution observée par rapport à un modèle donné. Un choix possible est la mesure d'entropie relative (ou « distance » de Kullback-Leibler) qui quantifie le degré de « surprise » associé à l'observation d'une distribution p alors qu'une distribution q était attendue :

$$KL_0(D_i, D_j) = \sum_{k | p_{i,k}.p_{j,k} \neq 0} P_{i,k}.\log(P_{i,k} / P_{j,k})$$

Mais on utilise plus souvent la version symétrisée de la distance de Kullback-Leibler :

$$KL(D_i, D_j) = \sum_{k \mid p_{i,k} \cdot p_{j,k} \neq 0} ((P_{i,k} - P_{j,k}).(\log(P_{i,k} - \log(P_{j,k}))$$

*2.9.4 **Mesure de Jaccard***

La mesure de Jaccard s'exprime comme suit :

$$sim(Di, Dj) = \frac{\sum_{k=1}^{t} \min(weight_{ik}, weight_{jk})}{\sum_{k=1}^{t} weight_{ik}^{\,2} + \sum_{k=1}^{t} weight_{jk}^{\,2} - \sum_{k=1}^{t} weight_{ik}.weight_{jk}}$$

Pour un système binaire, nous avons alors : $sim(Di, Dj) = \frac{a}{a+b+c}$

Avec cette approche, seule la présence ou l'absence de termes est porteuse d'information. Aucune analyse linguistique n'est utilisée, ni aucune notion de distances entre les mots : les documents sont représentés comme des ensembles de mots, désignés sous le vocable « sac de mots » ("bag of words" de [Salton 1983]).

De nombreuses solutions ont été proposées dans la littérature pour coder les composantes des vecteurs, c'est-à-dire pour attribuer un poids à chaque terme ([Salton et Buckley 1990]). Historiquement, le plus connu de ces codages s'appelle tf/idf, et donne parfois son nom à l'approche vectorielle ; cet acronyme signifie : term frequency / inverse document frequency.

Certains auteurs proposent également d'utiliser des fonctions différentes pour coder les termes de la requête et les termes des documents, ainsi qu'une fonction de similarité qui tienne compte des différences de longueurs des documents [Singhal 1996]. Lorsque l'on cherche à trouver quels sont les concepts que l'on trouve à proximité d'autres, les outils de statistiques traditionnelles sont particulièrement bien adaptés. Une mesure issue de la théorie de l'information [Fano 1961] connue sous le nom *d'information mutuelle*, permet de trouver des récurrences contextuelles en comparant la probabilité de trouver deux mots simultanément avec la probabilité de les trouver indépendamment :

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x).P(y)}$$

Dans cette formule, P(x) et P(y) désignent respectivement les probabilités de trouver les termes x et y, et P(x,y) désigne la probabilité de les trouver simultanément. Si les deux mots x et y sont dépendants l'un de l'autre, l'Information Mutuelle a une forte valeur positive. S'ils ne sont pas en relation, la valeur est proche de 0. Enfin, si deux mots sont en distributions complémentaires, la valeur est fortement négative. L'observation des mots x et y n'est pas forcément limitée aux occurrences adjacentes, il est possible de la faire dans une fenêtre textuelle fixe de largeur w ou variable (par exemple selon des règles ou sur des phrases).

Afin de prendre en compte l'ordre des mots dans la phrase, [Church et Hanks 1990] modifient légèrement cette mesure et introduisent le Rapport d'Association (Association Ratio) qui, à la différence de l'Information Mutuelle, n'est pas symétrique. Avec cette mesure, et pour une fenêtre de largeur w=5, ils extraient différents corpus (38, 15 et 8.6 millions de mots) des associations qui sont jugées pertinentes pour établir des concordances en lexicographie.

[Smadja et al. 1996] mettent en évidence que cette mesure accorde autant d'importance à la présence simultanée de deux mots à proximité l'un de l'autre qu'à la non cooccurrence de ces mots. Pour corriger cela, ils introduisent le coefficient de Dice [Dice 1945] :

$$Dice(x,y) = \frac{2P(x,y)}{P(x) + P(y)}$$

Dans cette mesure, si les deux mots x et y sont dépendants l'un de l'autre, le coefficient de Dice a une valeur proche de 1. Dans le cas contraire, le Coefficient de Dice a une valeur proche de 0. Même si le coefficient de Dice est moins sensible aux données de faibles fréquences que l'Information Mutuelle, elle nécessite néanmoins un volume de données important (un million de mots).

Cette mesure de l'Information Mutuelle a été utilisée à des fins de classifications de textes médicaux [Manning 1999] [Loose 1998] et [Romesburg 1990].

2.9.5 Le Khi-2

Usuellement, chaque document est représenté par son profil lexical, i.e. un tuple D_t qui contient les fréquences des unités textuelles dans le document ($D_i = (f_{i,j})_j$, où $f_{i,j}$ est la fréquence de la jème unité dans le document D_t). Le corpus est alors représenté par une matrice T dont la ième ligne est la représentation du ième document. La similarité entre les documents est mesurée par une distance, appelée la distance du Khi-2 (χ^2), très proche de la distance euclidienne (somme des carrés des différences entre les composantes des profils) mais avec une pondération (1/f.j) associée à chacun des termes de la somme.
Plus formellement, on a :

$d_{\chi^2}(D_i, D_{i'})^2 = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}}\right)$, où $f_{i.} = \sum_j f_{i,j}$ et $f_{j.} = \sum_i f_{i,j}$

2.9.6 L'indice d'équivalence

Dans sa thèse sur l'analyse des co-occurrences [Parmentier 98], Parmentier écrit : *"…plus deux termes apparaissent souvent ensemble, plus leur association est grande … plus leur influence réciproque est grande. De plus, un indice d'association entre deux termes ne doit pas augmenter si l'on rajoute à la base un enregistrement ne contenant que l'un des deux termes. Il serait en effet dommageable qu'un tel ajout modifie l'influence d'un terme sur un autre d'une telle façon : l'association de deux termes augmenterait alors que leur co-occurrence ne varierait pas*". Ces constats sont directement issus des travaux de Bayes et de l'algorithme des associations.

Soit Ci le nombre d'occurrences de l'objet i dans cette base de taille N et Cij le nombre d'enregistrements de cette base où les objets i et j co-occurrent. L'indice d'équivalence :

Eij = Cij ² / (Ci x Cj)

"*présente toutes les 'bonnes' propriétés ... : c'est un indice d'association local, homogène, défini par un monôme ...* ". Sachant qu'un indice d'association est homogène s'il reste constant quand on multiplie l'ensemble de ses variables par un facteur constant, et local s'il ne dépend pas de la taille de la base. Cet indice d'équivalence traduit la notion de proximité conceptuelle,

c'est-à-dire que deux termes apparaissant souvent dans le même enregistrement ont de fortes chances d'être liés, conceptuellement parlant.

Mais dans le cas des terminologies médicales, on veut qu'un terme 1 puisse influer sur un terme 2 d'une manière différente que le terme 2 sur le terme 1. En effet, prenons l'exemple d'un diagnostic principal et d'un diagnostic secondaire. Soit A1 le premier et A2 le second. Soit C1 le nombre d'apparitions de A1 dans la base, et C2 celui de A2 dans la même base. Soit C12 le nombre de résumés en commun des deux diagnostics. Donnons des valeurs à ces variables : C1 = 500, C2 = 50, C12= 40. Ce qui nous donne 0.064 pour l'indice d'équivalence. Or, on voit bien que A2 est bien plus lié à A1 que A1 à A2, puisque la presque totalité de ses références sont dans A1. L'indice d'inclusion [Michelet 1988] traduit bien mieux cette notion "d'influence" d'un terme sur un autre :

$Ii \rightarrow j = Cij / Ci$

Ici, $I1 \rightarrow 2 = 40/500 = 8\%$ alors que $I2 \rightarrow 1 = 40/50 = 80\%$. En effet, si le système met à jour A2, il y a une probabilité de 80% (en se basant sur les statistiques de la base d'apprentissage) que A1 se trouve aussi dans la référence à traiter, alors que si le système trouve A1, il n'a que 8% de chances de trouver A2 dans la même référence.

2.9.7 L'information mutuelle

La mesure d'information mutuelle quantifie le lien existant entre un mot et un thème. Plus précisément, elle évalue l'influence qu'a, sur le thème d'un texte, la présence d'un mot dans ce texte. Pour un mot et un thème donnés, elle est évaluée de la façon suivante [Seymore 1998] :

$$I(w_i, T_j) = \log(P(w_i | T_j) - \log(P(w_i))$$

Classiquement, pour un mot donné, on calcule sa valeur d'information mutuelle avec chacun des thèmes. Ensuite, ces valeurs sont combinées afin d'obtenir une valeur unique pour chaque mot. [Yang et Pedersen 1997] montrent que dans ce cas la meilleure façon de combiner consiste à retenir, pour chaque mot, la valeur d'information mutuelle maximale parmi l'ensemble des thèmes.

2.9.8 Le gain d'information

Le gain d'information (également appelé information mutuelle moyenne) [Mitchell 1996], permet, tout comme la mesure d'information mutuelle, de quantifier le lien existant entre un mot et un thème mais ne prend pas seulement en compte l'influence qu'a l'apparition d'un mot sur un thème, il considère également sa non apparition, etc. La mesure de gain d'information se calcule de la façon suivante :

$$IG(w_i,T_j)=\sum_{T\in\{T_j,\overline{T_j}\}}\sum_{w\in\{w_i,\overline{w_i}\}}P(w,T)\log\frac{P(w,T)}{P(w)P(T)}$$

Comme dans le cas de l'information mutuelle, pour un mot donné, on a une valeur par thème traité. Dans ce cas, [Yang & Pedersen 1997] montrent qu'il faut utiliser la moyenne pondérée des valeurs de gain d'information entre le mot et chaque thème.

2.10 Les algorithmes de classification par apprentissage

Les méthodes d'apprentissage se divisent en deux approches principales : l'approche numérique qui utilise des fonctions mathématiques et l'approche symbolique qui utilise plutôt des expressions compréhensibles, telles que les règles de production ou les arbres de décision.

Parmi les méthodes d'apprentissage les plus souvent utilisées figurent la régression logistique [Hull 1994], les réseaux de neurones [Wiener 1993] (et [Wiener et al. 1995]) [Schütze et al. 1995], l'algorithme du perceptron [Ng et al. 2000], les plus proches voisins [Yang et Chute 1994], les arbres de décision [Lewis et Ringuette 1994] [Quinlan 1996] [Apté et al. 1998], les réseaux bayesiens [Lewis 1992] [Lewis et Ringuette 1994] [Joachims 1998] [McCallum et Nigam 1998a] [Sahami 1998], les modèles de Markov Cachés [Zaragoza 1999], les machines à vecteurs supports [Dumais et al. 1998] [Joachims 1998] et plus récemment les méthodes basées sur la méthode dite de boosting [Schapire et al. 1998] [Iyer et al. 2000] ou de bagging. La plupart de ces approches utilisent une représentation des textes en sacs de mots issus du modèle vectoriel.

2.10.1 Les modèles vectoriels

2.10.1.1 Principes

Le modèle vectoriel introduit par [Salton et al. 1975] représente chaque document, ainsi que le document à classifier, par un vecteur. Il s'agit d'un vecteur dont chaque dimension correspond à un terme ou un concept du corpus que l'on trouve généralement affecté d'un poids. Puis on met en œuvre une fonction de calcul de distance (distance Euclidienne) ou de calcul d'un coefficient de similarité (appelé Retrieval Status Value ou RSV) entre le document à classifier et les documents déjà étiquetés. Pour calculer cette similarité, plusieurs formules existent mais la plus fréquente est la fonction Cosine, qui mesure l'angle entre le candidat à étiqueter et chacun des autres documents. Ceci revient à rechercher la plus forte colinéarité entre les deux. Cette fonction présente également l'avantage de retourner une valeur facile à lire et à interpréter : le cosinus retourne une valeur qui est comprise entre 0 et 1 et peut donc s'assimiler par abus de langage comme étant un pourcentage de similarité. Cette technique présente l'avantage de permettre de classer les documents par ordre de pertinence décroissante.

Dans ce modèle, chaque mot du corpus représente une dimension de l'espace et le codage des vecteurs se fait soit par une fonction booléenne, soit par une fonction du nombre d'occurrences d'un mot dans le document. Les composantes des vecteurs peuvent également être des concepts, des paires de mots ou des phrases ; les composantes des vecteurs sont appelées termes dans la terminologie de la recherche d'information. Le choix de l'approche booléenne ou en effectif dépend de l'algorithme utilisé ou du type de problème à traiter.

2.10.1.2 Tf/idf

TF/IDF signifie Term Frequency / Inverse Document Frequency, avec tf/idf(c,d)=tf(c,d)*idf(c,d). Soit :

$$TFIDF_{c,d} = TF_{c,d}.((\log_2 \frac{N}{DF_c}) + 1)$$

Équation 1. Formule de TF IDF

Avec c = un concept, d= document, TF(c,d) la fréquence d'apparition du concept dans le document, DF(c)= le nombre de documents du corpus contenant le concept, et N le nombre de documents du corpus. Ainsi, quand DF(c) est égal à 1 (concept n'apparaissant que dans ce document), le TFDIF sera fort ; quand DF(c) est proche de N (concept apparaissant dans tous les documents), le TFIDF sera faible.

Chaque document *d* est représenté par un vecteur [d] de R^n, où chaque coordonnée d_w est calculée par rapport à la fréquence Occ(w,d) du terme w dans d selon la formule :

$$d_\omega = TFIDF(\omega, d) = \log(1 + Occ(\omega, d)) * \log(\frac{N}{N(\omega)})$$

où N est le nombre de documents du corpus et N(w) est le nombre de documents dans lequel w apparaît au moins une fois. Un terme se voit donc attribuer un poids d'autant plus fort qu'il apparaît souvent dans le document et rarement dans le corpus complet. Chaque vecteur [d] est ensuite normalisé en [$\underline{d}$] afin de ne pas favoriser les documents les plus longs. Pour effectuer la normalisation, nous divisons chaque coordonnée d_w par la norme euclidienne du vecteur :

$$d_\omega = \frac{d_\omega}{\sqrt{\sum_\omega {d_\omega}^2}}$$

2.10.1.3 La formule de Rocchio

Rocchio [Rocchio 1971b] est un des plus vieux algorithmes de classification et l'un des plus simples. Un profil prototypique [c] est calculé pour chaque classe selon :

$$c_w = \frac{t}{N_c} \sum_{d \in c} d_w - \frac{1-t}{N_{\bar{c}}} \sum_{d \notin c} d_w$$

Équation 2. Formule de Rocchio

où N_c est le nombre de documents dans c, $N_{\bar{c}}$ est le nombre de documents n'appartenant pas à c, et t est un paramètre du modèle compris entre 0 et 1. Dans les situations où un document peut être attribué à une seule classe, t est souvent positionné à 1. Ces profils correspondent au barycentre des exemples (avec un coefficient positif pour les exemples de la classe et négatif

pour les autres). Ces vecteurs sont également normalisés de la même façon que les documents.

Le classement de nouveaux documents s'opère en calculant la distance euclidienne (équivalente au produit scalaire et à la similarité en cosinus puisque tous les vecteurs sont de norme 1) entre la représentation vectorielle du document et celle de chacune des classes ; le document est assigné à la classe la plus proche.

La formule de Rocchio est une extension du modèle vectoriel qui permet de trouver les documents répondant à une requête. On transforme automatiquement une requête initiale (représentée par un vecteur noté Q0) en une nouvelle requête (représentée par un vecteur noté Q1) plus performante.

Grâce au modèle vectoriel, un ensemble de documents répondant à la requête initiale est proposé à un utilisateur qui les étiquette (relevance feedback). La nouvelle requête Q1 est construite grâce à la formule de Rocchio [Rocchio 1971], dont l'idée est d'ajouter à la requête initiale les termes des documents pertinents et de lui retrancher les termes des documents non pertinents :

$$Q_1 = \alpha.Q_0 + \beta.\frac{1}{R}\sum_{d|P} d - \gamma.\frac{1}{N-R}\sum_{d|P} d$$

Dans cette formule, les documents sont représentés par un vecteur d, P est l'ensemble des documents pertinents, R son cardinal et N le nombre total de documents étiquetés ; le triplet est choisi en fonction de l'importance que l'on souhaite donner à chaque terme.

S'il existe en plus de la requête initiale une base de documents étiquetés, ces documents sont utilisés comme s'il s'agissait de documents jugés par un utilisateur. S'il n'existe pas de requête initiale, mais uniquement des documents étiquetés comme pertinents ou non pertinents (c'est-à-dire dans le cas de la catégorisation de textes), alors le premier terme est supprimé et une requête Q1 est construite grâce à la formule. Cette formule permet donc d'effectuer également de la catégorisation de textes. Il est également possible de simuler l'interaction d'un utilisateur en postulant que les dix premiers documents trouvés par une première recherche sont pertinents et les cent derniers sont non pertinents (pseudo relevance feedback).

2.10.2 L'approche probabiliste

Les modèles probabilistes sont des applications purement mathématiques de modèles, souvent basés sur les travaux de Bayes, à partir de la représentation vectorielle des documents, parfois sous la forme binaire.

2.10.3 Naïve Bayes

Il s'agit du modèle probabiliste le plus célèbre, qui trouve ses origines dans le théorème de Bayes [Bayes 1763], connu également sous le nom de théorème des probabilités conditionnelles. Si l'on considère A et B deux évènements indépendants, B étant de probabilité non nulle, la probabilité de A sachant B vaut :

Or, $P_B(A) = p(A|B) = \frac{p(A \cap B)}{p(B)}$

P(B) ≠ 0 car si tel était le cas, il serait inutile de calculer la probabilité de A sachant un évènement impossible.

Ce qui nous amène à la loi de Bayes :

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)}$$

Équation 3. Simple Naïve Bayes

soit : p(A ∩ B) = p(A|B). p(B)=p(B|A).p(A)

Si les évènements ne sont pas indépendants, nous devons estimer :

$\arg\max_{C_i}(p(c_i|a_1,...,a_n) \alpha \arg\max_{C_i}(p(c_i).p(a_1,...,a_n|c_i)$

où $c_i \in \{c_1,...,c_k\}$ représente l'ensemble des classes possibles.

L'application de ce théorème en classification de documents est connu sous plusieurs appellations telles que "*Independance Bayes*" ([Todd 1994]), "*Naïve Bayes*" ([Kononenko 1990]) ou "*Idiot's Bayes*" ([Ohman 1988]). L'utilisation du théorème de Bayes à fin de classification est dite naïve (on parle également de idiot's Bayes) car il est fait volontairement

abstraction de l'hypothèse d'indépendance de A et B. Or nous savons que ce n'est pas le cas, bien au contraire. Néanmoins, les résultats obtenus par cette implémentation du théorème de Bayes sont valides et démontrés par [Hand 2001].

Sous cette forme initiale, nous avons l'algorithme dit Simple Naïve Bayes. Plusieurs transformations ont été utilisées, que nous allons d'abord démontrer.

Il est possible d'améliorer la formulation initiale de Bayes. Si l'on cherche à calculer la probabilité de réalisation d'un événement A sachant la réalisation soit d'un événement B soit de son contraire noté $\overline{B}$. Par définition, on pose P(B ou $\overline{B}$)=1 et A = $(B \cap A) \cup (\overline{B} \cap A)$. On peut donc calculer p(A) comme suit :

p(A) = p[$(B \cap A) \cup (\overline{B} \cap A)$]

nous savons que p($A \cup B$) = p(A)+p(B)- p($A \cap B$), donc

p(A) = p $(B \cap A)$+ p($\overline{B} \cap A$) - p($B \cap A \cap \overline{B} \cap A$). Or, il est impossible qu'un événement et son contraire (en l'occurrence B et $\overline{B}$) soient conjointement réalisés, donc

p($B \cap A \cap \overline{B} \cap A$) = 0 d'où :

p(A) = p($B \cap A$)+ p($\overline{B} \cap A$)

Or, d'après la formule des probabilités conditionnelles de Bayes, on a :

p($A \cap B$) = p(A|B).P(B)

En remplaçant B par $\overline{B}$, la formule de la ligne précédente devient :

p($A \cap \overline{B}$) = p(A|$\overline{B}$).P($\overline{B}$)

Par ailleurs, nous avons par définition p(B ou $\overline{B}$)=1 ce qui revient à écrire :

p(B ou $\overline{B}$)=p(B)+ p($\overline{B}$) - p($B \cap \overline{B}$) = 1. Or il est par principe qu'un événement et son contraire soient simultanément réalisés. Donc p($B \cap \overline{B}$) = 0 et par conséquent

p(B ou $\overline{B}$) = p(B)+ p($\overline{B}$) = 1

Nous avions :

p(A) = p($B \cap A$)+ p($\overline{B} \cap A$), ainsi que p($A \cap B$) = p($B \cap A$)= p(A|B).p(B), donc :

p(A)=p(A|B).p(B) + p(A|$\overline{B}$).p($\overline{B}$), avec p(B)+p($\overline{B}$)=1

On améliore parfois le théorème de Bayes en remarquant que

p(B)= p(A$\cap B$)+ p($\overline{A} \cap B$)=p(B|A).p(A) + p(B|$\overline{A}$).p($\overline{A}$), avec p(A)+p($\overline{A}$)=1

Nous pouvons donc réécrire le théorème de Bayes comme suit :

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} = \frac{p(B|A) \times p(A)}{p(B|A)P(A) + p(B|\overline{A})p(\overline{A})}$$

Équation 4. Naïve Bayes complémentaire

où $\overline{A}$ est le complémentaire de *A*.

Si on généralise cette formule avec $\{A_i\}$ l'ensemble des partitions possibles, soit $A= A_1+ A_2+\ldots+A_n$, dont les sous ensembles sont disjoints 2 à 2 ($A_i \cap A_j = 0$), nous obtenons pour tout A_i de la partition :

$$P(B) = \sum_{i=1}^{n} P(B|A_i).P(A_i)$$

Cette loi s'appelle la loi de probabilité totale :

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} = \frac{p(B|A) \times p(A)}{\sum_{i=1}^{|C|} p(B|A_i)P(A_i)}$$

Équation 5. Naïve Bayes avec loi de probabilité totale

Dans l'approche du modèle probabiliste, le coefficient de similarité entre un document étiqueté et un document candidat est la probabilité que le document soit pertinent connaissant le candidat. [Robertson et Sparck Jones 1976] ont proposé un calcul de cette probabilité qui s'appuie sur les calculs de la probabilité qu'un terme soit présent sachant que le document est pertinent et la probabilité qu'un terme soit présent sachant que le document est non pertinent. Pour cela, ils basent leurs calculs sur le PRW (Probabilistic Relevance Weighting), calculé comme suit :

G(c|d)=log(p(c|d)/p(non c|d))

où ‘non c’ représente un document qui n’est pas dans la catégorie (classe) ‘c’. La suite de leur approche est purement fondée sur les théorèmes de Bayes (probabilités conditionnelles). Une application efficace de cette méthode peut être retrouvée dans [Iwayama 1994].

Une description détaillée de cette approche peut être trouvée dans [Sparck Jones 1999]. Ce modèle a donné lieu à beaucoup d'extensions et est à l'origine du système OKAPI qui est l'un des systèmes les plus performants de TREC (avec le modèle vectoriel) dont une description peut être trouvée dans [Robertson et Walker 2000].

[Eyheramendy 2003] compare quatre implémentations de l’algorithme Naïve Bayes en s’appuyant sur quatre lois : Poisson, Bernoulli, multinomiale, et binomiale négative, et trois jeux de données : MDR, newsgroup, et Reuters. MDR est une base de comptes rendus d’anomalies et de dysfonctionnements de matériels médicaux, avec 620179 rapports sur la période 1984-1996. Ces comptes rendus ont été classés selon trois classes en fonction de leurs conséquences (dysfonctionnement, décès, graves blessures). Selon le jeu de données, les résultats varient de 60% à 80% pour la précision, et de 73% à 86% pour le rappel. Les meilleurs résultats sont obtenus avec le jeu de données MDR. Cette comparaison est pertinente car elle montre une grande variabilité des résultats en fonction du type de problème à traiter et de la loi utilisée, bien que sur ce dernier point, la loi multinomiale donne globalement les meilleurs résultats.

2.10.3.1 Le modèle unigramme

Dans le modèle unigramme [McDonough 1994], une distribution de probabilités des mots est calculée pour chaque thème. Ensuite, la probabilité de chaque thème est calculée, le thème correspondant à la probabilité a posteriori la plus élevée sera le thème retenu.

$$P(Tj\left|W_i^{\,N}\right.) = \frac{P(T_j)P(W_i^{\,N}\left|T_j\right.)}{\sum_{k=1}^{j} P(T_k)P(W_i^{\,N}\left|T_k\right.)}$$

Équation 6. Le modèle unigramme

2.10.3.2 Le modèle cache

Le modèle cache de [Bigi 2000] dérive lui aussi d'une distribution de probabilités des mots dans chacun des thèmes, mais également dans le document de test (plus précisément d'une fenêtre cache des mots du test). Ensuite, la distance de Kullback-Leibler symétrique est calculée entre la distribution des mots dans le document de test et celle des thèmes. Le thème retenu sera celui correspondant à la distance la plus faible.

2.10.3.3 La perplexité

La perplexité [Jelinek 1980] est issue du domaine de la reconnaissance de la parole et utilise les chaînes de Markov. La mesure de perplexité permet de mesurer l'adéquation entre un modèle de langage et un document donné. Si l'on développe un modèle de langage par thème et que l'on calcule la valeur de perplexité pour chacun des modèles de langage de thème sur le document de test, alors le thème correspondant à la perplexité minimale sera considéré comme étant celui du thème.

2.10.4 Les modèles numériques

2.10.4.1 Analyse Factorielle des Correspondances et Latent Semantic Indexing

C'est une méthode statistique permettant de comparer des données qualitatives ou classifier les documents. Cette analyse est utilisée, dans la plupart des cas, pour schématiser (sur plusieurs graphiques en deux dimensions) un tableau lexical, c'est-à-dire une matrice représentant, pour chaque mot, sa fréquence d'apparition dans chaque document (appelé aussi table de contingence, tableau croisé).

Historique

Benzécri a publié dès 1973 un ouvrage présentant cette technique d'analyse exploratoire de données multidimensionnelles [Benzécri et al. 1973]. Avant lui, on peut citer [Guttman 1941] et [Hayashi 1956] qui ont, tous les deux, publié des travaux relatifs à ce qu'ils appelaient les "méthodes de quantification".

Principe général

Dans un tableau de données, les méthodes factorielles permettent de calculer des distances entre chaque point-ligne (exemple: un mot ou un concept) et chaque point-colonne (exemple : un document). Dans le cas d'un tableau lexical, nous pouvons exprimer un document comme étant un point dans un espace à n dimensions (chaque mot étant une dimension). De même, un mot peut être exprimé comme un point dans un espace à p dimensions (chaque document étant une dimension). L'interprétation d'un espace à plus de trois dimensions est difficile, voire impossible pour un humain, qui est plus habitué à lire des données dans un espace à deux dimensions (un graphique sur une feuille de papier par exemple).

Le principe de l'analyse factorielle des correspondances (AFC) est justement de faire une représentation de ces deux espaces (n et p dimensions) sur plusieurs graphiques à deux dimensions. Il va de soi, que beaucoup d'informations sont perdues quand on projette tous les points sur un espace plus réduit, mais la méthode permet justement de chercher les meilleurs plans de projection (ceux pour lesquels la perte est minimum). Comme un photographe cherche le meilleur angle de projection pour présenter en deux dimensions des objets du monde réel.

Pour ce faire, l'AFC cherche le premier axe (axe de covariance maximale) appelé axe 1, puis l'axe 2 (orthogonal au premier), etc… Jusqu'à arriver à une perte minimale d'information. Ainsi l'axe 1 et l'axe 2 forment le plan qui contient le plus d'information. Mais cela ne signifie pas que les autres plans (axe 1 et axe 3, axe 2 et 3 …) ne doivent pas être pris en compte,. Ceci fait partie du travail d'interprétation ultérieur à l'AFC. L'originalité de l'AFC est de "laisser parler les chiffres", sans établir a priori de classes bien établies. Comme le disent [Lebart et al. 1995] dans leur préface, grâce aux graphes d'analyse factorielle, "*Benzécri a rendu les individus à la statistique ; longtemps ignorés à force d'être confondus dans de vastes agrégats (…) les individus effectuent leur rentrée sur la scène statistique sous la forme de points dans un nuage*".

Comparaison avec le Latent Semantic Indexing

Le "Latent Semantic Indexing" (LSI) fonctionne de la même manière. On projette les documents et les mots sur chaque axe, et, au lieu de représenter un document par les mots qu'il contient, on le représente par ses coordonnées sur chacun des axes.

Exemple sur un espace à 3x3 dimensions

L'exemple, fictif (tiré de [Pouliquen 2002]), représente le tableau de contingence des mots "foie", "rein" et "pancréas" dans trois documents, l'un parlant d'"hépatomégalie", l'autre de "diabète" et le troisième de "gastroentérologie". Par exemple, le document "hépatomégalie" contient sept fois le mot "foie", deux fois "pancréas" et une fois "rein".

	Hépatomégalie	**Diabète**	**Gastroentérologie**
Foie	7	1	7
Rein	1	4	6
Pancréas	2	8	7

Figure 6. Tableau lexical d'un exemple simple

Chaque document s'exprime dans l'espace des mots, chaque mot est un point dans l'espace des documents.

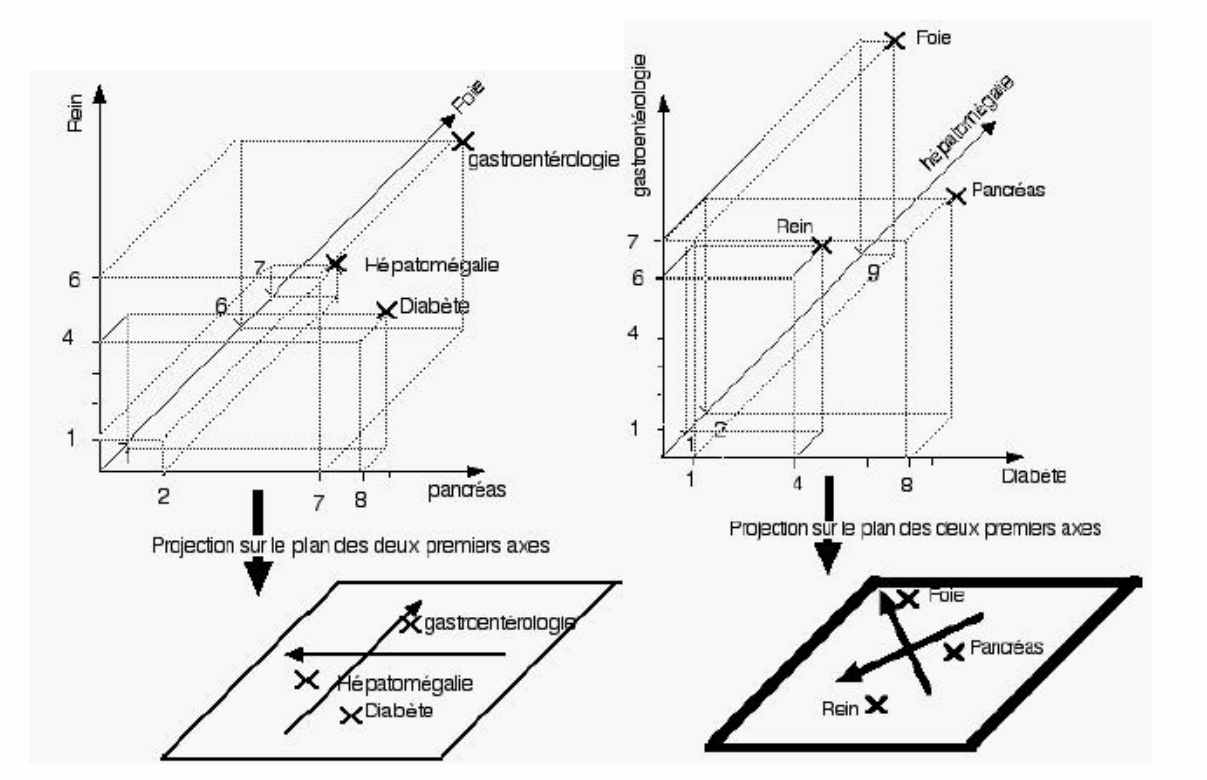


Figure 7. Exemple d'une projection sur un plan dans une AFC (source [Pouliquen 2002])

Les deux projections (à un coefficient près) ont les mêmes barycentres.

On prend l'habitude de représenter simultanément les deux projections :

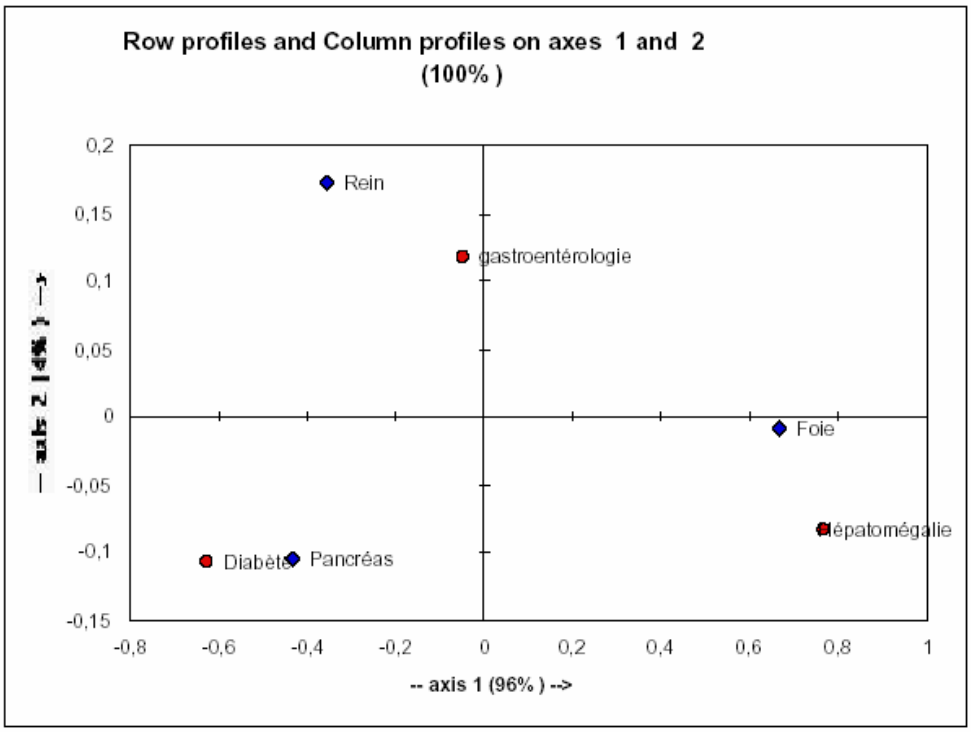


Figure 8. AFC : graphe résultant ([Pouliquen 2002])

Interprétation

C'est la partie la plus délicate de l'analyse des correspondances. Le but est de présenter de manière synthétique l'information. Mais cela ne dispense pas l'utilisateur d'interprétation, ce qui nécessite une bonne compréhension théorique et pratique de la méthode et du domaine. Si, dans notre exemple, seuls deux axes sont nécessaires pour représenter toute l'information, il est très fréquent que les deux premiers axes n'en représentent qu'une partie infime. Il faut alors visualiser et essayer de "comprendre" les autres axes, ce qui ne peut être fait automatiquement.

Ce dernier graphique permet bien d'avoir une vue synthétique d'un espace très complexe. On peut interpréter, les relations existant entre les mots, entre les documents, et les relations mots-documents. Ainsi le document Hépatomégalie est "attiré" par le mot "Foie", "Gastroentérologie" est plus partagé (plus attiré par le mot "Rein" car, dans notre exemple, aucun autre document ne parle particulièrement de rein …).

L'analyse des correspondances ne se justifie pas pour ce genre de données (peu de documents, peu de mots). Mais son intérêt est justement de pouvoir traiter un très grand nombre de

données. Inutile de "démontrer l'efficacité d'un filet de pêche dans un aquarium de salon" [Lebart et al. 1995].

La plupart du temps, l'analyse des correspondances est faite sur un découpage lexical des textes en mots, ce qui fait que l'on ignore délibérément de nombreuses informations de type sémantique ou syntaxique. On ne tient alors pas compte des synonymies ni des homonymies [Lebart et al. 1995, p 145]. Une solution envisageable est de recourir aux lemmes (toutes les flexions d'un mot sont ignorées et remplacées par le terme de référence).
Ceci est un sujet de polémique: En effet, si deux formes fléchies sont réellement utilisées indifféremment, elles apparaîtront proches sur le graphique final, si, par contre, elles sont sémantiquement différentes, elles seront dissociées. Le mot "Cancer" au singulier ne signifie pas tout à fait la même chose que le mot "cancers" au pluriel…

2.10.4.2 Attribution de mots-clés

L'attribution de mots-clés consiste à coder des documents en leur assignant différents mots-clés choisis dans un thésaurus. En médecine, par exemple, chaque patient faisant un séjour hospitalier, aura son dossier codé par des actes et des diagnostics. L'attribution des mots-clés est le plus souvent manuelle (cas des séjours hospitaliers). Parmi les méthodes automatiques d'attribution de mots-clés, on distingue essentiellement deux méthodes : par apprentissage ou par extraction. La première méthode (statistique) calcule, à partir d'un corpus indexé manuellement, une matrice de co-occurrence de chaque mot avec chaque mot-clé. L'attribution de mots-clés à un nouveau texte consistera à calculer la probabilité d'apparition des mots-clés en fonction des mots du texte. Ces méthodes donnent de relativement bons résultats si le corpus est suffisamment important (qualitativement et quantitativement). On peut les utiliser sur différentes langues (par exemple : [Steinberger 2001], [Steinberger et al. 2002]). Mais ces méthodes fonctionnent très mal sur des petits textes, encore moins sur des phrases (une phrase contient trop peu de mots pour que les cooccurrences aient une information sémantique pertinente).

La seconde méthode (linguistique) consiste à décrire chaque mot-clé par des termes (les différentes variantes linguistiques pouvant le représenter), et à essayer de reconnaître ces termes dans le texte. Cette méthode a l'énorme avantage de fonctionner sur de petites unités textuelles. La principale difficulté étant de définir la liste exhaustive des termes associés à chaque mot-clé. De plus, il faut recourir aux outils linguistiques pour reconnaître un terme quelle que soit sa forme syntaxique (par exemple: [Jacquemin et Tzoukermann 1999]).

2.10.4.3 K plus proches voisins (k-PPV ou KNN)

k-PPV (K Plus Proches Voisins, ou Knn pour K nearest neighbours) est un qui a prouvé son efficacité face au traitement de données textuelles [Yang 1997]. La phase d'apprentissage consiste à stocker les exemples étiquetés. Le classement de nouveaux textes s'opère en calculant la distance euclidienne entre la représentation vectorielle du document et celle de chaque exemple du corpus. Les K éléments les plus proches sont sélectionnés et le document est assigné à la classe majoritaire (le poids de chaque exemple dans le vote étant éventuellement pondéré par sa distance). Une variante de cette méthode est utilisée pour la classification automatique. Soit 'k' le nombre de classes à utiliser. On prend k documents initiaux que l'on considère comme centres de classes. On fait une première affectation de l'ensemble des documents à chaque classe, puis on calcule le barycentre de chacune des classes obtenues. Ce barycentre devient le nouveau centre de classe. On réitère le processus jusqu'à ce que le barycentre devienne stable, ou après un nombre d'itérations fixé à l'avance. Cette technique donne de très bons résultats mais impose de connaître à l'avance le nombre de classes. De plus, la principale difficulté consiste à déterminer une fonction de distance pertinente.

2.10.4.4 Support Vector Machine (SVM)

Pour une tâche d'apprentissage donnée, avec un nombre fini de données d'apprentissage, la meilleure performance en généralisation (c'est-à-dire sur des données non observées pendant l'apprentissage) est obtenue si on arrive à trouver un bon compromis entre la nature des données et la capacité du système choisi à apprendre ces données. Un système avec une grande capacité est comme un bachelier qui apprend par cœur mais qui se trompe souvent sur des exemples non appris et un système avec une faible capacité est comme son frère qui n'apprend pas du tout et qui est incapable de bien généraliser.

L'exploration et la formulation de ces concepts ont donné naissance à une théorie dans le domaine de l'apprentissage statistique [Vapnik 1982], d'où sont issues les Machines à Vecteurs Support (MVS). Dans ce qui suit nous allons présenter les MVS, développées par Vapnik, pour une tâche de classification et pour un problème à deux classes.

SVM [Vapnik 1995] est un des algorithmes les plus performants en classification textuelle [Joachims 1998]. L'idée principale est de trouver un hyperplan qui sépare au mieux les données et dont la séparation (ou marge : distance séparant la frontière du plus proche exemple) est aussi grande que possible. Cette recherche correspond à un problème d'optimisation au cours duquel des vecteurs supports (les exemples les plus proches de l'hyperplan) sont sélectionnés. L'hyperplan calculé permet ainsi de séparer l'espace en deux zones. Pour classer les nouveaux documents, on calcule dans quelle région de l'espace ils se situent et on leur attribue la classe correspondante. Plus précisément, la définition de la ligne de décision (celle qui sépare les hyperplans) est définie comme suit :

$\vec{w}.\vec{x} - b = 0$, où le vecteur x correspond au vecteur à classifier et le vecteur w et la constante b résultent de l'apprentissage. Voici un exemple en illustration :

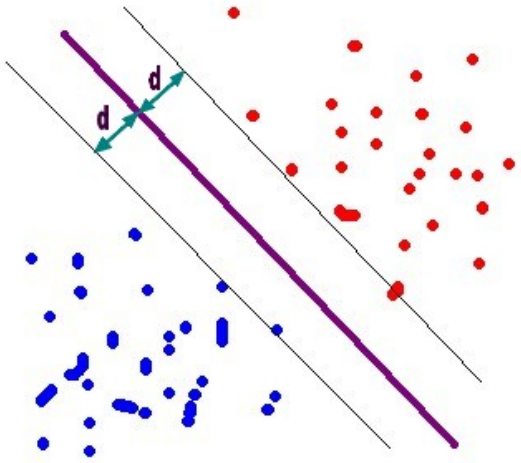


Figure 9. Support Vector Machine

On affecte ensuite le document à la classe dans laquelle il se trouve projeté. Cette méthode est très proche des travaux de Benzécri sur l'ACP [Benzécri et al. 1973] qui présentait cette technique de projection graphique à d'autres fins ...

2.10.4.5 LLFS

LLSF (Linear Least Squares Fit) est une méthode de classification basée sur la fréquence d'apparition des termes et les combinaisons linéaires développée par [Yang et Chute 1994] dont la première application fut la classification de comptes rendus hospitaliers.

[Burnside 2000] ont implémenté cette technique pour la classification automatique de comptes rendus de mammographies. Les résultats obtenus montrent une précision de 83.4% +- 5.3% et un rappel de 35.4% +- 5.6%. Le principe de cette méthode consiste à analyser la fréquence d'apparition d'un terme en fonction du diagnostic. Pour cela, on construit une matrice qui contient en colonne les documents, en ligne les termes, et à l'intersection des deux se trouve le nombre d'occurrences du terme dans le document. Cette matrice s'appelle matrice « A ». A partir d'un lexique médical, on construit une seconde matrice (matrice « B ») avec les documents en colonne et les diagnostics en ligne. A l'intersection des deux, on note 1 ou 0, selon que le document correspond au diagnostic ou pas. La méthode consiste alors à déterminer W tel que WA=B. Lorsque qu'un nouveau document est soumis à classification, on calcule WA, ce qui permet d'obtenir B. On peut néanmoins reprocher à ces applications le fait de n'avoir été réalisées que sur des petites volumétries. Par exemple, Burnside n'a utilisé qu'une centaine de documents pour constituer la matrice B.

2.10.4.6 Les approches neuronales pour la catégorisation de textes

Le perceptron multi-couches

Une approche fondée sur les réseaux de neurones a été proposée dans la thèse de [Wiener 1993] dont les résultats ont été repris dans [Wiener et al. 1995]. Deux architectures neuronales sont proposées et testées sur le corpus Reuters-22173 (qui est une ancienne version du corpus Reuters-21578 disponible aujourd'hui).

La première architecture est un perceptron multi-couche avec une couche de neurones cachés et un neurone de sortie ; un réseau de neurones différent est construit pour chaque catégorie. Les descripteurs sont sélectionnés soit par une méthode de sélection de termes, soit par la méthode LSI, soit par une méthode LSI améliorée (local LSI). Pour la deuxième architecture, les catégories du corpus Reuters sont regroupées en cinq grands ensembles (agriculture, energy, foreign exchange, government, metals). Un réseau est ensuite utilisé pour déterminer à quel ensemble appartient un document, puis cinq réseaux différents sont construits pour déterminer, à l'intérieur d'un ensemble, la catégorie exacte du document. Cette architecture a l'avantage de permettre à chacun des cinq réseaux d'être "spécialisé" et d'utiliser une représentation particulièrement adaptée pour distinguer des catégories proches. Cette deuxième architecture améliore les résultats, mais elle nécessite un découpage manuel

des catégories pour déterminer les ensembles et n'est réalisable que sur un corpus pour lequel le nombre de catégories est connu à l'avance et n'évolue pas. [Schütze et al. 1995] ont également effectué de la catégorisation de textes avec des réseaux de neurones comportant une couche de neurones cachés. Leur modèle est identique au premier modèle utilisé dans [Wiener 1993]; les entrées sont sélectionnées soit par la méthode du khi-2, soit par la méthode LSI.

[Pouliquen 2002] montre que même lorsque le nombre de neurones cachés est nul (le modèle est une simple régression logistique), le modèle peut être surajusté. La mise en œuvre d'une procédure d'arrêt prématuré limite ce surajustement et améliore significativement les résultats.

A partir de ces deux études, il est possible de tirer plusieurs conclusions :

- Malgré ses avantages théoriques, la méthode LSI n'apporte pas d'amélioration sur une méthode de sélection des termes.
- Dans ces études, l'ajout de neurones cachés n'améliore pas les résultats par rapport à une régression logistique.
- Il est nécessaire de se protéger du surajustement, même pour le modèle sans neurone caché, par une méthode de régularisation qui peut prendre la forme d'un terme de pénalisation dans la fonction de coût ou d'une procédure d'arrêt prématuré.

2.10.4.7 La recherche par mots clé

La recherche par mots clés est la plus naturelle. Cette technique de classification consiste à rechercher les termes du document à classifier dans une liste de mots clés pour identifier la classe d'affectation. Cette technique donne de mauvais résultats (F-Mesure de 40%) et présente l'inconvénient de tenir à jour cette liste de mots clés.

2.10.4.8 Chaînes de Markov

Les Modèles de Markov Cachés (MMC) sont des modèles stochastiques qui permettent de décrire ou d'analyser la génération de séquences discrètes. La théorie des MMC a été élaborée vers la fin des années soixante par Baum et al. [Baum 1966] et [Baum 1967]. Ces modèles ont été très étudiés pour la reconnaissance de la parole depuis le début des années soixante-dix [Baker 1975]. Mais ce n'est que plus récemment qu'ils ont été employés

pour l'analyse de la séquence dans des domaines comme la biologie, la reconnaissance de l'écriture manuscrite ou l'analyse de séquences textuelles. Un MMC est défini à partir d'une chaîne de Markov et de densités de probabilités. Une chaîne de Markov est un automate à états et est définie d'une part par un ensemble d'états S = {S1, S2,…, Sl} et d'autre part par une matrice de transitions spécifiant avec quelle probabilité on peut passer d'un état à un autre. Une probabilité de transition entre deux états égale à zéro signifie que cette transition n'est pas autorisée. En spécifiant l'ensemble des transitions autorisées, on obtient différentes topologies de MMC. Les deux structures de MMC les plus répandues sont les MMC ergodiques et les MMC du type Bakis [Bakis 1976]. Dans le cas ergodique, tous les états du modèle sont reliés entre eux. Pour certaines applications, notamment le traitement de signaux temporels comme la parole ou l'écriture, d'autres types de MMC, comme les MMC gauche-droite du type Bakis se sont montrés plus efficaces. La propriété fondamentale des MMC gauche-droite est qu'aucune transition n'est autorisée sur les états dont l'indice d'état est inférieur à celui de l'état courant. Les MMC utilisés en classification supervisée de document sont généralement utilisés pour optimiser la désambiguïsation des termes, en utilisant des CMM gauche-droite. [Brun 2003] a utilisé les MMC pour la reconnaissance automatique de la parole.

[Besançon 2001] a utilisé les chaînes de Markov pour intégrer les sens des mots dans un modèle probabiliste de représentation de textes, en vue d'optimiser la désambiguïsation des termes et améliorer leur rattachement à des concepts. Selon l'auteur, les premiers résultats pour l'évaluation de cette représentation pour une tâche de recherche documentaire montrent une amélioration des performances.

2.10.4.9 Algorithme EM

Une présentation très détaillée de l'algorithme EM se trouve dans [Amini 2001]. Afin de déterminer les paramètres Θ d'un modèle qui explique au mieux un ensemble d'observations *X*. On peut chercher à maximiser la vraisemblance du mélange (ou *LM*) suivant Θ. En général, il n'est pas possible de résoudre explicitement :

$$\frac{\delta L_M}{\delta \Theta} = 0$$

Une manière de maximiser *LM* est d'utiliser une classe générale de procédures itératives connues sous le nom d'algorithme EM [Dempster 1977]. Cet algorithme est l'un des

algorithmes d'optimisation les plus employés en statistiques. Parmi les applications classiques qui utilisent cet algorithme on peut citer les MMC et l'estimation des paramètres d'un mélange de densités. En général avec cet algorithme on ne peut trouver qu'un maximum local.

Le principe de cet algorithme est le suivant : A chaque itération, les valeurs de Θ sont ré-estimées de façon à accroître *LM* et ceci jusqu'à ce qu'un maximum soit atteint. [Bishop 1995], [McLachlan 1997], [Mitchell 1997], [Webb 1999] ont étudié en détail les différents points de l'algorithme EM, nous n'allons nous intéresser ici qu'à sa forme générale et nous donnerons une preuve de sa convergence. L'idée principale de l'algorithme EM est d'introduire des variables cachées *Z* de façon à ce que, si les *Z* étaient connus, la valeur optimale de Θ pourrait être trouvée facilement.

Ainsi, on peut écrire :

$$L_M\Theta = pX\Theta = \sum_Z pXZ\Theta = \sum_Z p(X/Z,\Theta)\,p(Z/\Theta)$$

[Nigam 1999] a utilisé l'algorithme EM pour améliorer la classification de documents, surtout dans un contexte de documents non étiquetés. La comparaison avec un classifieur de type Naïve Bayes montre un gain de performances (précision et rappel) variant de +1 à +5%, avec une moyenne de l'ordre de +2%.

2.10.4.10 Maximum de vraisemblance de classification

Il existe une autre approche pour faire de la discrimination en mode non-supervisé appelé le *maximum de vraisemblance de classification*. Dans cette approche les vecteurs indicateurs de classes *tk* sont traités comme des paramètres manquants et on maximise directement le logarithme de la vraisemblance de classification.

Ce critère est maximisé grâce à l'algorithme de classification EM (CEM) [McLachlan 1992] et [Celeux 1992]. Il est similaire à l'algorithme EM excepté une étape **C** de classification dans laquelle, à chaque donnée, est assignée une et une seule composante du mélange.

Dans ce cadre, Krishnan et Nardy [Krishnan 1987] et Titterington [Titterington 1987] ont proposé une méthode d'estimation de paramètres d'un mélange de densités pour un

problème de classification. Nous décrivons ici leurs travaux. Ils considèrent dans leur étude qu'à chaque forme est associée une étiquette incertaine, les vraies étiquettes des exemples étant inconnues. En particulier, [Krishnan 1987] considère en détail un problème à deux classes où à chaque forme *xi* (de dimension *p*) est associée l'hypothèse *zi* (*zi* $\in$ [0,1]). *Zi* représente la croyance du superviseur imparfait en l'appartenance de l'exemple *xi* à la classe 1.

2.10.5 Les arbres de décision

L'arbre de décision est une technique d'apprentissage symbolique qui peut être utilisée à des fins de classifications. [Grappa] explique les principales techniques pour créer les arbres. Parmi les algorithmes connus, citons CART (Classification And Regression Tree) et ID3, qui fut amélioré par C4.5.

On trouve plusieurs utilisations des arbres de décision à des fins de classification de données textuelles, en particulier dans [Lewis et Ringuette 1994] [Quinlan 1996] [Apté et al. 1998]. [David 2001] a utilisé cette technique sur la base Reuters, obtenant 80.5% de rappel et 87% de précision.

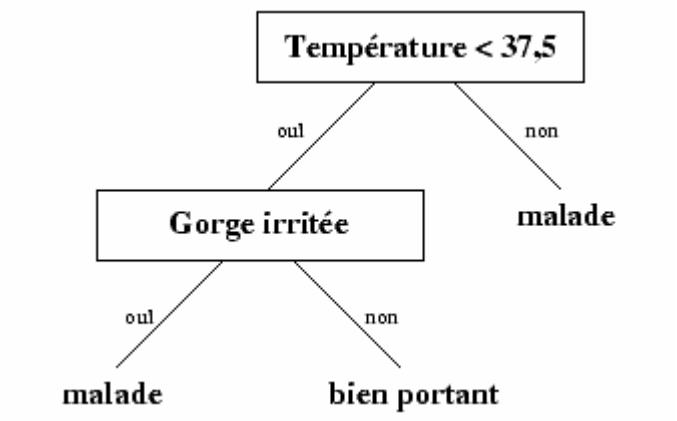


Figure 10. Arbre de décision

2.10.6 Optimisation

Certaines techniques permettent d'optimiser les résultats obtenus à partie des algorithmes classiques. Les meilleurs gains sont obtenus par l'ajout de connaissances depuis un système d'information externe.

2.10.6.1 Probabilistic word vector

[Satoshi 2003] utilise un dictionnaire pour construire les vecteurs. L'expérimentation, basée sur TF/IDF, ne se fait pas à partir des termes du document initial mais à partir des termes enrichis de ceux du dictionnaire, ce qui lui permet d'ajouter des coefficients sur les termes associés et synonymes. Les résultats présentés montrent une forte amélioration des performances.

2.10.6.2 Le vote majoritaire

Une des voies explorées depuis plusieurs années pour améliorer les capacités de généralisation des systèmes d'apprentissage consiste à combiner différents systèmes pour une même tâche. Nous allons nous intéresser ici à une famille de techniques qui font coopérer des systèmes identiques entraînés sur des ensembles d'apprentissage différents. Elles ont été développées vers le milieu des années 90, et ont connu un succès notable en apprentissage. Les deux représentants les plus connus sont le *Bagging* et le *Boosting*.

Notre intérêt pour ces techniques naît du fait que le Boosting a été employé avec succès en classification de textes [Schapire 2000]. Plusieurs appellations sont utilisées pour désigner les algorithmes de Boosting et de Bagging. Dans la littérature on parle soit d'algorithmes à base de vote (*Voting algorithms*) soit de moyennes pondérées (*weigthed average algorithms*). Nous nous conformerons à la terminologie utilisée par les auteurs. Ces méthodes comparées à celles sans vote réduisent d'une manière significative l'erreur en généralisation.

Les méthodes de vote peuvent être divisées en deux grandes familles : celles qui changent la distribution de probabilité $p(x)$ de l'ensemble d'apprentissage en se basant sur le

comportement des classifieurs précédents (les méthodes de *Boosting*) et celles qui ne le font pas (comme le *Bagging*). Chaque méthode utilise un algorithme de base (*weak algorithm*) et une base d'apprentissage. Le principe commun consiste à faire tourner cet algorithme plusieurs fois. Dans le cas du Boosting, la distribution des exemples de la base d'apprentissage est changée à chaque itération, tandis que dans le cas du Bagging les différentes bases d'apprentissage sont formées à l'avance. Les classifieurs générés sur chaque base sont alors combinés pour créer un classifieur final qui est utilisé sur la base de test.

2.10.6.3 Bagging

L'algorithme de Bagging (**B**ootstrap **agg**regat**ing**) a été introduit par Breiman [Breiman 1996]. Une présentation très détaillée se trouve dans [Amini 2001]. Cet algorithme se déroule en deux étapes :

1. Par bootstrap, on génère d'abord B bases, $D_l^1, D_l^2, \ldots D_l^B$ à partir de la base d'apprentissage initiale D_l, (un ensemble de bootstrap [Efron 1993] est généré en échantillonnant N exemples d'une base d'apprentissage par un tirage avec remise).
2. On entraîne ensuite B classifieurs Cl_1, Cl_2, …, Cl_B sur chacune de ces bases. Le classifieur final Cl^* est construit en combinant les Cl_i, la classe prédite par la combinaison est celle qui est majoritaire parmi ces classifieurs de base.

Pour une base de Bootstrap donnée, un exemple issu de la base d'apprentissage originale (de taille N) a une probabilité de $1-(1-1/N)^N$ d'être sélectionné aléatoirement au moins une fois. Pour N assez grand, cette probabilité est de l'ordre de $1-1/e = 63.2\%$. Ce qui signifie que chaque base de bootstrap contient seulement 63.2% d'exemples de la base d'apprentissage apparaissant au moins une fois.

Dans le cas d'algorithmes de base instables comme la descente de gradient stochastique, si les classifieurs ne sont pas corrélés la performance globale peut être améliorée [Breiman 1994]. Dans le cas d'algorithmes stables la performance globale peut se dégrader dans le cas où la taille de la base d'apprentissage serait trop limitée [Breiman 1996].

Entrée : Une base d'apprentissage D_l, un modèle de base M_w, un entier B (nombre d'échantillons bootstrapés).

Pour i = 1 à B faire

{

1. D_l^i = échantillon bootstrapé de D_l (tirage avec remise)
2. $Cl_i = M_w(D_l^i)$

}

$$Cl^*(x) = \arg\max_{t \in T} \sum_{i:Cl_i(x)=t} l$$

1 max arg (l'étiquette la plus souvent prédite)

Sortie : classifieur *Cl**

Breiman [Breiman 1996] explique que l'amélioration des performances est due en partie à la diminution de la variance. Mais cette affirmation a fait l'objet de controverses notamment par [Schapire 1997]. Il existe néanmoins un excellent papier de [Bauer 1999] qui tire la même conclusion que Breiman. Il s'agit donc encore aujourd'hui d'une question ouverte.

2.10.6.4 Boosting

[Schapire 1990] Schapire a introduit l'algorithme de Boosting au début des années quatre-vingt dix. Cet algorithme est considéré comme l'une des méthodes les plus performantes pour accroître les performances d'un classifieur de base. Après les améliorations apportées par Freund [Freund 1990], Freund et Schapire [Freund 1995] ont proposé l'algorithme d'AdaBoost (**Ada**ptive **Boost**ing) qui est l'un des algorithmes de boosting les plus populaires aujourd'hui. Une présentation très détaillée se trouve dans [Amini 2001]. Dans cette partie nous allons nous intéresser aux deux variantes de cet algorithme : AdaBoost.M1 et AdaBoost.M2 ([Freund 1996]).

[Schapire 2002] a montré comment l'ajout préalable de connaissances sur un domaine permet d'améliorer les performances des systèmes de classification. L'expérimentation sur quatre jeux de données a donné des résultats d'autant plus significatifs que les données d'apprentissage étaient réduites. Cette technique, appelée 'boosting', sera donc particulièrement recommandée pour les jeux de données réduits, ou pour optimiser les algorithmes.

Comme le Bagging, l'algorithme d'AdaBoost génère un ensemble de classifieurs et les combine par une méthode à base de votes. A la différence du Bagging les classifieurs sont entraînés séquentiellement ; le *ième* classifieur généré prend en compte les erreurs des classifieurs déjà construits. Ceci peut être fait par exemple en ré-échantillonnant la base d'apprentissage : un exemple mal classé par un classifieur va être affecté d'un poids plus élevé qu'un exemple bien classé. De cette façon les exemples mal classés ont plus de chance d'être présents dans la nouvelle base d'apprentissage. Le but est de forcer le classifieur à se focaliser sur les exemples difficiles. Un premier algorithme adaboost est décrit ci-dessous.

Entrée : une base d'apprentissage $D_l = \{x_i, t_i\}_{i=1,..,n}$, un modèle de base Mw, un entier I (nombre d'itérations).

1. $D_l^{(1)}=D_l$ avec le poids de chaque exemple égal à 1.

Pour i=1 à I faire

{

2. $D_l^{(i)}=Mw(D_l^{(i)})$ (Apprendre les paramètres du classifieur Ci à partir de la base $D_l^{(i)}$

3. $\varepsilon_i = \frac{1}{n} \sum_{x_j \in D_l^{(t)} / Cl(x_j) \neq t_j} poids(x_j)$

4. Si $\varepsilon_i \geq ½$, poser i=I+1 et sortir de la boucle

5. $\beta_i = \varepsilon/(1-\varepsilon_i)$

6. $\forall xj \in D_l^{(i)}, si\ Cl_i(x_j) \neq t_j$ alors $\text{poids}(x_j) = \text{poids}(x_j).\beta_i$

7. Construire la distribution de probabilité p(x) à partir du poids des exemples.

8. Construire la base $D_l^{(i+1)}$ en tirant les exemples suivant la distribution p(x).

}

$$Cl^*(x) = \arg\max_{t \in T} \sum_{i:C_i(x)=t} \log\frac{1}{\beta i}$$

Sortie : classifieur Cl*

Le principe de cet algorithme est le suivant : pour un nombre d'itérations **I**, on construit **I** bases d'apprentissage pondérées $D_l^{(1)}$, …, $D_l^{(I)}$ et **I** classifieurs Cl_1, Cl_2, …, Cl_I. Le classifieur final *Cl** est formé en utilisant un schéma de vote pondéré : le poids de chaque classifieur dépend de ses performances sur la base d'apprentissage utilisée pour l'entraîner. Les deux propriétés fondamentales de l'algorithme AdaBoost.M1 sont :

1. Les exemples mal classés à l'étape *i* sont pondérés par un facteur inversement proportionnel à l'erreur du classifieur sur la base d'apprentissage, i.e., $1/(2\varepsilon_i)$.
2. La proportion des exemples mal classés est ε_i, et le poids de ces exemples est amplifié par un facteur $1/(2\varepsilon_i)$. Ainsi le poids total des exemples mal classés après mise à jour sera la moitié du poids des exemples de la base d'apprentissage original. On peut ainsi avoir peu d'exemples mal classés ayant des poids élevés.

Le problème avec cet algorithme provient du fait que dès que le classifieur de base réalise une erreur $\varepsilon_i >= ½$ avec une base d'apprentissage donnée, l'algorithme s'arrête. Or ce cas de

figure se manifeste fréquemment avec un mauvais classifieur ou avec une base d'apprentissage bruitée.

A chaque itération *i* l'algorithme d'AdaBoost.M2 génère un classifieur $h : X \times T \rightarrow [0,1]$ qui donne une estimation de la probabilité a posteriori des classes *t* sachant la forme d'entrée *x*. Les sorties du classifieur *h* sont interprétées de la manière suivante :

Si $h_i(x_k,t_k) = 1$ et $h_i(x_k,t) = 0$, $\forall t \neq t_k$, alors h_i a prédit correctement l'étiquette de x_k.

Si $h_i(x_k,t_k) = 0$ et $h_i(x_k,t) = 1$, alors h_i a donné la prédiction opposée.

Si $h_i(x_k,t_k) = h_i(x_k,t)$ alors la classe de x_k est choisie d'une manière aléatoire entre *t* et t_k.

Cette interprétation a conduit à la définition de la *pseudo-perte* (*pseudo-loss*) du classifieur h_i en regard de la distribution Dis_i définie par :

$$\varepsilon_i = \frac{1}{2} \sum_{(k,t)\in\Psi} Dis_i(k,t)(1 - h_i(x_k,t_k) + h_i(x_k,t))$$

où Ψ est l'ensemble des exemples mal classés. Ainsi, au lieu d'utiliser la prédiction de l'erreur usuelle, la distribution de probabilités $p(x)$ est obtenue sur la pseudo-perte. En utilisant ce critère, l'algorithme de Boosting se focalise non seulement sur les exemples difficiles mais plus spécifiquement sur les classes qui sont plus difficiles à discriminer.

Chaque paire mal classée (k,t) est intuitivement interprétée comme la réponse à une question binaire de la forme : « *Est-ce que l'étiquette associée à l'exemple* x_k *est* t_k *(l'étiquette correcte) ou t (l'étiquette incorrecte) ?* » Avec cette interprétation, le poids $Dis_i(k,t)$ donne l'importance d'une étiquette incorrecte *t* pour l'exemple x_k.

Un résultat surprenant des algorithmes de Boosting est que l'erreur en généralisation décroît même lorsque l'erreur en apprentissage devient nulle. En utilisant la décomposition de l'erreur en biais et en variance, [Schapire 1997] montre que l'algorithme de Boosting combine deux effets :

1. il réduit le biais du classifieur de base en le forçant à se concentrer sur différentes parties de l'espace des exemples,
2. il réduit aussi la variance du classifieur de base en moyennant plusieurs hypothèses générées sur différents sous-ensembles de la base d'apprentissage.

[Schapire 1997] explique que la décroissance de l'erreur en généralisation est liée à la distribution de la marge des exemples en apprentissage. Ils démontrent qu'en se focalisant sur

les exemples les plus difficiles, les algorithmes de Boosting tentent de maximiser la marge de tous les exemples en apprentissage. En tenant compte soit de l'erreur soit de la pseudo-erreur, les différences entre le Bagging et le Boosting peuvent se résumer comme suit :

1. Bagging génère des classifieurs de bases en parallèle tandis que Boosting les génère séquentiellement,
2. Le Bagging utilise le pré-échantillonnage et le Boosting la pondération,
3. Pour le classifieur final, le Bagging donne un poids égal à chaque classifieur de base, alors que le Boosting les pondère.

2.11 Comparaison des méthodes

Pour faciliter les comparaisons entre les différents algorithmes, il fallait trouver des référentiels communs. Ainsi, nous avons vu rapidement les chercheurs se focaliser sur deux aspects : utiliser un corpus commun et une mesure de performance identique.
Plusieurs corpus sont utilisés, mais le plus connu est la base Reuters (base de données de dépêches de presses), et plus rarement Ohsumed (base de textes médicaux) ou NG20 (groupes de discussions).

En ce qui concerne les mesures de performances, on trouve essentiellement la F1 mesure de [Van Rijsbergen 79]. Préalablement, on utilisait surtout le BEP.

2.11.1 Analyse des corpus Reuters

Le problème de la base Reuters est que de nombreux chercheurs ont utilisé des versions et variantes différentes de ce corpus. En reprenant un important volume de documents de recherche, nous avons pu identifier 12 versions différentes de ce corpus :

N°	Version	Préparé par	Nombre de catégories	Nombre de documents en apprentissage	Nombre de documents en test	Nombre de documents	Documents étiquetés
1	Version 1	CGI	182	21450	723		80%
2	Version 2	Lewis	113	14704	6746		42%
3	Version 2.2	Yang	113	7789	3309		100%
4	Version 3	Apté	93	7789	3309		100%
5	Version 4	PARC	93	9610	3662		100%
6	Reuters1	Lam	135	14704	6746	21450	
7	Reuters2	Apté	93	10667	3680	14704	
8	Reuters3	Wiener	92	9610	3662	13272	
9	Reuters4	Dumais	90	9603	3299	12902	
10	Reuters5	Joachims	10	9603	3299	12902	
11	ModLewis	Lewis	135	7437	6188	13625	
12	Reuters6	Cai	50	9603	3299	21578	

Figure 11. Les différentes versions du corpus Reuters

2.11.2 Quelle est la meilleure méthode ?

Il paraît difficile de répondre à cette question tant la comparaison s'avère impossible. De nombreux algorithmes ont été implémentés pour la catégorisation de textes, mais une des questions récurrentes est : quelle est la meilleure méthode pour cette tâche ? Il existe, en pratique, plusieurs méthodologies pour tenter de répondre à cette question.

La première consiste à comparer différentes méthodes mises en œuvre par différents auteurs sur le même corpus. L'inconvénient de cette méthode est qu'il faut que tous les auteurs utilisent exactement le même découpage du corpus. Nous venons de voir qu'il existe différentes versions du corpus Reuters, mais même lorsqu'on utilise le même corpus, il ne l'est pas toujours dans sa totalité. Pour le corpus Reuters-21578, qui est souvent utilisé, certains auteurs considèrent 90 catégories [Joachims 1998], [Schapire et al. 1998], [Yang et Liu 1999], d'autres en considèrent 118 [Dumais et al. 1998]. De plus, la plupart des auteurs considèrent 3299 documents sur la base de test, mais [Yang et Liu 1999] en considèrent uniquement 3019 en supprimant tous les documents de la base de test qui n'appartiennent à aucune catégorie.

Finalement, ces légères différences de découpage rendent difficiles les comparaisons à travers ces publications. De plus, tous les auteurs n'utilisent pas les mêmes mesures de performances, et peuvent calculer les moyennes de manières différentes. Enfin, même dans le cas où les auteurs utilisent les mêmes mesures, il est nécessaire d'utiliser des tests statistiques pour vérifier que les différences ne sont pas dues au hasard [Hull 1993].

Une autre approche souvent proposée est l'utilisation de plusieurs méthodes par le même auteur ; de cette manière, le découpage et les mesures sont identiques pour toutes les méthodes. [Yang et Liu 1999] comparent ainsi les machines à vecteurs supports, les plus proches voisins, les réseaux de neurones, une combinaison linéaire, et des réseaux Bayésiens. Le résultat de leur comparaison est le suivant :

Méthode	RAPPEL	PRECISION	F-MESURE
SVM	81.20%	91.37%	85.99%
KNN	83.39%	88.07%	85.67%
LLSF	85.07%	84.89%	84.98%
Neural Network	78.42%	87.85%	82.87%
Naïve Bayes	76.88%	82.45%	79.56%

Figure 12. Comparaison des différentes méthodes de classification

Il apparaît que, hormis Naïve Bayes, les résultats sont sensiblement proches.

[Dumais et al. 1998] proposent également une série de comparaisons en mettant en compétition une variante de l'algorithme de Rocchio (appelée find similar), des arbres de décision, des réseaux Bayésiens et des machines à vecteurs supports :

	Findsim	NBayes	BayesNets	Trees	LinearSVM
Earn	92.9%	95.9%	95.8%	97.8%	98.0%
Acq	64.7%	87.8%	88.3%	89.7%	93.6%
money-fx	46.7%	56.6%	58.8%	66.2%	74.5%
Grain	67.5%	78.8%	81.4%	85.0%	94.6%
Crude	70.1%	79.5%	79.6%	85.0%	88.9%
Trade	65.1%	63.9%	69.0%	72.5%	75.9%
Interest	63.4%	64.9%	71.3%	67.1%	77.7%
Ship	49.2%	85.4%	84.4%	74.2%	85.6%
Wheat	68.9%	69.7%	82.7%	92.5%	91.8%
Corn	48.2%	65.3%	76.4%	91.8%	90.3%
Avg Top10	64.6%	81.5%	85.0%	88.4%	92.0%
Avg All Cat	61.7%	75.2%	80.0%	N/A	87.0%

Figure 13. Comparaison des méthodes [Dumais 1998]

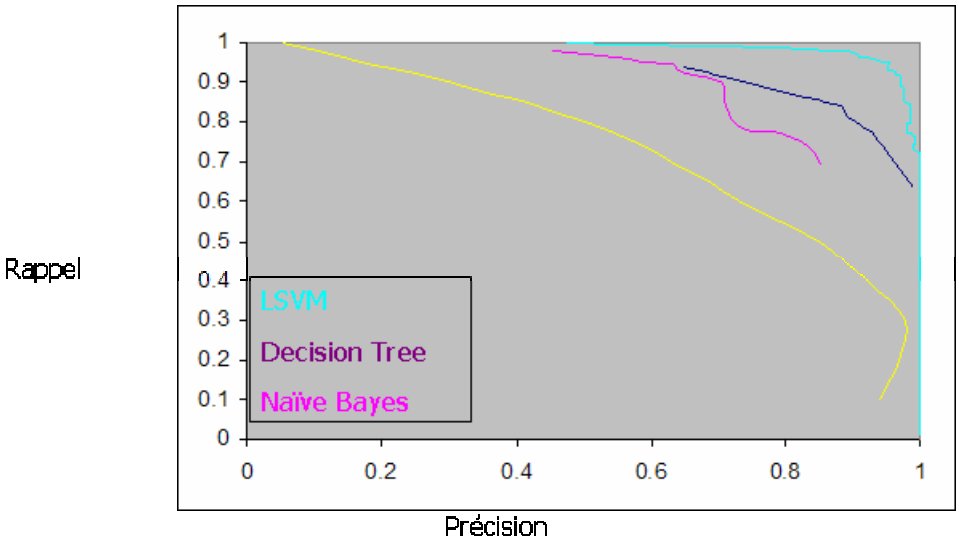


Figure 14. Evolution du rappel et de la précision depuis [Dumais 1998]

Il faut également intégrer que toutes ces méthodes sont délicates à mettre en œuvre et leurs performances dépendent fortement des algorithmes utilisés.

La conférence TREC est une conférence annuelle dont l'objectif est de mettre en compétition des algorithmes sur les mêmes problèmes. Il s'agit probablement d'une bonne solution pour comparer dans le temps différentes méthodes, car chaque participant propose des solutions qu'il connaît bien avec des algorithmes dont il a pu tester l'efficacité. Le corpus est évidemment identique pour tout le monde, ainsi que les méthodes d'évaluation et la répétition annuelle de cette conférence permet de juger les approches sur le long terme. De plus la conférence TREC a l'avantage de proposer un état de l'art à un instant donné contrairement aux comparaisons faites à partir des publications pour lesquelles le décalage dans le temps peut rendre certaines conclusions obsolètes.

Méthode de détection de thème	**Méthode de sélection de vocabulaire**	**Taille de vocabulaire**	**Performances**
Unigramme	Fréquence de document	30K	83.1%
TFIDF	Fréquence de mots	30K	74.3%
Cache	Fréquence de mots	34K	82.5%
Perplexité	Fréquence de mots	64K	79.0%
SVM	Information mutuelle	40K	78.3%

Figure 15. Comparaison de performances selon [Brun 2003]

Mais ces documents sont, selon l'inventeur du web, "*destinés aux humains plutôt que des données et informations qui peuvent être analysées automatiquement*" [Berners-Lee et al. 2001], le défi est justement d'extraire automatiquement de l'information de ces documents écrits en langage naturel. "*La puissance de la langue naturelle crée en même temps un obstacle à son utilisation pour le traitement de l'information*" [Zweigenbaum 1999].

2.11.3 Comparatif des performances et conclusion

L'analyse de différentes publications utilisant la base Reuters nous a conduits au tableau suivant :

Type	Système	Auteur	Reuters1	Reuters2	Reuters3	Reuters4	Reuters5	Reuters6
Batch Linear	FINDSIM	[Dumais 1998]				0.617	0.646	
	ROCCHIO	[Cohen et Singer 1999]	0.660	0.748		0.776		
		[Joachims 1998]					0.799	
		[Lam et Ho 1998]				0.781		
		[Li and Yamanishi 1999]				0.625		
Bayesian Network		[Dumais 1998]				0.800	0.850	
		[Lam et al 1997]	0.542					
Committee	ADABOOST.MH	[Schapire et Singer 2000]		0.860				
		[Weiss et al 1999]				0.878		
Decision Rules	CHARADE	[Moulinier et al 1996]		0.783				
		[Moulinier et Ganascia 1996]		0.738				
	DL-ESC	[Li and Yamanishi 1999]				0.820		
	RIPPER	[Cohen et Singer 1999]	0.683	0.811		0.820		
	SLEEPING EXPERT	[Cohen et Singer 1999]	0.753	0.759		0.827		
	SWAP-1	[Apté et al 1994]		0.805				
Decision trees		[Dumais 1998]					0.884	
	C4.5	[Joachims 1998]					0.794	
	IND	[Lewis et Ringuette 1994]	0.670					
Example based Knn	GIS-W	[Lam et Ho 1998]				0.860		
	KNN	[Joachims 1998]					0.823	
		[Lam et Ho 1998]				0.820		
		[Yang 1999]	0.690	0.852	0.820			
		[Yang et Liu 1999]				0.856		
Neural Network		[Wiener et al 1995]			0.820			
	CLASSI	[Ng et al 1997]		0.802				
	NNET	[Yang et Liu 1999]				0.838		
On line linear	BALANCED WINNOW	[Dagan et al 1997]	0.747	0.833				
	WIDROW-HOFF	[Lam et Ho 1998]				0.822		
probabiliste		[Dumais 1998]				0.752	0.815	
		[Joachims 1998]					0.720	
		[Lam et al 1997]	0.443					
	Bim	[Li and Yamanishi 1999]				0.773		
	Naïve Bayes	[Yang et Liu 1999]				0.795		
	PropBayes	[Lewis 1992a]	0.650					
Regression	LLSF	[Yang 1999]		0.855	0.810			
		[Yang et Liu 1999]				0.849		
Sans apprentissage	Word	[Yang 1999]	0.150	0.310	0.290			
SVM		[Dumais 1998]				0.870		
	SVMLIGHT	[Dumais 1998]					0.920	
		[Joachims 1998]					0.864	
		[Li and Yamanishi 1999]				0.841		
		[Yang et Liu 1999]				0.859		
Probabilistic LSA Term	ADABOOST.MH	[Cai et Hofman 2003]						0.858
	ADABOOST.MR	[Cai et Hofman 2003]						0.843
Probabilistic LSA MixA	ADABOOST.MH	[Cai et Hofman 2003]						0.870
	ADABOOST.MR	[Cai et Hofman 2003]						0.861

Figure 16. Comparaison des performances des algorithmes

Malgré la difficulté de comparaison, il semble que l'on puisse tirer quelques enseignements et conclusions. Incontestablement, les méthodes par apprentissage donnent les meilleurs résultats, en particulier la méthode SVM (support vector machine). Mais il convient également d'être prudent sur l'interprétation des résultats. Par exemple, l'implémentation des machines à vecteurs supports proposée par [Dumais 1998] obtient des résultats nettement meilleurs que celle proposée par [Joachims 1998]. Joachims considère l'ensemble des termes du corpus Reuters, après suppression des mots les plus fréquents et l'utilisation de racines lexicales (les stems), soit 9962 termes distincts utilisés pour représenter les textes en entrée de son modèle. Il considère que l'ensemble de ces termes apporte de l'information, et qu'il est indispensable de les inclure tous. [Dumais et al. 1998] utilisent également les machines à vecteurs supports mais ils ne considèrent que 300 descripteurs pour représenter les textes. Ils obtiennent néanmoins de meilleurs résultats sur le même corpus. Nous observons que les réseaux de neurones testés par [Yang et Liu 1999] sont des perceptrons multi-couche avec une couche cachée comportant 64 neurones, 1000 descripteurs en entrées et 90 neurones de sorties correspondant aux 90 catégories ; ils considèrent un seul réseau pour l'ensemble des catégories comportant plus de 64000 poids (l'algorithme d'apprentissage n'est pas précisé). Il n'est pas surprenant, dans ces conditions, que les performances obtenues ne soient pas très bonnes : de telles démarches jugent plus la capacité des auteurs à mettre en œuvre des méthodes, que les capacités des méthodes elles-mêmes. L'algorithme de Rocchio est considéré comme ancien, mais [Schapire et al. 1998] a montré que cet algorithme obtient d'excellents résultats pour la catégorisation de textes à condition d'utiliser un codage efficace, de bien choisir les documents non pertinents, et d'effectuer une optimisation des poids ("*a state of the art version of Rocchio's algorithm is quite competitive with modern machine learning algorithms for text filtering*"). Leurs conclusions vont à l'encontre d'autres comparaisons qui montrent que cet algorithme n'est pas performant par rapport aux méthodes fondées sur l'apprentissage numérique [Schütze et al. 1995] [Lewis et al. 1996] [Cohen et Singer 1996]. Ces différentes remarques prouvent que le succès d'une méthode dépend d'un ensemble complexe de paramètres qui vont du codage des documents au choix des algorithmes et de leur implémentation, du choix des corpus et de leurs variantes, …

Il est par conséquent, extrêmement difficile de tirer des conclusions définitives sur une approche. Chaque publication nous apporte des éléments d'appréciations complémentaires mais il n'existe pas de méthode universelle.

Globalement, nous pouvons considérer que SVM est plus performant que les autres méthodes. Mais ce qui est vrai sur le corpus Reuters ou Ohsumed l'est-il sur d'autres corpus ? Nous verrons dans le chapitre suivant que ce n'est pas le cas sur notre base de comptes rendus hospitaliers.

Enfin, nous constatons que les traitements linguistiques des corpus ainsi que l'optimisation de l'apprentissage par boosting ou bagging améliorent les résultats obtenus.

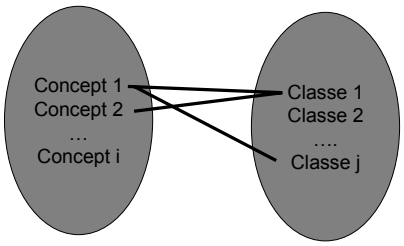
CHAPITRE 3
LA CLASSIFICATION AUTOMATIQUE DES COMPTES RENDUS HOSPITALIERS

Le but de CIREA est de déterminer les codes diagnostics à partir des comptes rendus hospitaliers. Ces derniers sont peu ou pas structurés et rédigés en langage naturel. Nous avons implémenté des algorithmes de textmining. Une première phase d'apprentissage utilisant les comptes rendus codés permet de modéliser les connaissances, en particulier de valoriser les relations entre les concepts et les diagnostics. Cette base de connaissance est utilisée pour permettre de proposer des codes à partir d'un nouveau compte rendu.

3.1 La formalisation du problème

Nous avons été confrontés à un problème de formalisation : comment devions-nous représenter le problème de classification automatique pour favoriser son exploitation par des algorithmes ?

Un problème de classification de document textuel se présente généralement sous une forme simple : un ensemble de concepts est relié à un ensemble de classes, avec une classe par document :

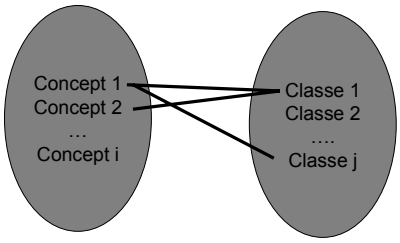


Modèle classique :
Un document = une classe

	Document j
Concept i	Effectif ij

La relation entre les classes (car un document = une classe) et les concepts conduit à réaliser une matrice dans laquelle on comptabilise le nombre de fois où la classe et les concepts sont associés.

Dans le cas d'une classification multi labels, on utilise généralement un produit cartésien pour démultiplier le problème d'affectation :

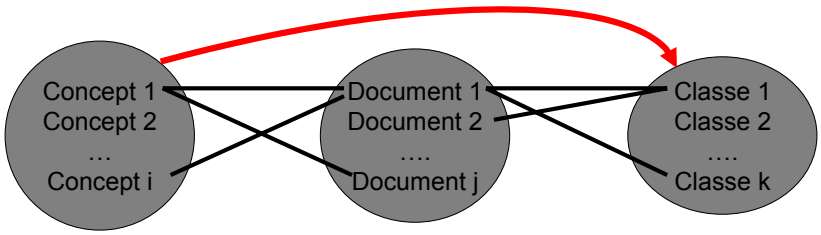


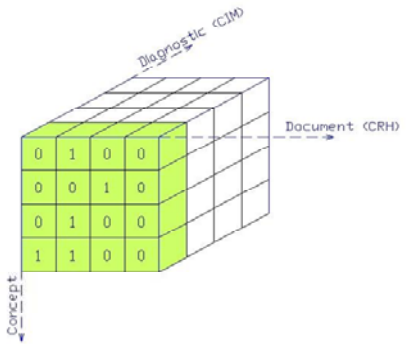
Modèle classique :
Un document = quelques classes
On démultiplie (produit cartésien)

Le fait d'avoir quelques classes possibles par document revient à ajouter une dimension au tableau.

	Document j	
	Classe a	Classe b
Concept i	Effectif ij	Effectif ij

La problématique de CIREA est en réalité un problème pour lequel nous ajoutons une dimension : nous supputons qu'il existe une relation entre les documents et les classes. Nous savons également que les documents contiennent des concepts. Pouvons nous en déduire que nous pouvons trouver un algorithme qui permette de mettre en relation les concepts avec les classes ? Nous sommes dans une représentation du problème sous trois dimensions, donc sous forme cubique :





Modèle Cirea :
Un document = plusieurs classes et plusieurs concepts.
L'enjeu devient de trouver un lien direct entre concepts et classes

Figure 17. Représentation du problème CIREA

A cette problématique s'ajoute celle de la faible reproductivité du codage, ce qui se traduit par un niveau d'incertitude fort sur les liaisons entre les ensembles.

3.2 Analyse du problème

Pour aboutir à réaliser le travail de classification, nous avons identifié plusieurs grandes phases, chacune représentant une difficulté particulière :

- la représentation du langage médical,
- la constitution du corpus d'apprentissage,
- la mise au point du plan expérimental,
- l'analyse des comptes rendus et de leur reproductivité,
- la création d'un algorithme de désuffixation,
- les algorithmes de classification,
- le bilan comparatif.

3.3 La représentation du langage médical

Il existe de nombreuses possibilités pour représenter le langage médical. Ce chapitre se propose de présenter les différentes techniques utilisées. Nous avons identifié six principales techniques : les thésaurus, les catalogues, les dictionnaires, les terminologies, les nomenclatures et les classifications.

3.3.1 Les thésaurus

Un thésaurus est un ensemble structuré de termes normalisés d'un domaine. Les thésaurus sont à visée d'indexation et de recherche de documents, dans un vocabulaire structuré et normalisé. Les relations sémantiques et conceptuelles sont importantes. Le thésaurus le plus important est le MeSH. Il sert de support à de nombreuses applications. Le MeSH (Medical Subject Headings) est un thésaurus conçu en 1960 par la NLM (National Library of Medicine). Il permet l'indexation des articles du domaine biomédical et d'accéder à la base de données Medline. Il contient plusieurs centaines de descripteurs (subject headings). Un descripteur est un terme représentant sans ambiguïté, un concept. Le concept en question

peut être : élémentaire (par exemple, vessie) ou complexe (par exemple, maladies artérielles occlusives). Le MeSH 2007 [MeSH 2007] comprend :

- 24 357 descripteurs
- 83 qualificatifs (Subheadings/Qualifiers)
- 164 331 concepts chimiques supplémentaires (SCRs)

avec

- 494 nouveaux descripteurs
- 99 descripteurs modifiés
- 22 descripteurs supprimés

Son arborescence est présentée en annexe 2.

3.3.2 Les catalogues

Un catalogue est un ensemble d'objets nommés décrivant, à un instant donné, un champ d'activité avec plus ou moins de précision. Ils ne possèdent pratiquement pas d'organisation sémantique ou hiérarchique. Ils sont souvent utilisés à des fins de tarification. Citons, par exemple, le CDAM (catalogue des actes médicaux), remplacé par la CCAM depuis 2002 (classification commune des actes médicaux).

3.3.3 Les dictionnaires

Un dictionnaire est un ensemble de mots avec leurs catégories grammaticales, leur sens et définition. Parmi les dictionnaires les plus connus, citons :

- Dictionnaire ADM
- Dictionnaire Internist/QMR (Quick Medical Reference)
- Dictionnaire DXplain
- Dictionnaire AI/RHEUM

Dans le domaine décisionnel, la valeur ajoutée par les dictionnaires est faible car ils sont peu exploitables en traitement automatisé. Par contre, ils permettent de lever toute ambiguïté.

3.3.4 Les terminologies

La terminologie est un ensemble de termes d'un domaine spécialisé organisé et structuré. La terminologie médicale possède des propriétés intéressantes car la structure des mots est porteuse de sens. En effet, de nombreux préfixes, suffixes, et affixes décrivent des actes ou désignent des localisations.

3.3.5 Les nomenclatures

Une nomenclature est un système de noms utilisés dans une branche de connaissance particulière. La nomenclature la plus représentative est la SNOMED. A l'origine, fut créée la SNOP (Systematized Nomenclature of Pathology) en 1965, par le Collège des Anatomo-pathologistes. Il s'agissait d'un système multi-axial (topographie, étiologie, morphologie, fonction), étendu à tout le vocabulaire médical par Roger Côté. Plusieurs versions successives ont été créées : la SNOMED I en 1975, SNOMED II (7 axes) en 1979, SNOMED International (11 axes) en 1990, et SNOMED RT en 1998.
Actuellement, la SNOMED comprend 200 000 termes.

Dans le domaine médical, une nomenclature existe à des fins de tarification : la NGAP (nomenclature générale des actes professionnels).

3.3.6 Le méta thésaurus UMLS

Un méta thésaurus est un sur-ensemble de toutes ces représentations. Pour représenter le langage médical, le méta thésaurus UMLS (Unified Medical Language System) est un acteur majeur. UMLS est un projet des USA de la National Library of Medicine (NLM) né en 1986. Il est révisé tous les ans depuis 1991. Son objectif est de constituer un langage unifié à partir des thésaurii, classifications, nomenclatures, … existants et de favoriser l'indexation automatique des textes biomédicaux.

Il a été construit à partir du MeSH, auquel on a rajouté un réseau sémantique pour constituer un méta thesaurus, étendu progressivement par l'ajout de nouveaux concepts, termes, liens et relations. UMLS comprend deux composants : le métathésaurus et le réseau sémantique.

Le métathésaurus constitue la base unifiée des concepts médicaux. Les termes sont reliés aux concepts : 730 000 concepts, 1 330 000 termes, 1 718 000 chaînes (Version 2000). Le MESH est intégré dans sa totalité et constitue la base du métathésaurus (426 000 chaînes). Les termes du MESH sont traduits en Français (INSERM), Allemand, Espagnol, Portugais.

Le réseau sémantique comprend 134 types sémantiques et comprend :

- les nœuds du réseau,
- des relations d'héritage is-a (est-un),
- des fonctions biologiques à 2 fils : physiologique et pathologique,
- des relations sémantiques (54 relations),
- des relations non hiérarchiques,
- des relations spatiales, temporelles, conceptuelles, physiques, fonctionnelles.

Les relations peuvent aussi être reliées hiérarchiquement.

Les autres sources importantes sont :

- les classifications ICD 9 (CIM 9), ICD9 CM (30000), ICD 10 (13000) DSM IV,
- les nomenclatures SNOMED International (164000),
- NHS Clinical terms (Read Codes 347000),
- UWDA (Digital Anatomist 27000),
- les terminologies et vocabulaires,
- actes : Procedures CPT 4 (15000),
- PDQ (Physician Data Query) (17 000),
- DXplain (10000),
- les termes venant de catalogues spécialisés,
- UMD (Universal Medical Device) (10000),
- LOINC (laboratoires) (64000),
- des dictionnaires de termes,
- LCH Librairie du Congres USA (6000),

L'unité sémantique est le concept identifié par un code unique (UI). Elle regroupe toute l'information spécifique au concept :

- définition,
- terme préférentiel,
- termes synonymes,
- variants lexicaux,
- types sémantiques.

Pour relier les concepts entre eux, on définit des liens sémantiques :

- relations hiérarchiques (relations PARent et CHilD),
- relations non hiérarchiques, notées RR si validées par un expert ou UR sinon.

Le réseau sémantique UMLS est constitué de nœuds (concepts) et d'arcs (relations sémantiques qui relient les concepts). Les types sémantiques sont rattachés aux concepts. Il s'agit d'un niveau d'abstraction supplémentaire par rapport au concept. La relation implicite est « est un » (is-a). Le réseau regroupe les relations sémantiques qui existent entre les catégories de concepts mais qui ne sont pas vérifiées pour chaque instance des types sémantiques.

Les types sémantiques permettent d'hériter des propriétés du type (est-un). Un concept est toujours relié à un type sémantique. Il existe 134 types Sémantiques. Exemples :

- types anatomiques : body part, organ, organ component (foie est-une partie du corps),
- type concept spatial: body space or junction (cavité péritonéale est-un espace du corps),
- type matériel ou instrument médical : medical device (stéthoscope est-un instrument médical).

Il existe 5 catégories de relations sémantiques :

- physically related to,
- spatially related to,
- temporally related to,

- functionally related to,
- conceptually related to.

Notons également 54 relations sémantiques, reliées également par des relations hiérarchiques (is-a). Par exemple, 'Affects' possède 6 relations filles : 'manages', 'treats', 'disrupts', 'complicates', 'interacts with', et 'prevents'.
Dans certains cas, les relations ne peuvent être applicables pour certaines instances. De même l'héritage de certaines relations peut être bloqué.

3.3.7 La CIM 10

La classification des maladies est une problématique ancienne. Dès 1853, W. Farr réalise une Nomenclature des causes de décès. En 1893, le Dr Jacques Bertillon établit la classification des causes de mort. Il faudra attendre 1946 pour que l'OMS (Organisation Mondiale de la Santé) établisse une première classification complète des maladies. Celle-ci sera révisée tous les 5 ans jusque la version 9, parue en 1975. En 1988, apparaît la version ICD 9-CM (Clinical Modification). A partir de cette date, nous avons assisté à diverses adaptations, souvent liées à l'usage d'un système assurantiel ou social, comme par exemple ICD10-AM pour Australian Modification qui intègre les actes, ou la CIM10-CA pour son adaptation Canadienne. A ce jour, la France utilise la CIM dans sa 10ème version, parue depuis 1994, soit la dernière en date. Depuis cette version, la mise à jour est continue. Les actes sont codés par la CCAM (Classification Commune des Actes Maladies). La CIM-10 (Classification internationale des Maladies, 10ème version) est un catalogue hiérarchisé et codifié des maladies. La hiérarchie complète des codes est sur 5 niveaux et couvre environ 51000 codes diagnostics. Mais elle est utilisée dans la majorité des cas sur les 4 premiers niveaux seulement, ce qui couvre environ 15 000 codes diagnostics.

Voici le premier niveau de la CIM10 :

I **Maladies infectieuses et parasitaires**
II **Tumeurs**
III **Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire**
IV **Maladies endocriniennes, nutritionnelles et métaboliques**
V **Troubles mentaux et du comportement**
VI **Maladies du système nerveux**
VII **Maladies de l'œil et de ses annexes**
VIII **Maladies de l'oreille et de l'apophyse mastoïde**
IX **Maladies de l'appareil circulatoire**
X **Maladies de l'appareil respiratoire**
XI **Maladies de l'appareil digestif**
XII **Maladies de la peau et du tissu cellulaire sous-cutané**
XIII **Maladies du système ostéo-articulaire, des muscles et du tissu conjonctif**
XIV **Maladies de l'appareil génito-urinaire**
XV **Grossesse, accouchement et puerpéralité**
XVI **Certaines affections dont l'origine se situe dans la période périnatale**
XVII **Malformations congénitales et anomalies chromosomiques**
XVIII **Symptômes, signes et résultats anormaux d'examens cliniques et de laboratoire, non classés ailleurs**
XIX **Lésions traumatiques, empoisonnements et certaines autres conséquences de causes externes**
XX **Causes externes de morbidité et de mortalité**
XXI **Facteurs influant sur l'état de santé et motifs de recours aux services de santé**

Il est utile de souligner que le chapitre XXI, dont les codes commencent tous par la lettre 'Z' ne donne pas de diagnostic mais permet de préciser des facteurs environnementaux au diagnostic. Certains diagnostics peuvent se retrouver dans plusieurs chapitres, mais ces cas tendent à disparaître à chaque révision. Par exemple, la syphilis cardio-vasculaire peut être codée de plusieurs façons différentes, selon qu'elle est une cause ou une conséquence, donc selon l'approche médicale. Si on l'aborde comme étant une maladie de l'appareil circulatoire, elle sera codée en I98 (autres troubles de l'appareil circulatoire au cours de maladies classées ailleurs), et plus précisément en I98.0 (syphilis cardio-vasculaire). Mais si on la considère comme une maladie infectieuse ou parasitaire dont le mode de transmission est essentiellement sexuel, elle sera codée en A52.0 (syphilis cardio-vasculaire), voire en A50.5 (autres formes tardives de syphilis congénitale, symptomatique). Cet exemple souligne à nouveau la difficulté de l'évaluation.

Les codes des maladies infectieuses et parasitaires vont de A00 à B99. Voici le détail de ce second niveau de hiérarchie :

A00-A09	Maladies intestinales infectieuses
A15-A19	Tuberculose
A20-A28	Certaines anthropozoonoses bactériennes
A30-A49	Autres maladies bactériennes
A50-A64	Infections dont le mode de transmission est essentiellement sexuel
A65-A69	Autres maladies à spirochètes
A70-A74	Autres maladies à *Chlamydia*
A75-A79	Rickettsioses
A80-A89	Infections virales du système nerveux central
A90-A99	Fièvres virales transmises par des arthropodes et fièvres virales hémorragiques
B00-B09	Infections virales caractérisées par des lésions cutanéo-muqueuses
B15-B19	Hépatite virale
B20-B24	Maladies dues au virus de l'immunodéficience humaine [VIH]
B25-B34	Autres maladies virales
B35-B49	Mycoses
B50-B64	Maladies dues à des protozoaires
B65-B83	Helminthiases
B85-B89	Pédiculose, acariase et autres infestations
B90-B94	Séquelles de maladies infectieuses et parasitaires
B95-B97	Agents d'infections bactériennes, virales et autres
B99	Autres maladies infectieuses

Dans le chapitre A00-A09, nous retrouvons les maladies infectieuses intestinales. Le niveau suivant donne un niveau de précision complémentaire :

A00 Choléra

A00.0	A *Vibrio cholerae* 01, biovar cholerae Choléra classique
A00.1	A *Vibrio cholerae* 01, biovar El Tor Choléra El Tor
A00.9	Choléra, sans précision

A01 Fièvres typhoïde et paratyphoïde

A01.0	Fièvre typhoïde, infection due à *Salmonella typhi*
A01.1	Paratyphoïde A
A01.2	Paratyphoïde B
A01.3	Paratyphoïde C

…

Le niveau .9 est généralement réservé au diagnostic "sans autre précision" ou "SAI : sans autre indication".

Malheureusement, le codage des actes et des diagnostics n'est pas le même pour tous les pays, bien que ces différences tendent à se résorber, de par l'apparition de normes (HL7) et la nécessité d'harmoniser les échanges électroniques. Voici la synthèse des codifications retenues par pays à ce jour :

Pays	Codage des actes et procédures	Codage des diagnostics
Etats-Unis	CPT-4 /5	
	ICD9-PCS	
		ICD-10
Grande Bretagne	Read 4	
	CTV3	CTV3
	OPCS-4	
		ICD-9
		ICD-10
Allemagne[3]	ICPM	
		ICD-10 SGB version 1.3
		ICD-10 SGB version 2.0
	ZEK, AVB	
Espagne		ICD-10
France	CDAM / NGAP / CCAM	ICD-10
Portugal		ICD-10
Italie	ICD9CM (Italian)	ICD-10
Autriche	ICPM	
Suisse	ICPM	ICD-10
Belgique (Francophone)	INAMI	ICD-9
Belgique (Flandres)	RIZIV	ICD-9
Pays Bas	COTG	
Danemark	SKS	
Norvège	NCSP (Norwegian)	ICD-10 (Norwegian)
Suède	NCSP (Swedish)	ICD-10 (Swedish)
Finlande	NCSP (Swedish, Finish var.)	
Australie	ICD10-AM	ICD10-AM
Canada	ICD9-CCA	ICD9-CA ou ICD10-CA

Figure 18. Codage des maladies dans les différents pays

[3] AVB et ZEK sont utilisés pour coder les évènements critiques.

3.4 Constitution des corpus, des bases d'apprentissage et de test

3.4.1 Constitution d'une base de CRH

Pour couvrir le domaine de recherche, il nous a fallu recueillir les comptes rendus médicaux provenant de plusieurs sources et se constituer une base de données favorisant le traitement automatique du langage médical.

Les médecins doivent rédiger un compte rendu hospitalier (CRH) pour chaque patient admis en service de réanimation. Celui-ci contient de nombreuses informations, organisées autour du schéma suivant :

- motif de l'hospitalisation,
- antécédents médicaux et chirurgicaux,
- histoire de la maladie,
- résultats des différents examens cliniques et complémentaires,
- évolution de l'état du patient pendant son séjour en réanimation,
- conclusion donnant le diagnostic et l'issue du séjour,
- codification des pathologies.

Nous avons recueilli environ 40 000 CRH provenant de 15 établissements hospitaliers de France avec des origines géographiques étendues. Nous avons été confrontés à une grande variété de formats de bases de données, de codages (unix, mac, PC, …) et de systèmes d'informations (logiciels commerciaux ou propriétaires). De plus, certains établissements possédaient d'anciens CRH codés en CIM9, ou simplement n'avaient pas de diagnostics codés mais rédigés en langage naturel. Nous avons développé un programme permettant de transposer certains codes CIM9 en CIM10 grâce à une table de concordance partielle. **Au final, nous avons pu constituer une base de données avec 33 000 comptes rendus médicaux exploitables, structurés et étiquetés**.

3.4.2 Constitution d'une base pour le traitement du langage

Pour constituer une base de données permettant le traitement du langage médical, nous avons réuni les éléments suivants :

- une table des termes du vocabulaire français,
- une table des concepts médicaux,
- une table des préfixes et suffixes médicaux,
- une liste de stopwords.

La table du vocabulaire français contient 543 418 mots, dont l'origine principale est l'ABU (http://abu.cnam.fr/). Nous l'avons complété avec la liste des termes autorisés au scrabble. L'avantage de la liste ABU est que chaque terme est fourni sous toutes ses formes et le lemme associé est précisé.

Nous avons réalisé une base de données de concepts médicaux. Nos principales sources sont les suivantes :

- terminologie du MeSH,
- terminologie extraite de la CIM 10,
- ajouts manuels,
- SNOMED.

Au final, 100 882 concepts médicaux (au sens de [Dubois 1994]) constituent notre base de concepts techniques, allant du terme simple à des concepts complexes.

Nous nous sommes procuré une liste de stopwords de 224 items.

Grâce au concours des médecins réanimateurs et des départements d'informatique médicale, nous avons réalisé manuellement (à partir d'internet et des contenus des CRH), une table de 957 acronymes médicaux fréquents. Pour cela, nous avons recherché dans le corpus tous les termes de moins de 5 lettres avec au moins une majuscule et les avons triés par fréquence d'apparition.

Voici un extrait du catalogue des acronymes :

Acronyme	Sens
AA	acide aminé
AAA	Anévrisme de l'aorte abdominale
AAG	Asthme aigu grave
AAS	Anthrax Anti Serum
ABP	Antibioprophylaxie
ABPA	Aspergillose bronchopulmonaire allergique
ABVD	Adriamycine, Bléomycine, Vincristine et DTIC (déticène)
AC	AntiCorps
ACE	Antigène Carcino-Embryonnaire (marqueur de tumeur cancéreuse)
ACR	Arrêt cardio-respiratoire
ACTH	Hormone corticotrope hypophysaire
ACTP	Allocation Compensatrice pour Tierce Personne
ACV	Accident Cérébro-Vasculaire
AD	Virus Adénovirus
AFP	Alpha -foetoprotéine

Figure 19. Catalogue des acronymes médicaux

Enfin, nous avons compilé plusieurs sources ([Pol 2003] [Likita 2004] [Pouliquen 2004] [Dolisi 2005] et [Goonarch 2005]), pour obtenir un dictionnaire composé de 1445 racines lexicales de termes médicaux : 687 préfixes, 528 suffixes, 129 affixes, et 101 termes retirés car susceptibles de générer des erreurs (par exemple le préfixe privatif 'a' ne doit pas s'appliquer à tous les termes commençant par 'a').

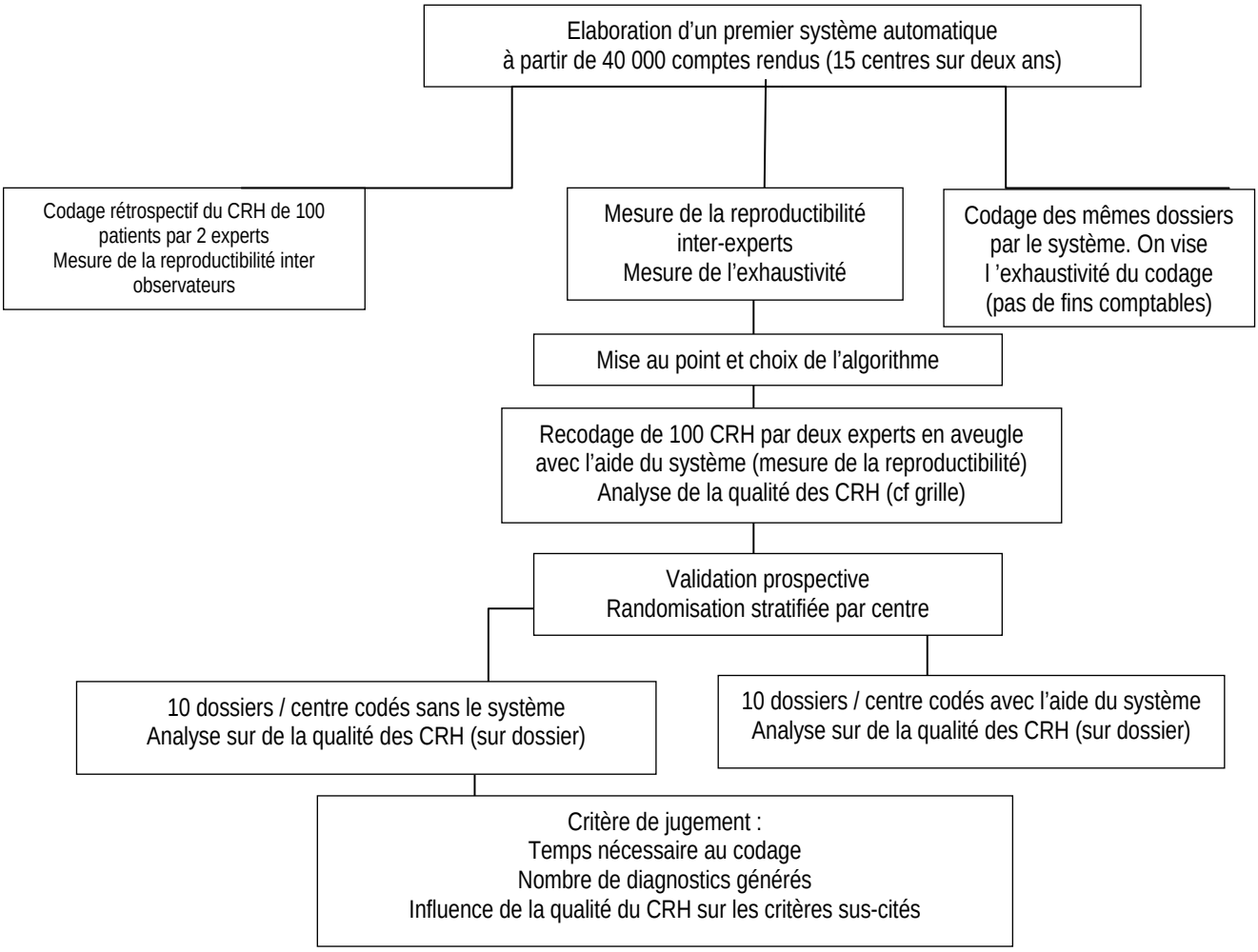
Racine	Origine	Relatif à
thym(o)-	Du grec thumia, de thumos [thym(o)-, -thymie, -thymique], cœur, affectivité, relatif à la passion.	cœur
amygdal(o) -	Du latin amygdala, du grec amugdalê [amygdal(o)-], amande; en anatomie, relatif aux amygdales, organes lymphoïdes de la gorge.	amygdale
tom(o)-	Du grec tomê [tom(o)-, -tome, -tomie], section, coupe.	couper
toc(o)-	Du grec tokos [toc(o)-, -tocie], accouchement.	accouchement
ampull(o) -	Du latin ampulla [ampull(o)-, -ampullaire], petite ampoule ; en médecine, relatif à l'ampoule de Vater, dans le duodénum, où aboutissent le canal cholédoque et le canal de Wirsung (pancréas).	duodenum
thromb(o)-	Du grec thrombos [thromb(o)-], caillot, en rapport avec la coagulation du sang.	caillot
stapéd(o)-	Du latin stapes [stapéd(o)-, -stapéal], étrier, relatif à l'osselet du même nom situé dans l'oreille moyenne.	oreille
athér(o) -	Du latin atheroma, du grec athêra [athér(o)-], bouillie ; terme qui s'applique à la tunique interne d'une artère.	artère

Figure 20. Exemples de préfixes médicaux

Cet ensemble de données nous a permis de disposer d'une base de données avec des comptes rendus hospitaliers étiquetés, ainsi que d'un ensemble de vocabulaire, termes, outils permettant de les exploiter efficacement.

3.5 Plan expérimental

La démarche expérimentale est décomposée comme suit :



La qualité globale du CRH est évaluée en fonction des recommandations suivantes :

- décret n°2002-637 du 29 avril 2002 relatif à l'accès aux informations professionnelles détenues par les professionnels et les établissements de santé en application des articles L 1111-7 et L. 1112-1 du code de la santé publique,
- décret n° 92-329 du 30 mars 1992 relatif au dossier médical et à l'information des personnes accueillies dans les établissements de santé publics et privés modifiant le code de santé publique,
- manuel d'accréditation des établissements de santé, février 1999, ANAES.

3.6 Analyse des comptes rendus et de leur reproductivité

3.6.1 Analyse statistique et dénombrements

Comme tout problème d'analyse de données, nous avons procédé à une première analyse des comptes rendus hospitaliers sous forme de statistique descriptive. Notre première approche consiste à dénombrer les effectifs de diagnostics par compte rendu :

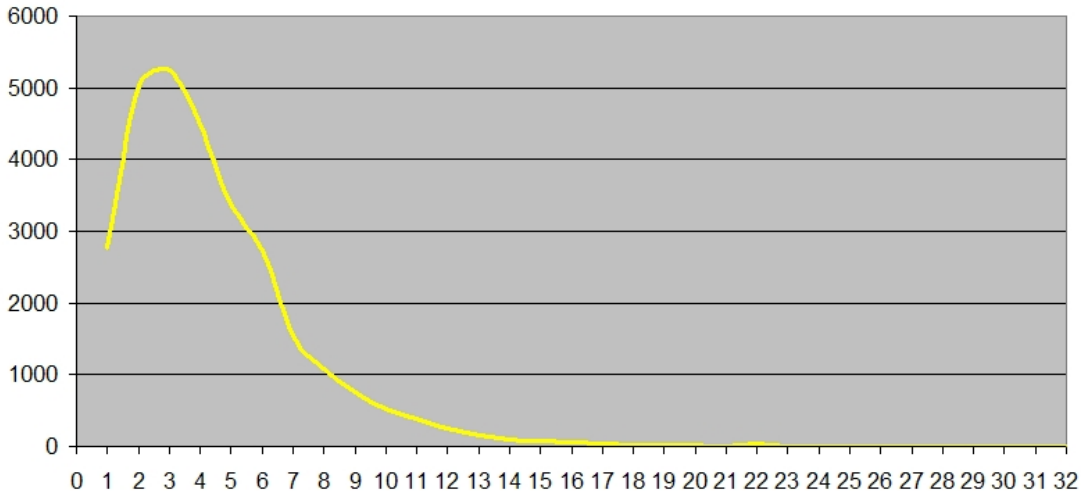


Figure 21. Nombre de diagnostics par compte rendu

Nous dénombrons entre 1 et 32 codes diagnostics par CRH. Ces résultats sont conformes à ce qui a été mesuré dans d'autres études [Cubrea].

Nous avons ensuite analysé la concentration des codes diagnostics :

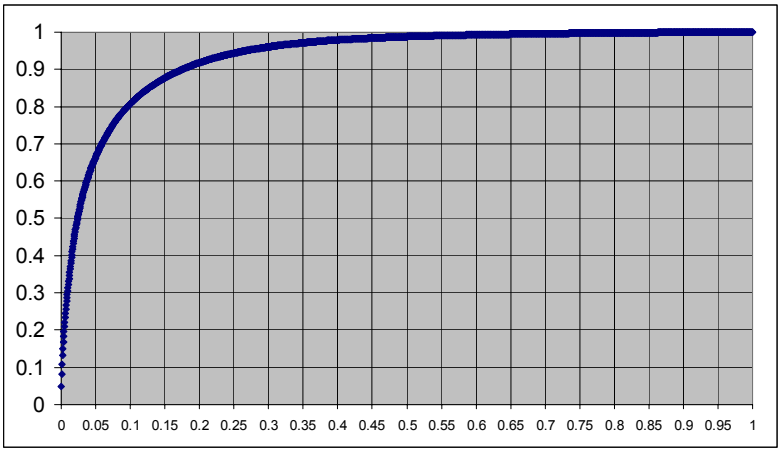


Figure 22. Concentration des diagnostics

Nous constatons également une forte concentration sur quelques diagnostics. Par exemple, 10% des codes se retrouvent dans 80% des comptes rendus. Les diagnostics les plus fréquents sont les suivants :

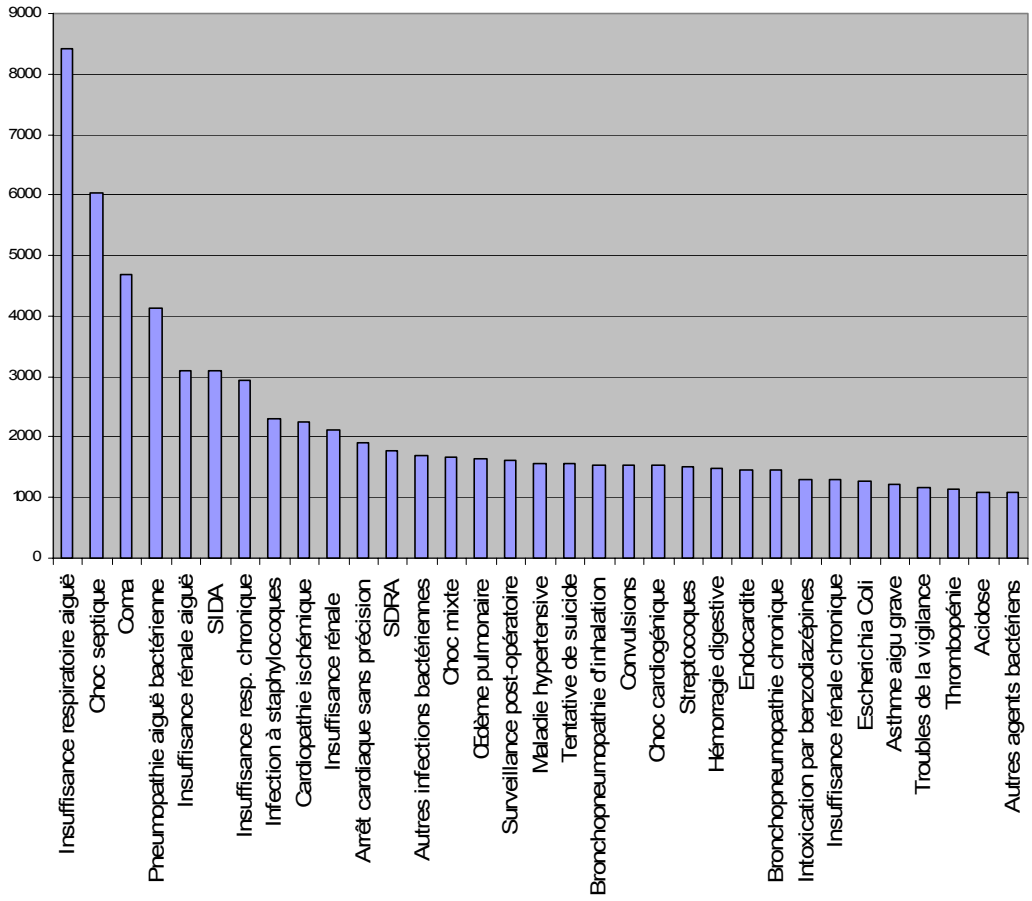


Figure 23. Liste des diagnostics les plus fréquents

Ce premier constat nous pose un premier problème : un système informatique qui proposerait systématiquement les codes les plus fréquents obtiendrait une bonne précision mais serait totalement inutile. Le rappel serait mauvais et la pertinence d'un tel système bien faible. Il nous faudra par conséquent avoir une approche qualitative forte.

Enfin, nous avons voulu savoir s'il pouvait exister une corrélation entre le nombre de termes utilisés dans le CRH et le nombre de diagnostics :

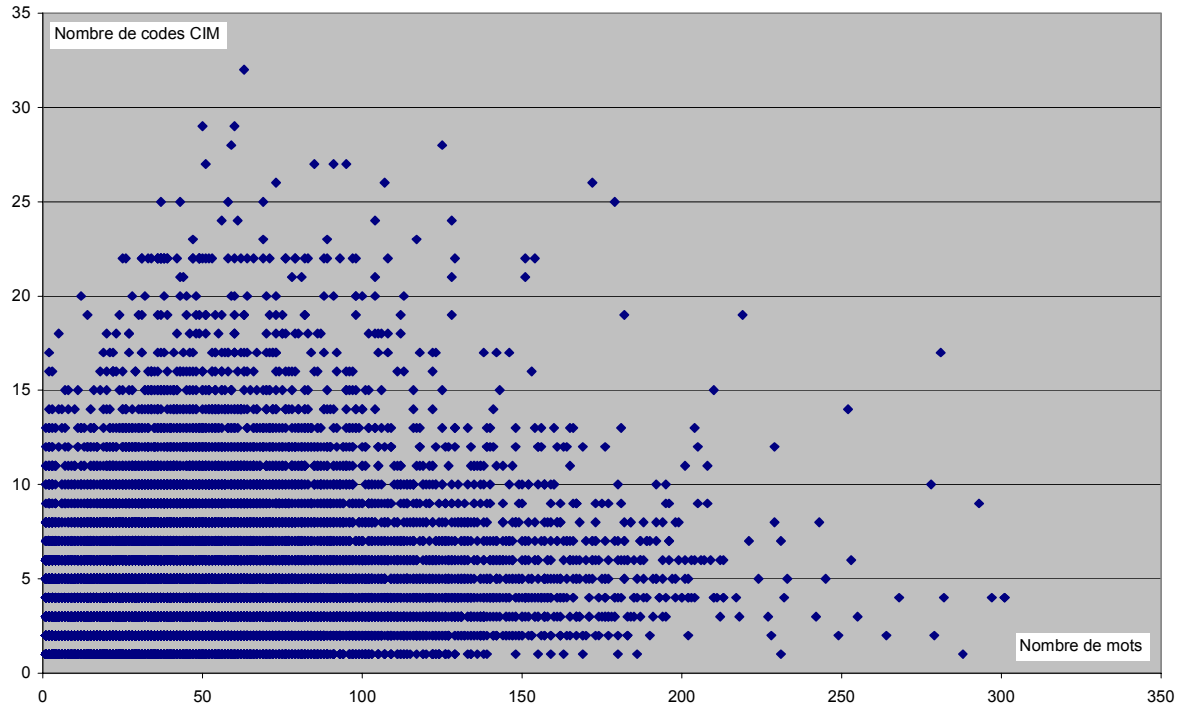


Figure 24. Nombre de diagnostics et nombre de mots

L'analyse graphique révèle qu'il n'existe aucun lien entre ces deux éléments. Le coefficient de corrélation est de 34.59%, le coefficient de détermination n'est que de 11.96% ce qui est insuffisant pour être significatif.

L'analyse des co-occurrences ne permet pas non plus d'extraire de la connaissance utile car les effectifs ne sont pas assez significatifs non plus :

cim1	Libellé	cim2	Libellé	Nbfois
J960	Insuffisance respiratoire aigüe	R402	Perte de conscience SAI	1393
J159	Pneumopathie bactérienne, sans précision	J960	Insuffisance respiratoire aigüe	1382
A419	Septicémie, sans précision	J960	Insuffisance respiratoire aigüe	1167
A419	Septicémie, sans précision	J159	Pneumopathie bactérienne, sans précision	1090
A419	Septicémie, sans précision	A490	Infection à staphylocoques, sans précision	915
A419	Septicémie, sans précision	N179	Insuffisance rénale aigüe, sans précision	906
J449	Maladie pulmonaire obstructive chronique, sans précision	J961	Insuffisance respiratoire chronique	755
A419	Septicémie, sans précision	A498	Autres infections bactériennes, siège non précisé	726
I255	Myocardiopathie ischémique	J960	Insuffisance respiratoire aigüe	695
J960	Insuffisance respiratoire aigüe	N179	Insuffisance rénale aigüe, sans précision	681
A419	Septicémie, sans précision	J80	Syndrome de détresse respiratoire de l'adulte	641
A498	Autres infections bactériennes, siège non précisé	J159	Pneumopathie bactérienne, sans précision	638
R402	Perte de conscience SAI	Z915	Antécédents personnels de Tentative de suicide	637
T424	Intoxication par Benzodiazépines	Z915	Antécédents personnels de Tentative de suicide	586
A498	Autres infections bactériennes, siège non précisé	J960	Insuffisance respiratoire aigüe	586
R402	Perte de conscience SAI	T424	Intoxication par Benzodiazépines	572
I255	Myocardiopathie ischémique	J159	Pneumopathie bactérienne, sans précision	517
A491	Infection à streptocoques, sans précision	J159	Pneumopathie bactérienne, sans précision	506
A490	Infection à staphylocoques, sans précision	J159	Pneumopathie bactérienne, sans précision	500

Figure 25. Co-occurrences des codes CIM

3.6.2 Analyses basiques

Nous avons fait plusieurs analyses statistiques de base. En particulier, nous avons calculé des indicateurs de gravité qui nous ont permis de calculer une probabilité de décès. A partir de ceux-ci, nous avons comparé le taux de décès prévu d'après le SAPS (voir annexe 3 page 212) avec la mortalité constatée. Pour calculer la mortalité prédite à partir du SAPS, nous utilisons la formule : mortalité prédite =e(Logit)/(1+e(Logit)), où 'e' représente la fonction exponentielle et avec Logit = -14.4761 + 0,0844 * SAPS + 6.6158 * log(SAPS+1).

Probabilité théorique	Décès constatés	Ecart	Effectif concerné	Individus
5%	2.59%	-2.41%	116	-2.8
15%	23.91%	8.91%	46	4.1
25%	22.73%	-2.27%	22	-0.5
35%	39.13%	4.13%	23	0.95
45%	26.67%	-18.33%	15	-2.75
55%	43.48%	-11.52%	23	-2.65
65%	52.63%	-12.37%	19	-2.35
75%	**63.64%**	**-11.36%**	**22**	**-2.5**
85%	76.19%	-8.81%	21	-1.85
95%	85.19%	-9.81%	27	-2.65
		TOTAL	**334**	**-13**

Par rapport au modèle théorique (qui est un bon modèle de prédiction), 13 personnes de plus auraient du décéder. Soit un gain de 3.89%.

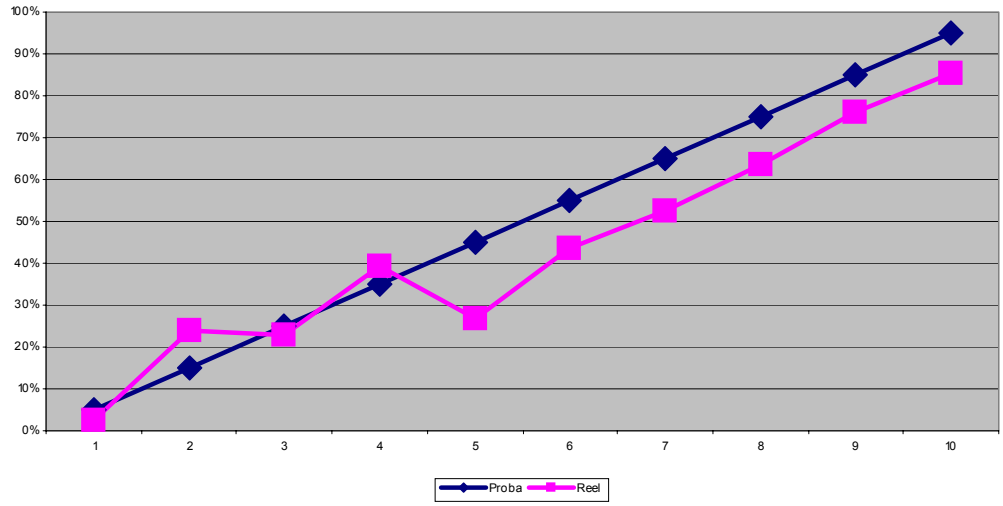


Figure 26. Probabilité de décès et décès constatés

Sur le plan qualitatif, nous observons que cet établissement soigne très bien les patients à pronostic sévère. Le praticien concerné a souhaité analyser pourquoi le taux est trop élevé pour la tranche des 15%.

3.6.3 De la reproductivité du codage

Nous nous sommes ensuite penchés sur l'analyse de la reproductivité du codage. Pour cela, nous avons extrait 100 comptes rendus codés dans notre base de données et avons demandé à ce que chacun soit recodé en aveugle par deux médecins experts du codage. Les comptes rendus extraits l'ont été selon des techniques d'échantillonage en se basant sur une grande dispersion d'établissement, de codes, de nombre de diagnostics, de concepts, … Puis nous avons comparé le codage initial avec les deux recodages par les médecins experts. Le niveau de reproductibilité totale s'est avéré très faible (18% de consensus total) [Misset 2005].

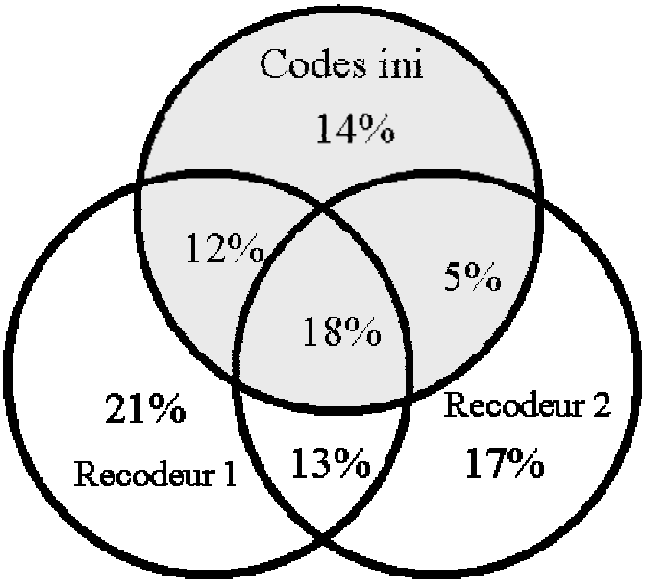


Figure 27. Reproductibilité du codage

Cette observation permet de soulever le problème de l'évaluation. Le codage n'étant pas reproductible, il faudra réaliser une analyse qualitative fine des résultats obtenus par l'algorithme.

3.7 Proposition d'un algorithme de désuffixation : EDA

Pour faciliter le traitement sémantique, nous avons élaboré un algorithme de désuffixation du langage médical : EDA. La désuffixation fait partie des techniques simples de traitement des données textuelles facilitant la classification automatique de documents. Le traitement sémantique de la terminologie médicale soulève trois catégories de difficultés :

- d'une part, de nouveaux termes apparaissent régulièrement, ce qui nécessite des mises à jour fréquentes,
- d'autre part, celles-ci nécessitent un consensus d'experts car ces concepts demandent une très grande précision dans leur définition, ce qui allonge les délais de disponibilité des mises à jour,
- enfin, il s'agit d'une forme de langage particulier, très technique, avec peu de flexions et de conjugaisons.

Pour contourner ces problématiques, nous avons voulu savoir si les techniques de désuffixation pouvaient améliorer les résultats et s'il était possible d'optimiser les algorithmes existants. Nous avons également mesuré l'impact de l'enrichissement du texte en fonction des racines lexicales rencontrées dans la terminologie médicale.

Nous nous sommes demandés si l'implémentation d'un algorithme de désuffixation dans ce contexte pouvait améliorer significativement les résultats obtenus et dans quelles proportions.

3.7.1 L'algorithme EDA

L'algorithme EDA fonctionne en deux phases principales : une phase de préparation et d'harmonisation de la forme, suivi d'une phase de traitement.

3.7.1.1 Phase 1

La première phase consiste à préparer le mot en appliquant quelques modifications :

1. transformation en minuscules,
2. séparation des caractères ligaturés ('cœur' devient 'coeur') et des traits d'unions,
3. suppression des signes diacritiques (ex : accents 'dégénéré' devient 'degenere'),
4. suppression des doubles lettres,
5. remplacer 'ck', 'cqu', 'qu', et 'c' par 'k', et 'y' par 'i'.

La transformation en minuscules permet de ne pas se préoccuper de la casse, en rendant identiques des termes qui ne l'auraient pas été. Ainsi 'Cardiopathie' et 'CARDIOPATHIE', considérés initialement comme différents deviennent tous deux 'cardiopathie'. La séparation des caractères ligaturés et des traits d'unions poursuit le même objectif, à savoir rendre identiques des termes considérés comme différents.

Les trois dernières règles de préparation des données permettent uniquement de corriger les erreurs de frappe et erreurs induites par le changement de casse. En effet, les mots en majuscules ne sont généralement pas accentués, et les mots 'Thérapie' et 'THERAPIE' auraient été différents après changement de la casse. [Zweigenbaum 2002] a décrit une méthode de réaccentuation du langage médical (termes du MESH) obtenant de très bons résultats (break even point à 84.2 ± 4.4 %). Le remplacement des lettres 'ck', 'cqu', 'qu', et 'c' par la lettre 'k', ainsi que 'y' par 'i' s'explique par le grand nombre de noms propres utilisés en médecine, avec des orthographes différentes. L'objectif est donc de se rapprocher d'une prononciation phonétique.

Cette première phase permet de nettoyer le terme et le mettre dans une forme 'standard'. La suppression des accents nous a permis de regrouper de nombreux termes qui étaient considérés au préalable comme différents. Dans l'expérimentation présentée en section 4, les trois premières règles ont été appliquées à tous les algorithmes.

A l'issue de cette première phase, nous regardons si le mot restant contient au moins 5 caractères. Si ce n'est pas le cas, on supprime un éventuel 's' et/ou 'e' final et la procédure s'arrête à cette étape. Dans le cas contraire, la procédure continue.

La seconde phase consiste à exécuter séquentiellement un ensemble de 37 règles (voir détail en annexe). Il est important de respecter ces règles dans l'ordre, jusqu'à ce que le mot

résultant contienne 5 caractères ou, dans le cas contraire, jusqu'à la dernière règle. Chaque règle s'applique sur le résultat obtenu par la règle précédente. Voici un exemple de plusieurs déclinaisons qui aboutissent au même résultat :

Terme de départ	Règles appliquées	Terme obtenu
INTESTIN	Aucune	INTESTIN
INTESTINS	1	INTESTIN
INTESTINE	2	INTESTIN
INTESTINES	1 et 2	INTESTIN
INTESTINAL	5	INTESTIN
INTESTINAUX	3 et 6	INTESTIN
INTESTINALES	1, 2, et 5	INTESTIN
INTESTINALE	2 et 5	INTESTIN

Figure 28. Exemples de résultats de désuffixation

3.7.1.2 Phase 2 : enrichissement des termes médicaux

La terminologie médicale possède une structure sémantique forte. [Jujols 1991] et [Pacak 1983] ont depuis longtemps identifié les racines grecques et latines. Nous nous sommes demandé si cette structure spécifique ne pourrait pas permettre d'enrichir un corpus de textes médicaux pour améliorer les résultats d'un algorithme de classification. L'avantage de cette méthode est qu'elle ne demande aucune mise à jour et permet même d'induire de la connaissance sur des termes nouveaux ou inconnus.

Nous sommes partis de l'hypothèse que cette structure pouvait apporter un supplément de connaissance. [Zweigenbaum 1999] a exploré avec succès cette piste avant de traduire UMLS en Français. [Namer 2005] a utilisé cette structure pour définir automatiquement les termes médicaux. Chaque fois qu'un préfixe est rencontré, nous ajoutons le terme correspondant. Par exemple, le terme 'pneumopathie' contient le préfixe 'pneumo', ce qui génère l'ajout du mot 'poumon'. Chaque fois qu'un suffixe est rencontré, le terme correspondant est ajouté, à condition qu'un préfixe ait été rencontré. Par exemple, le suffixe 'ite' désigne une inflammation. Le mot 'évite' ne sera pas traité, alors que le mot 'laryngite' génèrera 'larynx' et 'inflammation'.

Pour désigner les organes, voici quelques exemples de préfixes utilisés :

Anus	proct...
Genou	gon...
Hanche	cox...
Bassinet	pyélo...
Caecum	typhlo...
Cordon	funicul...
Doigt	dactyl...
Diaphragme	phréno...
Foie	hépat...
Intestin	entéro...
Langue	glosso...

Péritoine	cœlio...
Rate	spléno...
Rein	néphr...
Testicule	orch...
Tête	céphal...
Trompe	salpin...
Utérus	hystéro, métro...
Vagin	colpo...
Veine	phléb..., véno...
Vésicule biliaire	cholécyst...
Vessie	cysto...

Figure 29. Désignation des organes

Nous avons identifié deux types de suffixes : certains sont des indicateurs d'un état pathologique, alors que les autres sont indicateurs de gestes techniques.

Les suffixes indicateurs d'états pathologiques

Associés au nom d'un organe ou d'un tissu, ils dénotent un état pathologique. Exemple : 'ite' témoigne d'une inflammation (pancréatite, appendicite, gastrite), 'algie' ou 'odynie' dénotent une douleur, …

Les suffixes indicateurs de gestes techniques

Associés au nom de l'organe concerné, ils permettent de préciser les gestes techniques réalisés. Par exemple : 'centèse' signifie ponction, 'ectomie' représente l'ablation, 'plastie' la réparation, …

3.7.2 Expérimentation et résultats

Pour expérimenter notre algorithme, nous avons choisi d'utiliser Naïve Bayes comme algorithme de classification, en se limitant au niveau 1 de la CIM, avec la formule à posteriori utilisant la loi de probabilité totale [Schneider 2004] :

$$p(c_j / d) = \frac{p(c_j) * p(d / c_j)}{p(d)} \text{ avec } p(d) = \sum_{j=1}^{|C|} p(c_j) p(d / c_j)$$

Pour évaluer les résultats, nous avons utilisé deux indicateurs de synthèse : la K-mesure [Nakache 2005] et la F-mesure ([Sebastiani 2004] et [Van Rijsbergen 1979]).

La K-mesure est utilisée avec les paramètres alpha= 1.2 et beta = 1.5, pour mieux traduire le besoin exprimé par les médecins, à savoir privilégier le rappel et pénaliser les mauvais scores. Nous avons défini 5 seuils (2, 4, 6, 8, et 10) correspondant aux nombres de réponses proposées par l'algorithme. L'algorithme appliqué (Naïve Bayes) donne les résultats suivants :

Seuil	Précision	Rappel	F-mesure	K-mesure
2	0.980	0.472	0.637	0.587
4	**0.715**	**0.671**	**0.692**	**0.602**
6	0.576	0.810	0.673	0.559
8	0.473	0.887	0.617	0.489
10	0.396	0.929	0.555	0.421

Figure 30. Résultats EDA sans aucun algorithme

Pour l'ensemble des résultats qui suivent, nous ne présentons que le meilleur résultat (seuil = 4). Nous avons ensuite expérimenté l'algorithme Carry [Carry 2002], adaptation au Français de l'algorithme de Porter :

Précision	Rappel	F-mesure	K-mesure
0.747	0.700	0.723	0.639

Figure 31. Résultats de désuffixation avec l'algorithme Carry

Et enfin l'algorithme EDA sans ajout de termes :

Précision	Rappel	F-mesure	K-mesure
0.753	0.706	0.729	0.646

Figure 32. Résultats de désuffixation avec l'algorithme EDA

Nous constatons que l'algorithme Carry apporte effectivement un gain sur l'indicateur de performance global de l'algorithme de classification de l'ordre de 3.1 % et qu'EDA l'améliore de 3.7 %.

Nous avons ensuite expérimenté l'enrichissement des données à partir des racines lexicales des termes médicaux. Le tableau de synthèse récapitulatif est le suivant :

Désuffixation	**Enrichi**	**Précision**	**Rappel**	**F-mesure**	**K-mesure**
Aucune désuffixation	Non	0.715	0.671	0.692	0.602
Aucune désuffixation	Oui	0.744	0.697	0.720	0.635
Désuffixation avec Carry	Non	0.747	0.700	0.723	0.639
Désuffixation avec Carry	Oui	0.766	0.718	0.741	0.662
Désuffixation avec EDA	Non	0.753	0.706	0.729	0.646
Désuffixation avec EDA	Oui	0.772	0.724	**0.747**	0.669

Figure 33. Synthèse des résultats obtenus avec EDA

Ce qui donne les résultats suivants pour la F-mesure :

Désuffixation	**Non Enrichi**	**Enrichi**
Aucune désuffixation	69.23%	71.97%
Désuffixation avec Carry	72.27%	74.12%
Désuffixation avec EDA	72.87%	74.72%

Figure 34. Gains sur la F-mesure selon la méthode utilisée

Nous constatons que les résultats obtenus sont meilleurs lorsque l'on ajoute des termes en fonction des racines lexicales. Le gain varie de +1.85 % lorsque les termes ont été désuffixés à +2.74 % sans traitement de désuffixation. Au final, le gain total de l'algorithme EDA enrichi est de +5.49 % par rapport à une classification sur les termes non désuffixés et non enrichis. La performance liée à la seule désuffixation par EDA est légèrement meilleure que Carry (+0.60%). Ceci s'explique par le fait que Carry propose un traitement important des verbes conjugués et adjectifs qui n'ont aucun impact sur la recherche de diagnostics. Si l'on observe les résultats sur la K-mesure qui intègre le besoin et le niveau d'exigence de l'utilisateur, le gain est plus important.

3.7.3 Conclusion et perspectives

Sur 25 275 termes différents présents dans 30 000 comptes rendus, 10 602 ont été regroupés, soit 42%. L'utilisation de cet algorithme de désuffixation nous a permis de mesurer une amélioration de 5.49 %. Les deux tiers du gain résultent de la désuffixation, le dernier tiers de l'enrichissement des documents par la recherche de racines lexicales des termes médicaux. Ces résultats sont supérieurs à ceux obtenus avec le désuffixeur Carry (+3.1 %) non enrichi. Cet algorithme est très simple à implémenter et ne nécessite aucune base de données, aucun traitement lourd, ni aucune mise à jour. Il est très rapide à implémenter et en exécution.

3.8 Proposition d'un algorithme de classification : CLO3

Avant de développer un algorithme spécifique, nous avons testé et évalué les algorithmes traditionnellement utilisés pour la classification automatique des documents textuels. Par exemple, l'expérimentation de Naive Bayes donne les résultats suivants :

Précision	Rappel	F-Mesure	K-Mesure
0.734	**0.669**	**0.700**	**0.490**

Figure 35. Classification avec Naive Bayes

Ces résultats ne nous sont pas apparus satisfaisants, aussi avons-nous cherché à explorer de nouvelles pistes.

3.8.1 L'algorithme CLO3

L'algorithme CLO3 se base sur le principe que les termes utilisés sont en rapport direct avec les codes diagnostics. Par exemple, nous supposons que le terme 'coma' est associé au diagnostic de coma. Pour quantifier cette relation, nous allons donc calculer pour chaque terme une première variable appelée 'Poids Brut', définie comme suit :

Poids brut = Variance de la fréquence (couple) / moyenne de fréquence (couple)

Un couple est l'association d'un code CIM10 avec un concept.
Il s'agit donc d'un coefficient de variation qui va nous permettre de quantifier la concentration de chaque terme ou concept.

Mais étant donné le nombre de diagnostics associés, il fallait trouver une méthode qui permette de supprimer les mots trop rares (pour ne pas leur donner un poids injustement trop fort). L'étape suivante de l'algorithme consiste à calculer un second poids, appelé 'poids net' qui permette de quantifier la relation entre un terme et un diagnostic (que nous appelons couple) :

Poids net = Poids brut * fréquence (couple) * effectif (couple)

Nous multiplions par la fréquence et l'effectif pour faire apparaître en priorité les co-occurrences fréquemment rencontrées. Ainsi, les diagnostics associés seront supprimés automatiquement. Supposons en effet qu'un patient souffre à la fois de diabète et d'une crise cardiaque. Dans le compte rendu, nous trouverons par exemple le mot diabète et le mot cardiaque. Nous allons donc avoir les quatre relations suivantes :

Terme	Diagnostic
Cardiaque	**Crise cardiaque**
Diabète	**Crise cardiaque**
Cardiaque	**Diabète**
Diabète	**Diabète**

Mais sur l'ensemble des comptes rendus, la fréquence d'apparition des couples 'terme cardiaque – diagnostic diabète' et 'terme diabète – diagnostic crise cardiaque' sera faible.

	Diagnostic crise cardiaque	Diagnostic diabète
Concept crise cardiaque	**Effectif fort à majorer**	**Effectif faible à minorer**
Concept diabète	**Effectif faible à minorer**	**Effectif fort à majorer**

Donc en multipliant par cette fréquence, on supprime ces relations non désirées. En multipliant par l'effectif, on amplifie la même fonctionnalité.

La troisième étape consiste à standardiser les valeurs calculées pour les poids nets, en calculant le poids final, appelé 'PoidsA'. Pour cela, nous divisons le poids de chaque couple par la moyenne du poids de chaque classe de diagnostic. Notre objectif est de pouvoir accentuer ce poids en le mettant à une puissance 'n', cette formule nous permet d'avoir des valeurs inférieures à 1 si elles sont inférieures à la moyenne. Ainsi, lorsque nous élèverons ce résultat à une puissance quelconque supérieure à 1, les valeurs inférieures à la moyenne seront diminuées alors que les autres seront augmentées. Cette fonction puissance va donc permettre d'amplifier le résultat.

La quatrième étape consiste à créer un second indicateur baptisé 'PoidsB' inspiré des probabilités simples et Naive Bayes mais en extrapolant ce raisonnement. Pour chaque couple (terme, diagnostic), nous calculons :

PoidsB = Effectif du couple / Effectif total du terme

A partir de ces deux termes, nous cherchons une fonction qui optimise le résultat de la classification en se basant sur la fonction suivante :

$Poids = PoidsA^{a} * PoidsB^{b}$

Nous avons obtenu les meilleurs résultats avec a=2 et b=5, ce qui nous donne :

$Poids\ CLO3 = PoidsA^{2} * PoidsB^{5}$

Le poids CLO3 donne une relation entre un concept et un diagnostic (ou plus généralement entre un descripteur et une classe). Pour Classifier un nouveau document, il faut sommer les poids CLO3 et retenir les meilleurs scores supérieurs à 0,0005.

L'expérimentation de cet algorithme (sans désuffixation) donne les résultats suivants :

Précision	**Rappel**	**F-Mesure**	**K-Mesure**
0.804	**0.733**	**0.767**	**0.589**

Figure 36. Résultats obtenus avec CLO3

La comparaison entre Naive Bayes et CLO3 sur le niveau 2 de la CIM donne donc le tableau final suivant :

Algorithme	Précision	Rappel	F-Mesure	K-Mesure
CLO3	0.804	0.733	0.767	0.589
Naive Bayes	0.734	0.669	0.700	0.490
Gain	0.070	0.064	0.067	0.099

Figure 37. Comparaison entre Naïve Bayes et CLO3

Nous observons que les gains par rapport à Naive Bayes sont importants tant pour le rappel que pour la précision, avec un gain global sur la F-mesure de 6,7 % et de 9,9% pour la K-mesure. Ce qui signifie que l'accroissement de la satisfaction de l'utilisateur est bien plus important. Nous obtenons des résultats similaires avec OHSUMED.

3.8.2 Bilan comparatif

Pour finir, nous avons souhaité comparer CLO3 avec les principales autres méthodes. Nous avons comparé avec Knn (en faisant varier K), SVM, Naive Bayes, et TF IDF / RM (Régression Matricielle) [Sandu popa 2007] qui améliore l'algorithme déjà TF/IDF classique. Les résultats sont les suivants, au niveau 4 de hiérarchie de la CIM :

Méthode	Précision	Rappel	F-Mesure
Knn	0.3300	0.2400	0.2779
SVM	0.3300	0.2980	0.3132
Naive Bayes	0.3295	0.3549	0.3417
TF IDF / RM	0.3992	0.3671	0.3825
CLO3	0.3902	0.4161	**0.4027**

Figure 38. Comparaison de CLO3 avec les autres algorithmes

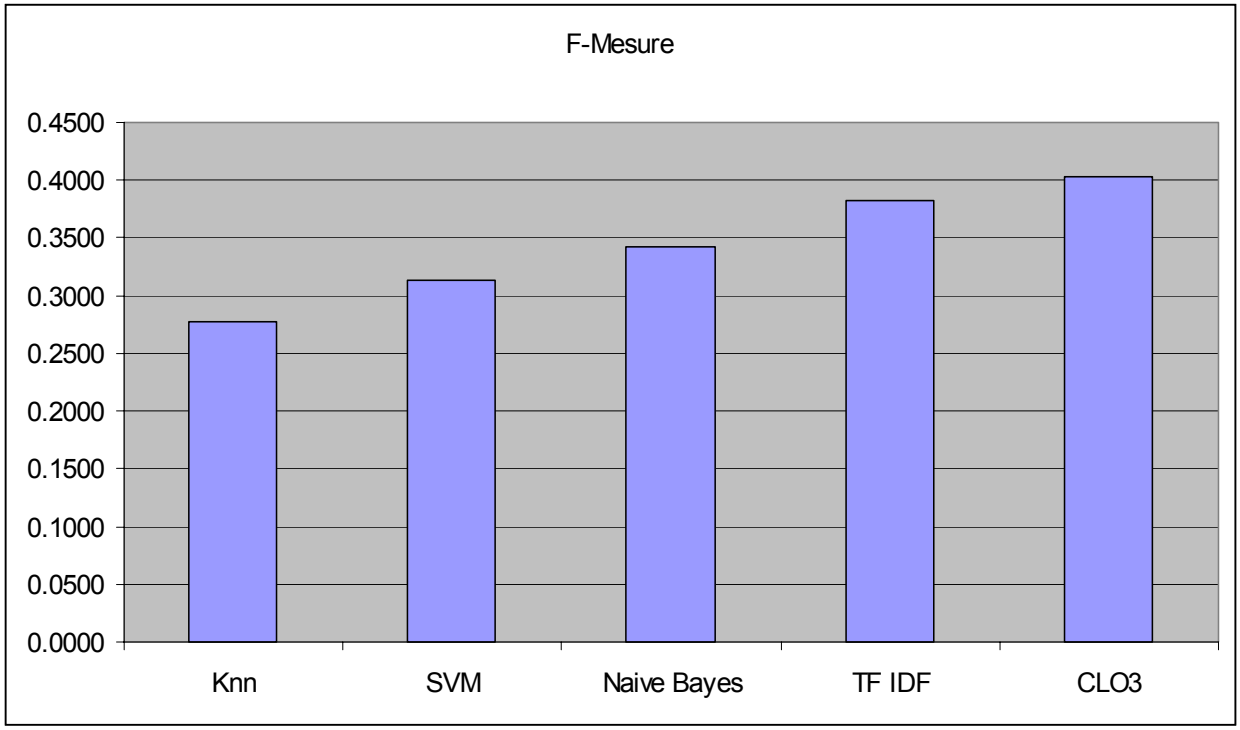


Figure 39. Comparaison graphique de CLO3 avec les autres algorithmes

Nous observons également qu'avec CLO3 les bons codes diagnostics apparaissent dans les premières positions :

	Position moyenne sur 7 codes proposés (optimum de la F-mesure)
CLO3	3.32
TFIDF / RM	3.53

Figure 40. Position des bonnes réponses avec CLO3

3.8.3 Pourquoi obtenons-nous de meilleurs résultats avec CLO3 ?

Knn donne des résultats médiocres à cause des caractéristiques des patients admis en réanimation. En effet, la diversité et la complexité des cas ne permet pas d'assimiler le voisinage comme facteur de similitude diagnostique.

SVM ne donne pas de bons résultats à cause du bruitage. Il n'existe pas de forte concentration principalement à cause de l'étiquetage multiple qui bruite les performances.

Nous avons le même problème avec Naïve Bayes dont les résultats sont meilleurs.

Enfin TF IDF donne de meilleures performances, à condition d'utiliser la variante de régression matricielle, c'est-à-dire en sommant les vecteurs résultats. Le motif principal est qu'il existe bien une relation entre les termes ou concepts utilisés et les diagnostics codés.

Nous obtenons de meilleurs résultats avec CLO3 car notre méthode s'inspire à la fois de TF IDF et de Naïve Bayes, mais les phases préalables d'accentuation permettent de diminuer le bruitage.

3.8.4 Conclusion et perspectives

Cet algorithme nous permet d'obtenir de très bons résultats puisqu'il améliore de près de 7% les algorithmes analogues. Néanmoins, nous devons encore procéder à une évaluation manuelle par les praticiens. En effet, [Misset 2005] a démontré que le taux de reproductibilité du codage était très faible, donc certains codes considérés comme faux sont peut être exacts. De plus, nous pensons que la reproductibilité devrait être améliorée.

Pour le praticien, l'avantage de cette méthode est que les codes générés peuvent être explicités. Nous offrons la possibilité qu'il puisse manuellement modifier les poids en vue d'une personnalisation et/ou d'une optimisation.

Enfin, nous pensons explorer deux autres pistes d'optimisation : la première concerne l'ajout de règles, la seconde consiste à identifier le nombre de diagnostics total à proposer en fonction du texte et des premiers codes diagnostics identifiés.

3.9 Optimisation de l'apprentissage

Il existe de nombreuses possibilités permettant d'optimiser l'apprentissage, que ce soit au niveau de la précision, du rappel, voire des temps de traitement et de réponses. Ce chapitre propose de présenter les différents paramètres que nous avons explorés pour optimiser nos résultats.

3.9.1 Exploitation de la structure hiérarchique

[Koller et Sahami 1997] a implémenté Naïve Bayès en exploitant la structure hiérarchique de la classification. Nous avons implémenté cette méthode pour vérifier si les gains sur la F-mesure étaient sensibles. L'amélioration apportée n'est pas significative et n'a pas été retenue.

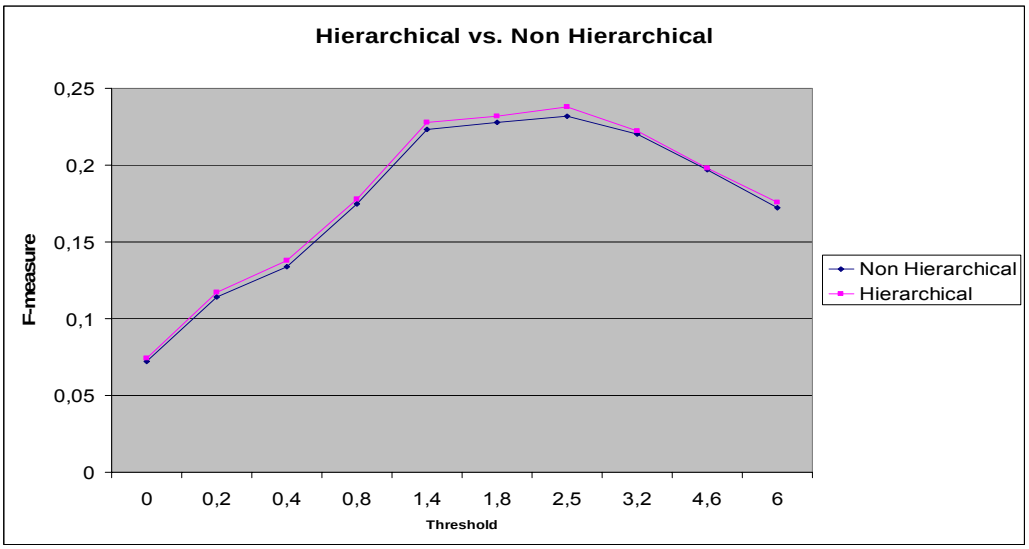


Figure 41. Comparaison k-NN hiérarchique/non hiérarchique – F-mesure

3.9.2 Importance de la source

Nous avons constaté que la source d'apprentissage améliorait la reproductibilité du codage. Par exemple, les prédictions pour un praticien sont meilleures si le corpus d'apprentissage ne contient que ses propres comptes rendus. L'algorithme CLO3 reproduit donc les habitudes des praticiens. Mais cela ne signifie par pour autant que le codage proposé est plus pertinent, simplement qu'il est plus conforme aux habitudes du praticien. Nous avons obtenu une amélioration sensible, mais un des objectifs étant d'améliorer la reproductibilité du

codage, nous avons préféré conservé notre base d'apprentissage avec la totalité des comptes rendus :

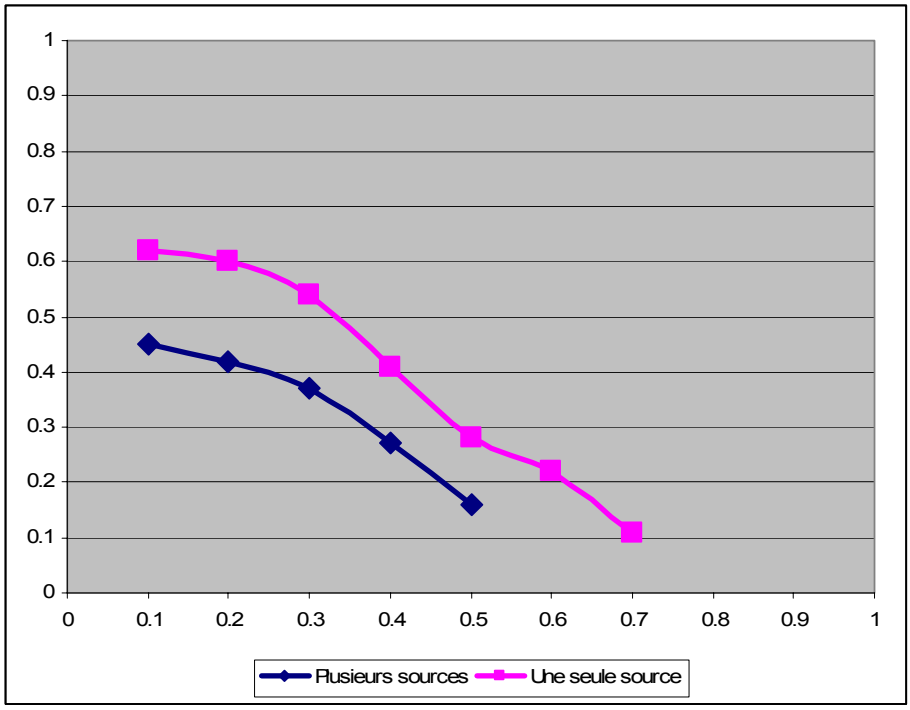


Figure 42. Impact de la source d'apprentissage

3.9.3 Analyse du résultat en fonction du nombre de CRH en apprentissage

Nous avons souhaité mesurer l'impact de la taille d'apprentissage sur les résultats obtenus :

Nombre de CRH appris	F-Mesure	Précision	Rappel
30000	0.4055	0.4266	0.3864
25000	0.3860	0.4070	0.3671
20000	0.3725	0.3659	0.3794
15000	0.3719	0.3949	0.3514
10000	0.3510	0.3743	0.3304
8000	0.3373	0.3930	0.2955
6000	0.3230	0.3191	0.3269
4000	0.2917	0.3403	0.2552
2000	0.2857	0.2847	0.2867
1500	0.2884	0.3105	0.2692
1000	0.2828	0.3044	0.2640
800	0.2769	0.2978	0.2587
600	0.2669	0.2681	0.2657
400	0.2716	0.2903	0.2552
300	0.2610	0.3066	0.2273
250	0.2415	0.2401	0.2430

Figure 43. Evolution des performances selon la taille du corpus d'apprentissage

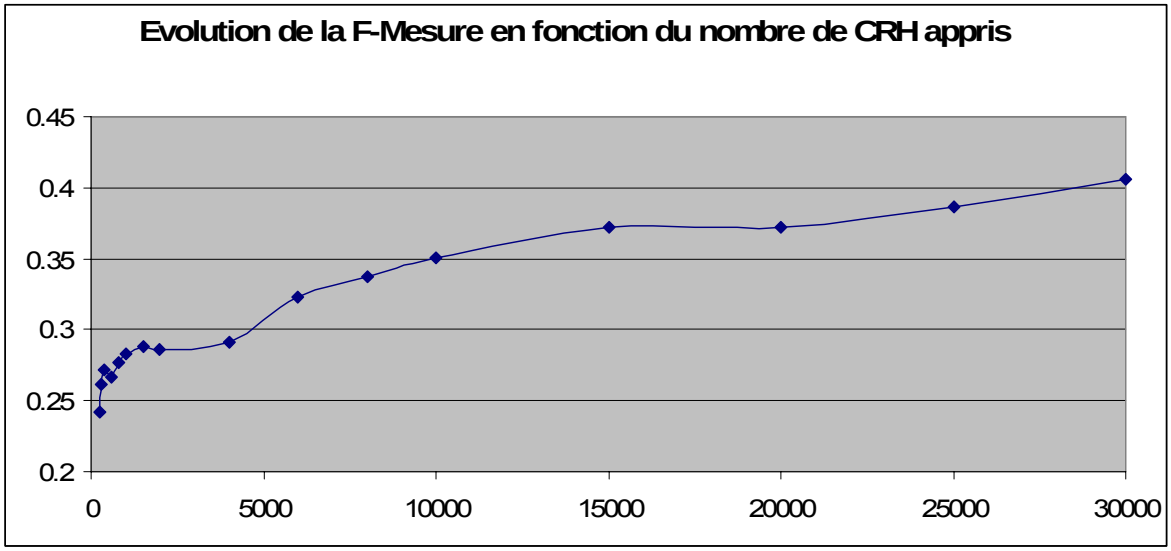


Figure 44. Evolution de la F-mesure en fonction de la taille d'apprentissage

Nous constatons que plus la base d'apprentissage est importante, plus les performances sont élevées. La courbe semble néanmoins ressembler à une fonction de type logarithme, c'est-à-dire que pour gagner un peu en performance, il faudrait produire beaucoup de CRH. Compte tenu de ces résultats, nous avons choisi de réaliser l'apprentissage sur la totalité du corpus.

3.9.4 Analyse du nombre de couples

Nous avons souhaité vérifier si le nombre de couples prédicteurs (Code CIM, Concept) conservés pouvait impacter la qualité de la prédiction et dans quelle mesure les temps de prédiction en étaient affectés. Nous avons réalisé nos tests sur 100 CRH prédits.

Nombre de lignes	Temps de traitement (en secondes)	Précision	Rappel	F-Mesure
344787	73	.4205	.3881	.4036
300000	62	.4205	.3881	.4036
250000	52	.4205	.3881	.4036
200000	47	.4205	.3881	.4036
150000	43	.4205	.3881	.4036
100000	36	.4205	.3881	.4036
80000	35	.4205	.3881	.4036
60000	32	.4205	.3881	.4036
40000	29	.4205	.3881	.4036
30000	26	.4213	.3881	.4040
25000	24	.4221	.3881	.4044
20000	22	.4221	.3881	.4044
15000	22	.4221	.3881	.4044
10000	18	.4221	.3881	.4044
8000	18	.4221	.3881	.4044
6000	17	.4237	.3881	.4051
4000	14	.4245	.3881	.4055
2000	11	.4277	.3881	.4070
1500	11	.4266	.3864	.4055
1000	11	.4322	.3846	.4070
800	11	.4339	.3846	.4078
600	11	.4369	.3811	.4071
400	11	.4392	.3724	.3946
300	10	.4600	.3619	.4051
250	10	.4732	.3549	.4056

Figure 45. Evolution selon le nombre de prédicteurs conservés

Sur les 344 787 couples générés par l'algorithme, nous constatons que le rappel est inchangé jusqu'à ce que l'on ne conserve que 2000 concepts. Dans ce cas la précision augmente, et par conséquent la F-mesure. Cette amélioration est factice car en diminuant le nombre de couples, le logiciel propose moins de codes candidats, ce qui améliore la précision. Ces résultats sont conformes aux travaux de [Dumais et al. 1998]. Par contre, les temps de traitement sont nettement améliorés. Nous avons utilisé ces résultats dans l'implémentation finale du logiciel.

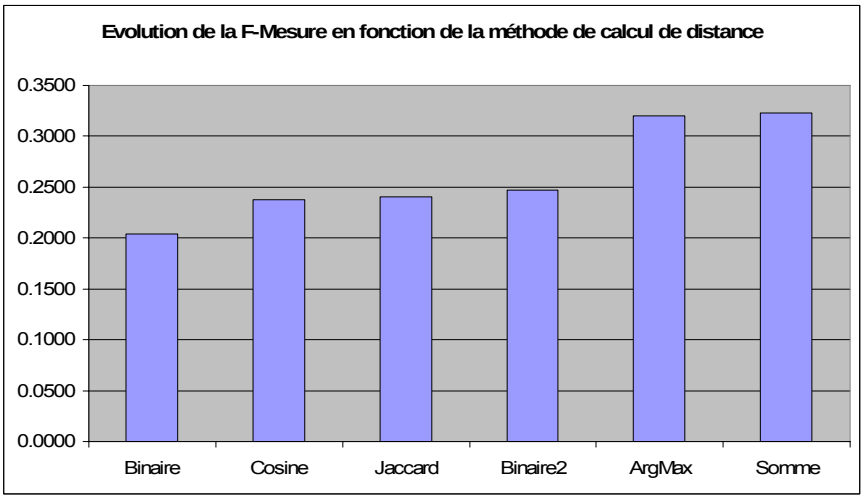
3.9.5 Comparaison des fonctions de distance

Nous avons cherché à identifier quelle était la meilleure fonction de calcul de distance pour optimiser la F-Mesure. Pour ce comparatif, nous avons utilisé les fonctions de distance suivantes :

- somme,
- cosine,
- distance de Jaccard,
- argument maximum,
- binaire,
- binaire2.

Evaluateur	F-Mesure	Précision	Rappel	Mode de calcul
Binaire	0.2041	0.1759	0.2430	somme du nombre de descripteurs communs (en positif et négatif)
Cosine	0.2381	0.1970	0.3007	$sim(D_i, D_j) = \frac{\sum_{k=1}^{t} weight_{ik} . weight_{jk}}{\sqrt{\sum_{k=1}^{t} weight_{ik}^2 weight_{jk}^2}}$
Jaccard	0.2401	0.1934	0.3164	$sim(Di,Dj) = \frac{\sum_{k=1}^{t} \min(weight_{ik}, weight_{jk})}{\sum_{k=1}^{t} weight_{ik}^2 + \sum_{k=1}^{t} weight_{jk}^2 - \sum_{k=1}^{t} weight_{ik}.weight_{jk}}$
Binaire2	0.2473	0.2576	0.2378	somme du nombre de descripteurs présents et communs
ArgMax	0.3198	0.3392	0.3024	nous ne retenons que le meilleur résultat
Somme	0.3230	0.3191	0.3269	Sommation de tous les résultats

Figure 46. Evolution des performances selon la fonction de distance



La fonction somme donne les meilleurs résultats.

3.9.6 Choix du dictionnaire d'apprentissage

Nous nous sommes posé la question de savoir quel serait le dictionnaire de termes le plus efficient pour extraire les concepts. Nous avons utilisé 2 dictionnaires : un dictionnaire des concepts médicaux, et un dictionnaire complet de tous les termes du français. Nous les avons utilisés tels quels mais en essayant deux variantes. La première variante consiste à compter chaque concept autant de fois qu'il apparaît dans le compte rendu (avec doublons). La seconde variante consiste à fusionner les deux dictionnaires et ne garder que les concepts présents dans le corpus d'apprentissage. C'est cette dernière option qui obtient les meilleurs résultats.

Dictionnaire	F-Mesure
Synthèse des dictionnaires	0.4074
Dictionnaire médical	0.4055
Dictionnaire Français complet sans doublons	0.4035
Dictionnaire Français complet avec doublons	0.3916
Dictionnaire médical hors mots du dictionnaire complet	0.3384

Figure 47. Evolution des performances selon le dictionnaire utilisé

Nous observons par ailleurs que le fait de considérer un terme autant de fois qu'il apparaît n'apporte aucune information supplémentaire, voire introduit un bruit supplémentaire, statistiquement non significatif. Ces résultats sont conformes à ce qui avait été observé par [Brun 2003].

3.9.7 Comparaison des rubriques d'apprentissage

Enfin, nous avons cherché à identifier quelles rubriques du compte rendu sont les plus chargées d'informations susceptibles d'être classificatoires, que ce soit en apprentissage ou en prédiction. Pour cela, nous avons à partir d'un même dictionnaire, réalisé l'apprentissage sur plusieurs rubriques et testé les prédictions sur l'exploitation d'autres rubriques. Les résultats sont très nets, quels que soient les dictionnaires d'apprentissage : les rubriques motif du séjour et conclusions, en apprentissage et en prédiction donnent les meilleurs résultats. Tous les autres choix montrent une détérioration rapide des résultats.

Champs apprentissage	Champs évalués	F-Mesure
Motif et conclusions	Motif et conclusions	0.4035
Motif, histoire de la maladie, examen clinique et conclusions	Motif, histoire de la maladie, examen clinique et conclusions	0.3522
Motif	Motif et conclusions	0.3519
Motif	Motif	0.3206
Motif	Motif, examen clinique, évolution et conclusions	0.2560
Tous	Tous	0.2114

Figure 48. Evolution des performances selon les rubriques du CRH utilisées

3.9.8 Impact du nombre de mots composant un concept

Nous avons imaginé que la longueur d'un concept, plus précisément le nombre de termes qui le composent seraient susceptible d'améliorer la qualité de la prédiction. Par exemple, nous pouvions supposer que le concept "syndrome de détresse respiratoire aigu de l'adulte" ou " accident vasculaire cérébral ischémique" puisse être plus riche d'informations qu'un concept simple composé d'un seul terme. Pour vérifier cette possibilité, nous avons essayé d'évaluer CLO3 rehaussé par le nombre de termes. Le tableau suivant donne les résultats obtenus :

Fonction de valorisation	**F-mesure**
CLO3	0.4058
CLO3*nombre de termes	0.4048
CLO3*10 $^{\text{nombre de termes}}$	0.3956

Figure 49. F-mesure selon nombre de termes du concept

Nous constatons que cette hypothèse est fausse et que les résultats obtenus se détériorent. Cette piste a donc été abandonnée.

3.9.9 Conclusions sur l'optimisation

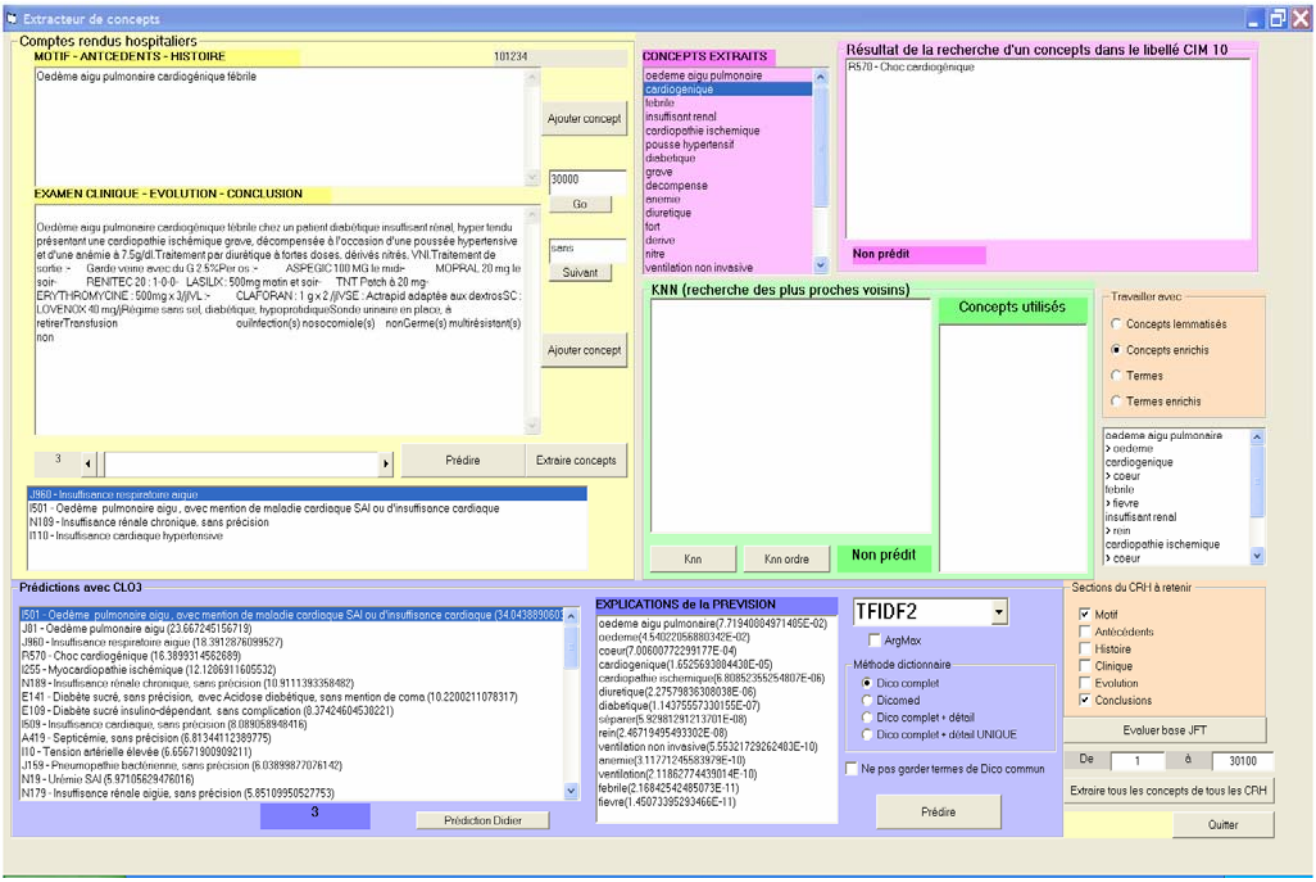
Au final, les meilleures performances sont obtenues en procédant à l'extraction des concepts depuis un dictionnaire ad hoc en ne comptabilisant qu'une seule fois un concept apparaissant plusieurs fois. Ce dictionnaire est un mélange des termes du vocabulaire français et des termes médicaux. L'apprentissage et la restitution ne portent que sur les rubriques motifs du séjour et conclusions. L'apprentissage s'est opéré sur la totalité de notre base de données. Enfin, seuls 90000 couples sont conservés pour optimiser les temps de réponse. Enfin, il est important de noter que ces paramètres d'optimisation pour l'algorithme CLO3 donnent les mêmes résultats avec TF/IDF.

3.10 Exemple d'utilisation du logiciel

Nous avons dans un premier temps développé un prototype de recherche qui nous a permis de mettre au point les algorithmes, de comparer les résultats obtenus en fonction des différents paramètres, de calculer les valeurs optimales pour la F-mesure. Ce travail nous a permis d'implémenter une version finalisée dans le logiciel client.

3.10.1 Le prototype de recherche

Voici un des écrans du prototype de recherche :



Dans la zone en haut à gauche, nous trouvons le compte rendu original, ainsi que le codage initial. Les concepts extraits apparaissent en haut au centre. En choisissant un concept, le logiciel le recherche dans le libellé de la CIM et l'affiche. La zone centrale permet de faire le codage en utilisant Knn.

La zone basse permet de choisir un algorithme pour faire la prévision (CLO3, TFIDF, SVM, Naïve Bayes). Si CLO3 est utilisé, le logiciel peut expliquer les raisons de ses choix, c'est-à-dire montrer la force de la relation entre un diagnostic et les concepts extraits. Les zones à

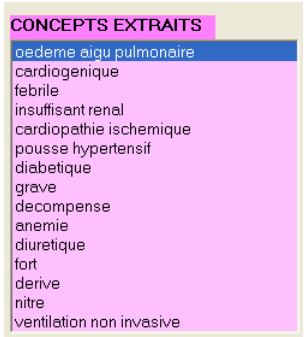
droite permettent d'utiliser des variantes. Par exemple choisir certaines sections du CRH ou la méthode utilisée pour la représentation vectorielle (lemmes, concepts, …). Enfin, une dernière option permet de choisir entre différents dictionnaires.

Le compte rendu original est celui-ci :

Motif : Œdème aigu pulmonaire cardiogénique fébrile.

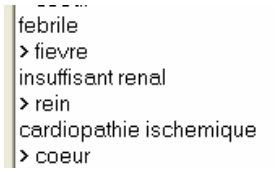
Conclusions : Œdème aigu pulmonaire cardiogénique fébrile chez un patient diabétique insuffisant rénal, hyper tendu présentant une cardiopathie ischémique grave, décompensée à l'occasion d'une poussée hypertensive et d'une anémie à 7.5g/dl.Traitement par diurétique à fortes doses, dérivés nitrés, VNI.

Concepts extraits :

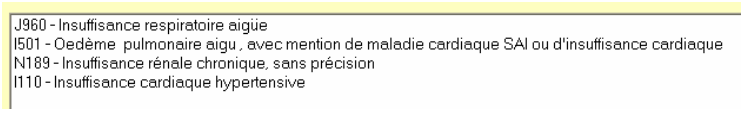


Il s'agit des concepts lemmatisés.

L'enrichissement donne les termes suivants :

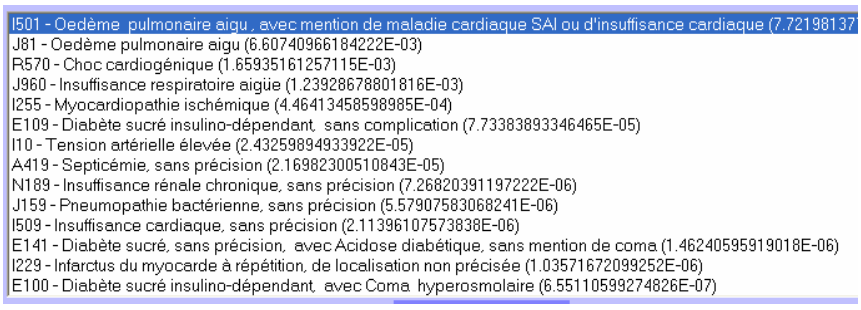


Diagnostics portés par le praticien :

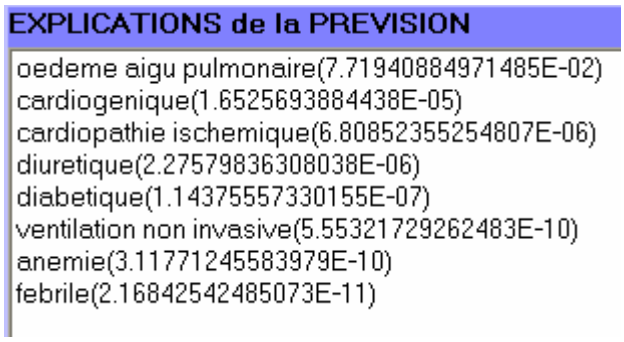


Effectuons maintenant une prédiction avec CLO3. Les résultats obtenus sont les suivants :

Prédictions avec CLO3 :

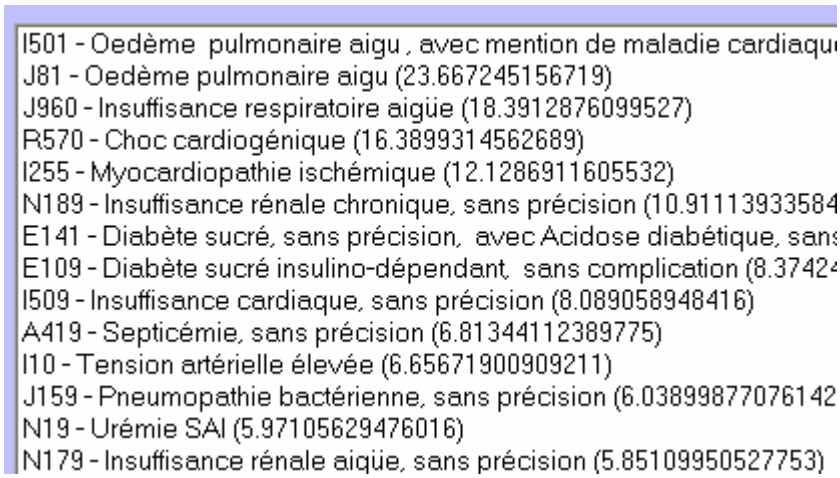


Le premier diagnostic proposé est "I501 – Œdème pulmonaire aigu, avec mention de maladie cardiaque". Nous pouvons obtenir des précisions sur ce diagnostic :

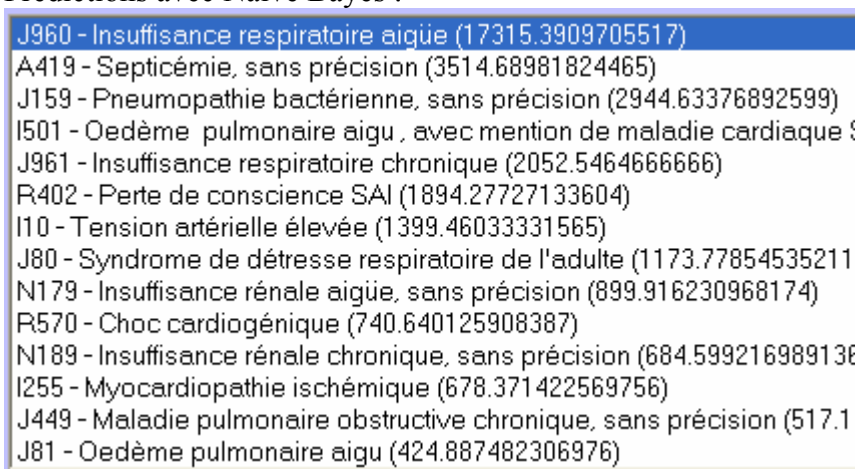


Nous observons que le diagnostic est proposé car les concepts "œdème aigu pulmonaire", "cardiogénique", …. Ont été trouvés dans le compte rendu.

Prédictions avec TF/IDF Matrix Regression :



Prédictions avec Naïve Bayes :



Le prototype de recherche dispose également d'un outil d'évaluation. Pour une série de comptes rendus déjà étiquetés, il compare les codes proposés avec les codes prédits et calcule la F-Mesure. Par ailleurs, il calcule également l'optimum du nombre de codes à proposer et la F-mesure correspondante :

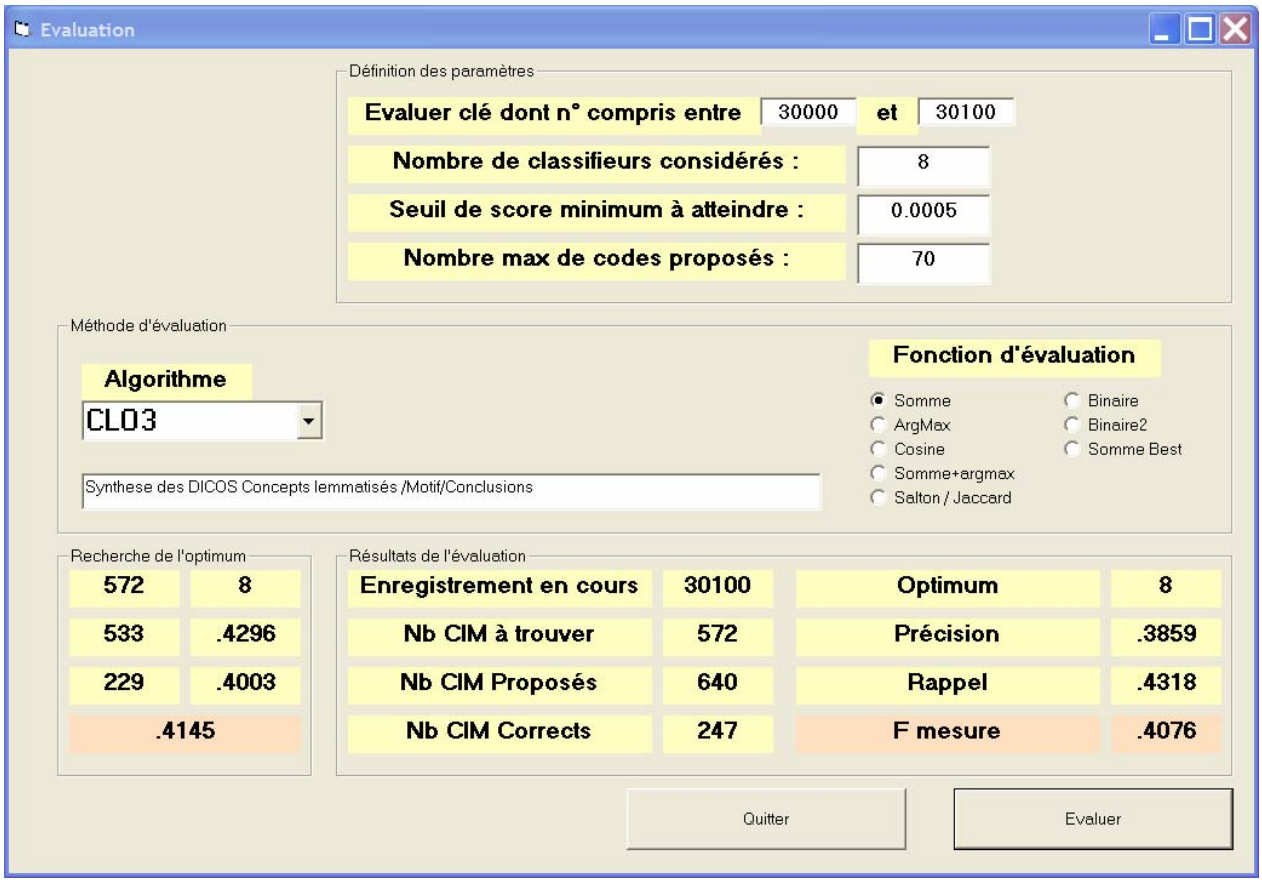


Figure 50. Prototype pour l'évaluation

3.10.2 Le logiciel client

Une version finalisée est implémentée dans le logiciel client. Le compte rendu est repris depuis les données saisies en observation. L'utilisateur peut choisir les rubriques à analyser. Le logiciel lui propose automatiquement la liste des concepts extraits depuis le compte rendu, la liste des codes CIM proposés, les explications correspondant à chaque code CIM :

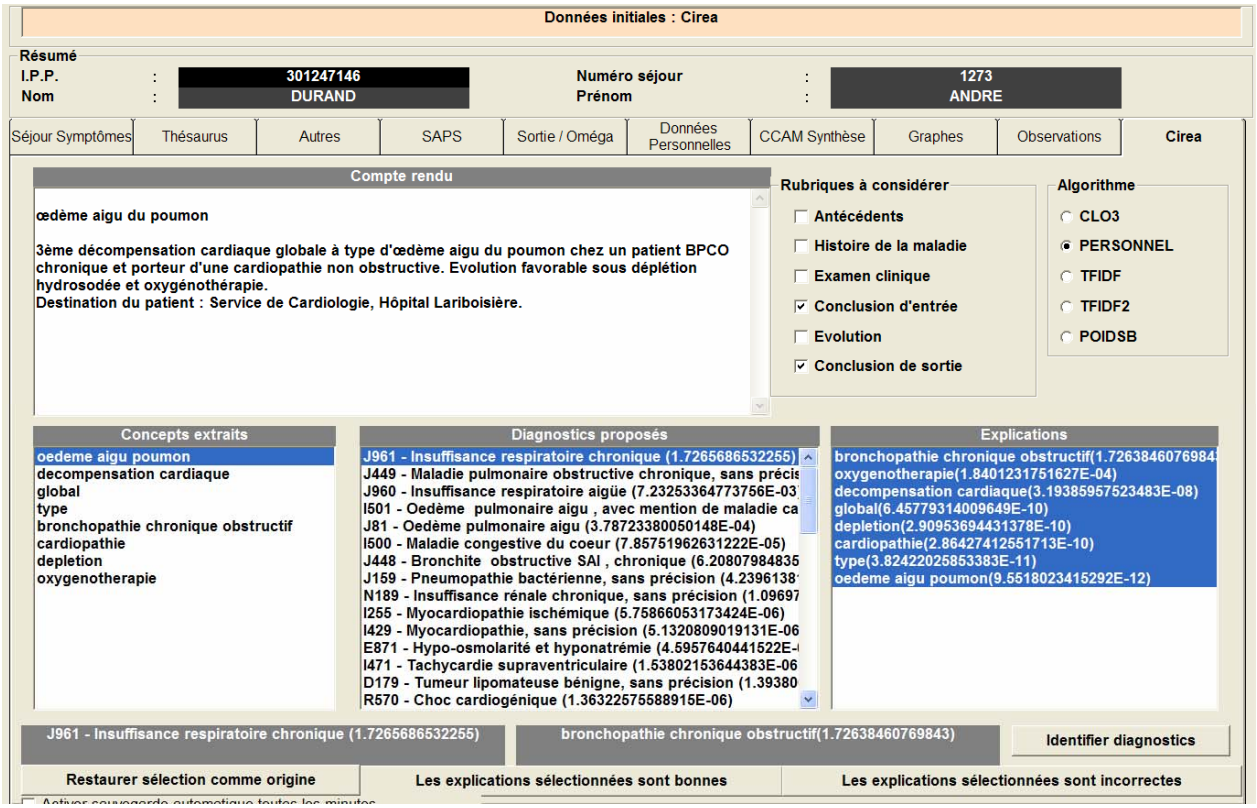


Figure 51. Le logiciel client de classification automatique

L'utilisateur peut personnaliser la prédiction en réajustant les poids.

CHAPITRE 4
LA PROBLEMATIQUE DE L'EVALUATION

4.1 Présentation de la problématique

L'évaluation est probablement le point le plus délicat puisque celui qui conçoit ou implémente une technique est souvent celui qui évalue. Les paramètres sont nombreux et certains problèmes sont récurrents :

- la modélisation ne sera-t-elle pas trop proche de la base d'apprentissage ? Une technique obtenant de bons résultats dans un contexte donné ne signifie pas que l'algorithme est bon mais que la modélisation et son application dans le contexte sont satisfaisantes. Les résultats peuvent changer dans un autre contexte. De plus, on a rarement une idée précise de la qualité de la base d'apprentissage. Donc si elle est mauvaise, les résultats seront à son image.
- Quel est le niveau de performance acceptable dans un contexte industriel ? Dans certains cas, un niveau bas peut offrir un niveau de service suffisant alors que pour les problèmes de médecine par exemple, cela devient une exigence très forte.
- Ce n'est pas parce que le logiciel proposera une classification non conforme au modèle d'apprentissage que celui ci sera faux. Or l'évaluation se fait à partir de documents étiquetés. On va donc, dans ce cas, considérer une erreur qui n'en sera pas une.

Nous considérons que l'évaluation consiste à mesurer la différence entre un résultat attendu et un résultat obtenu. Aucun consensus de métrique n'y est associé mais on utilise en général un indicateur compris entre 0 et 1 pour en faciliter l'interprétation, et souvent avec une approche booléenne.

4.2 Etat de l'art de l'évaluation des classifications

4.2.1 Premières approches

En matière de recherche d'informations, de nombreux chercheurs considèrent que l'évaluation consiste en une mesure de la pertinence [SAR 70]. Tefko Saracevic la définit comme suit :

- la correspondance entre un résultat et une demande, une mesure d'informativité du document à la requête,
- un degré de relation (chevauchement, relativité, …) entre le document et la requête,
- un degré de la surprise qu'apporte un document, qui a un rapport avec le besoin de l'utilisateur,
- une mesure d'utilité pour l'utilisateur,
- …

Mais les notions utilisées (informativité, relativité, surprise, …) sont très subjectives et difficilement associables à une métrique automatisée. Tefko Saracevic propose la définition de synthèse suivante pour tenir compte de cette influence multiple sur la pertinence : la pertinence est la A d'un B existant entre un C et un D jugé par un E.

A = mesure de la pertinence

B = la pertinence absolue. Cette relation de pertinence absolue peut exister, ne pas exister, ou exister à certain degré, entre un document et le contexte de recherche.

C = un document

D = le contexte dans lequel la pertinence est mesurée (y compris le besoin d'information). Le contexte D contient aussi bien l'expression du besoin que les facteurs contextuels qui influencent le jugement de pertinence.

E = le juge (l'utilisateur).

Il met en avant l'importance du contexte sur la pertinence, ainsi que l'utilisateur (le juge). Par conséquent, si l'un de ces facteurs varie, la notion de pertinence change aussi. Cette définition, bien qu'elle ne donne aucune échelle ou méthode de calcul, permet d'identifier les acteurs.

[Van Rijsbergen 79] aborde le problème de la mesure de l'évaluation en se focalisant sur la précision et le rappel mais il écarte le problème de la mesure de la pertinence car il considère que les facteurs susceptibles de la représenter ne sont pas mesurables car trop subjectifs. Aussi poursuit-il son étude sur les couples (précision, rappel) pour aboutir à la F-mesure.

Un système peut répondre à une requête selon le modèle suivant :

	Pertinent	Non pertinent	Total
Retrouvé (ou proposé)	a	b	a+b
Non retrouvé (ou non proposé)	c	d	c+d
	a+c	b+d	a+b+c+d=N

A partir de ce tableau de contingence, la communauté TALN calcule diverses mesures :

la précision (Precision) = $\frac{a}{a+b}$, le rappel (Recall) = $\frac{a}{a+c}$, la pertinence= $\frac{a+d}{a+b+c+d}$ L'erreur = $\frac{b+c}{a+b+c+d}$, le taux de chute, ou taux de documents non pertinents ramenés (Fallout) = $\frac{b}{b+d}$, le silence = 1 – rappel = $\frac{c}{a+c}$, la spécificité = $\frac{d}{b+d}$, le bruit = $\frac{b}{a+b}$ et l'overlap = $\frac{a}{a+b+c}$

Une autre mesure, la généralité (G) représente la densité de documents pertinents par rapport au nombre total de documents dans le référentiel et est défini par G=a/N

Au final, 4 mesures (a, b, c, d) génèrent 10 indicateurs de base, eux-mêmes combinés pour donner d'autres mesures. Dans la grande majorité des cas, on se limite à la précision et au rappel.

A partir de ces différentes mesures, plusieurs indicateurs ont été créés, mais le plus usuel est la F-Mesure de [Van Rijsbergen 79] :

Mesure F = ((1+ß²)*Précision*Rappel) / (ß²*(Précision+Rappel))

Mais elle est souvent utilisée sous cette forme, (paramètre ß² = 1) :

Mesure F = (2*Precision*Rappel) / (Precision+Rappel)

Lorsque la précision et le rappel sont identiques (et c'est ce que l'on souhaite atteindre), on a F = précision = rappel, ce qui en facilite l'interprétation. On cherche habituellement un algorithme qui optimise à la fois la précision et le rappel, si possible en leur attribuant la même valeur. En général, l'augmentation de l'un se fait au détriment de l'autre.
La difficulté est alors de comparer ces indicateurs mais sur une même base ou jeu de données. On trouve principalement des comparaisons sur la base Reuters, qui est une classification de dépêches de presse. Mais les comparaisons sont souvent faussées car cette base évolue et les auteurs ne comptent pas toujours de la même manière. Pour les requêtes dans le domaine médical, on se réfère également à la base OHSUMED. Enfin, citons dans le domaine des requêtes la conférence TREC qui met en compétition chaque année des équipes qui expérimentent divers algorithmes dont elles ont la maîtrise.

[Lewis 1991] présente une analyse comparative de la précision, du rappel, et du taux de chute (*fallout*). Mais comment calculer la précision globale (ou le rappel) d'un algorithme sur un ensemble de documents ? [Salton 1983] propose la méthode des 11 points de moyenne, à partir de la meilleure valeur dans chaque tranche de rappel variant de 0 à 1, par incrément de 0.10 et par extrapolation.

Dans le cas de l'évaluation d'un algorithme de classification, la F-mesure est fournie comme un résultat global de l'algorithme. Elle est utilisée pour évaluer la performance d'une ou de chacune des classes ([Dumais 98] et [Yang et Liu 99] par exemple).

On distinguera la macro moyenne (*macro average*) de la micro moyenne (*micro average*). La macro moyenne calcule une moyenne globale, tandis que la micro moyenne calcule la moyenne des mesures de chaque catégorie [Yang 1998]. [Sebastiani 2004] propose la synthèse suivante :

	Micro averaging	Macro averaging
Precision	$\pi = \frac{\sum_{i=1}^{\lvert C\rvert} TP_i}{\sum_{i=1}^{\lvert C\rvert} TP_i + FP_i}$	$\pi = \frac{\sum_{i=1}^{\lvert C\rvert} \pi_i}{\lvert C\rvert} = \frac{\sum_{i=1}^{\lvert C\rvert} \frac{TP_i}{TP_i + FP_i}}{\lvert C\rvert}$
Rappel	$\rho = \frac{\sum_{i=1}^{\lvert C\rvert} TP_i}{\sum_{i=1}^{\lvert C\rvert} TP_i + FN_i}$	$\rho = \frac{\sum_{i=1}^{\lvert C\rvert} \rho_i}{\lvert C\rvert} = \frac{\sum_{i=1}^{\lvert C\rvert} \frac{TP_i}{TP_i + FN_i}}{\lvert C\rvert}$

Figure 52. Micro moyenne et macro moyenne pour la F-mesure
TPi, TNi, FPi, FNi représentent respectivement les vrais positifs, vrais négatifs, faux positifs, et faux négatifs.

[Moschitti 2003] expose en détail son mode de calcul de l'approche du Break Even Point (BEP) de [Joachims 98]. Cet indicateur se calcule à partir de la précision et du rappel. Il consiste à trouver le couple (précision, rappel) où les deux mesures sont maximisées et leurs valeurs identiques.

[Sparck Jones 2001] propose de repenser l'évaluation car, dit-elle, au cours des 20 ou 30 dernières années, les problématiques, les besoins, et les méthodes se sont diversifiés et souvent complexifiés.

Enfin, citons le score utilisé par [Voohrees 2003] $\frac{1}{Q}\sum_{i=1}^{Q}\frac{n}{i}$, ou n représente le nombre de réponses exactes au rang i, et Q le nombre de questions. Ce score privilégie donc le fait de proposer en premier les bonnes réponses. Il est désormais utilisé pour analyser les résultats de la conférence TREC.

4.2.2 Le test de Kappa

Le test non paramétrique Kappa de Cohen permet de chiffrer l'accord entre deux ou plusieurs observateurs lorsque les jugements sont qualitatifs, contrairement au coefficient *u* de Kendall par exemple, qui évalue le degré d'accord entre des jugements quantitatifs. L'accord observé résulte de la somme d'une composante «aléatoire» et d'une composante d'accord «véritable». [Bonnardel 1995] présente une étude approfondie du test de Kappa et de ses variantes.

Le coefficient Kappa *K* propose de chiffrer l'intensité ou la qualité de l'accord réel entre des jugements qualitatifs appariés. Il exprime une différence relative entre la proportion d'accord observée P_o et la proportion d'accord aléatoire P_e qui est la valeur espérée sous l'hypothèse nulle d'indépendance des jugements, divisée par la quantité disponible au-delà de l'accord aléatoire [Grenier 1993].

En définitive, *K* est un pourcentage de l'accord maximum corrigé de ce qu'il serait sous le simple effet du hasard.

Dans le cas d'une étude d'accord entre deux observateurs statistiquement indépendants ayant *r* modalités de jugement, avec $r \geq 2$, le coefficient Kappa s'écrit :

$$K = \frac{P_o - P_e}{1 - P_e}$$

Équation 7. Coefficient K de Kappa

avec

P_o : la proportion d'accord observée.

P_e : la proportion d'accord aléatoire ou concordance attendue sous l'hypothèse d'indépendance des jugements.

Pour calculer cette valeur, on établit un tableau de contingence comme suit

		Juge A				
	Catégorie	1	2	...	*r*	Total
Juge B	1	n_{11}	n_{12}	...	n_{1r}	$n_{1.}$
	2	n_{21}	n_{22}	...	n_{2r}	$n_{2.}$
	.					
	.					
	R	n_{r1}	n_{r2}	...	n_{rr}	$n_{r.}$
	Total	$n_{.1}$	$n_{.2}$	...	$n_{.r}$	n

Figure 53. Tableau de contingence pour Kappa

On appelle concordance observée P_o, la proportion des individus classés dans les cases diagonales de concordance du tableau de contingence, soit la somme de ces effectifs diagonaux divisée par la taille de l'échantillon (*n*).

$$P_o = \sum_{i=1}^{r} p_{ii} = \frac{1}{n} \sum_{i=1}^{r} n_{ii}$$

Équation 8. Concordance observée de Kappa

avec *r* le nombre de modalités de jugement.

Et la concordance aléatoire P_e qui est égale à la somme des produits des effectifs marginaux divisée par le carré de la taille de l'échantillon.

$$P_e = \sum_{i=1}^{r} p_{i.} p_{.i} = \frac{1}{n^2} \sum_{i=1}^{r} n_{i.} n_{.i}$$

Équation 9. Concordance aléatoire de Kappa

Le coefficient Kappa est un nombre réel, sans dimension, compris entre -1 et 1. L'accord sera d'autant plus élevé que la valeur de Kappa est proche de 1 et l'accord maximal est atteint (K=1) lorsque $P_o = 1$ et $P_e = 0{,}5$.

Lorsqu'il y a indépendance des jugements, le coefficient Kappa est égal à zéro ($P_o = P_e$), et dans le cas d'un désaccord total entre les juges, le coefficient Kappa prend la valeur -1 avec $P_o = 0$ et $P_e = 0{,}5$.

[Landis et Koch 1977] ont proposé un classement de l'accord en fonction de la valeur de Kappa présenté dans le tableau suivant :

Accord	*Kappa*
Excellent	≥ 0,81
Bon	0,80 - 0,61
Modéré	0,60 - 0,41
Médiocre	0,40 - 0,21
Mauvais	0,20 - 0,0
Très mauvais	< 0,0

Figure 54. Degré d'accord et valeur de Kappa

Les limites de ce classement sont arbitraires et peuvent varier selon l'étude réalisée.

4.2.3 Distance sémantique

[Rodriguez 2003] aborde la question de la distance sémantique de manière globale : le voisinage sémantique N d'une entité de classe est un jeu de classes se trouvant à une distance inférieure ou égale à un entier positif r, dénommé le radius du voisinage sémantique. [Roddick 2003] propose une formule générique de calcul de distance sémantique, mais qui nécessite toutefois de connaître la distance entre les nœuds. [JIN 2003] présente plusieurs mesures, en particulier une distance basée sur la corrélation entre classes. [Laurent 2005] utilise la distance de Levenshtein dans le projet industriel Lucene. [Budanitsky 2001] compare cinq distances sémantiques : la mesure de Hirst and St-Onge's [Hirst 1998] basée sur le nombre de changements de directions, la mesure de Leacock and Chodorow [Leacock 1998] basée sur le plus court chemin, la mesure de Resnik [Resnik 1995] fondée sur le niveau d'information commune, la mesure de Jiang et Conrath [Jiang 1997] calculée sur la probabilité d'apparition d'une classe, et enfin [Lin 1998] qui en est une autre déclinaison. Les meilleurs résultats sont obtenus avec la formule de Jiang and Conrath. [Jian-Yun Nie 2004] met l'accent sur le choix de la mesure de distance et cite le cas d'une désambiguïsation détériorée par l'usage d'une distance calculée selon le nombre de liens entre les synsets dans l'arbre de hypernyme-hyponyme. Il met en cause l'utilisation du nombre de liens comme mesure de distance sémantique. [Amr Helmi 2002] aborde la difficulté de l'interprétation et pense que quelle que soit la méthode utilisée pour mesurer la distance sémantique entre deux termes, le résultat n'est pertinent que si l'outil descriptif est capable d'associer son interprétation du différentiel sémantique à la configuration morpho-syntaxique dans laquelle apparaissent ces termes. Enfin, [Bousquet 2001] utilise une distance sémantique dans le domaine des classifications médicales, en projetant les diagnostics sur les axes de la SNOMED.

4.3 Paradigme de l'évaluation : vers une nouvelle approche ?

La F-mesure est l'indicateur de synthèse communément utilisé pour évaluer les algorithmes de classification, à partir de la précision et du rappel. Actuellement cet indicateur est systématiquement utilisé pour évaluer les algorithmes. Nous proposons une nouvelle approche de l'évaluation en présentant une mesure de convergence entre différentes orientations. Ces nouvelles mesures, la K-mesure et la D-mesure, permettent également de prendre en compte des niveaux d'exigence des juges.

Le traitement du langage naturel donne naissance à de nombreux algorithmes, dont la recherche d'informations, la classification, la catégorisation, … L'efficacité de la plupart de ces algorithmes est calculée à partir de diverses mesures, dont les plus connues sont la précision et le rappel. Pour faciliter la lecture de la performance, [Van Rijsbergen 79] a créé une mesure synthétique : la F-mesure, qui combine ces deux indicateurs. Aujourd'hui, les besoins se sont diversifiés, les problématiques se sont complexifiées, mais l'on conserve les mêmes indicateurs de référence depuis 25 ans. Certes, cela facilite la comparaison entre différentes méthodes mais leur emploi est-il toujours pertinent ?

4.4 De la difficulté d'évaluer

L'intérêt des indicateurs de performance est de permettre de comparer la qualité des algorithmes. Malheureusement, les auteurs n'utilisent pas exactement les mêmes découpages. Pour le corpus Reuters-21578, certains auteurs considèrent 90 catégories [Joachims 1998], [Schapire et al. 1998], [Yang et Liu 1999], d'autres en considèrent 118 [Dumais et al. 1998]. De plus, la plupart des auteurs considèrent 3299 documents sur la base de test, mais [Yang et Liu 1999] en considèrent uniquement 3019 en supprimant tous les documents de la base de test qui n'appartiennent à aucune catégorie.

Ces différences de découpages rendent difficiles les comparaisons à travers ces publications. De plus, tous les auteurs n'utilisent pas les mêmes mesures de performances, et peuvent calculer les moyennes de manières différentes. Enfin, même dans le cas où les auteurs utilisent les mêmes mesures, il est nécessaire d'utiliser des tests statistiques pour

vérifier que les différences ne sont pas dues au hasard [Hull 1993].

Une autre approche souvent proposée est l'utilisation de plusieurs méthodes par le même auteur, sur la même base ([Dumais et al. 1998], [Yang et Liu 1999]). Ainsi, le découpage et les mesures sont identiques pour toutes les méthodes. Mais les algorithmes sont souvent délicats à mettre en œuvre et certains sont peut être mieux implémentés que d'autres. Par exemple, l'implémentation des machines à vecteurs supports proposée par [Dumais 1998] obtient des résultats nettement meilleurs que celle proposée par [Joachims 1998]. Les réseaux de neurones testés par [Yang et Liu 1999] sont des perceptrons multi couches avec une couche cachée comportant 64 neurones, 1000 descripteurs en entrée et 90 neurones de sorties correspondant aux 90 catégories ; ils considèrent un seul réseau pour l'ensemble des catégories comportant plus de 64000 poids (l'algorithme d'apprentissage n'est pas précisé). Il n'est pas surprenant, dans ces conditions, que les performances obtenues ne soient pas très bonnes : de telles démarches jugent plus la capacité des auteurs à mettre en œuvre des méthodes, que les capacités des méthodes elles-mêmes. L'algorithme de Rocchio est ancien, mais [Schapire et al. 1998] ont montré qu'il obtient d'excellents résultats pour la catégorisation de textes à condition d'utiliser un codage efficace, de bien choisir les documents non pertinents, et d'effectuer une optimisation des poids. Leurs conclusions vont à l'encontre d'autres comparaisons qui montrent que cet algorithme n'est pas performant par rapport aux méthodes fondées sur l'apprentissage numérique [Schütze et al. 1995] [Lewis et al. 1996] [Cohen et Singer 1996]. Ces différentes remarques prouvent que le succès d'une méthode dépend d'un ensemble de paramètres qui vont de l'étiquetage des documents au choix des algorithmes et de leur utilisation, et qu'il est, par conséquent, extrêmement difficile de tirer des conclusions définitives sur une approche.

[Voorhees 2003] (ainsi que [Voorhees 2000], [Voorhees 2001], et [Voorhees 2002]) s'est penchée sur l'analyse de l'évaluation réalisée dans le cadre des conférences TREC. Elle souligne en particulier la mauvaise reproductibilité de l'évaluation entre experts humains : à la question « Who is Tom Cruise married to ? », certains juges acceptaient « Kidman » comme bonne réponse, mais d'autres la considéraient comme inexacte. Quelques juges ont validé « actress Nicole Kidman », d'autres non. [Voorhees 2003] demande : quand une bonne réponse est « 5 octobre 1996 », pouvons nous considérer comme faux un système qui donnerait « octobre 1996 » ?

[Dimitrova 2004] a analysé les classifieurs de documents web. Elle a préféré l'évaluation humaine, ce qui lui a permis de mettre en évidence l'émergence de la dimension cognitive et pose le problème de la difficulté de la tâche : la longueur du texte, la présentation, le niveau d'expertise du lecteur, … sont autant de critères qui vont impacter le jugement. Elle illustre parfaitement l'importance du juge et la complexité de cette approche.

Les documents étant, selon l'inventeur du web, "destinés aux humains plutôt que des données et informations qui peuvent être analysées automatiquement" [Berners-Lee et al. 2001], le défi est justement d'extraire automatiquement de l'information de ces documents écrits en langage naturel. Tim Berners-Lee y répond en introduisant le semantic web, pour faciliter l'exploitation du web par les outils.

4.5 Vers de nouveaux indicateurs

4.5.1 Analyse de la F-Mesure

Nous allons démontrer dans un premier temps que la F-mesure correspond en fait à une moyenne harmonique, éventuellement pondérée par le coefficient β, puis en analyser les propriétés.

Van Rijsbergen [Van Rijsbergen 79] définit la E et F-mesure comme suit :
Mesure F = ((1+ß²)*Précision*Rappel) / ((ß²*Précision)+Rappel)

Le paramètre ß permet de donner un poids plus important à la précision ou au rappel, mais on positionne généralement le paramètre ß à 1. La mesure devient :
F1-Mesure = (2*Précision*Rappel) / (Précision+Rappel)

L'avantage de ce choix est que lorsque la précision est égale au rappel, on obtient : Précision = Rappel = F-mesure. Ceci facilite la lecture et en général, on recherche à maximiser la F-mesure (en maximisant la précision et le rappel), et équilibrer précision et rappel (ce qui nous ramène au 'Break Even Point'). En effet, il deviendrait difficile d'apprécier un algorithme qui fournirait une bonne précision et un mauvais rappel ou inversement.

Considérons maintenant la moyenne harmonique M de la précision (P) et du rappel (R) :

$\frac{2}{M} = \frac{1}{P} + \frac{1}{R}$ donc $\frac{2}{M} = \frac{P+R}{P*R}$ ce qui donne : $\frac{M}{2} = \frac{P*R}{P+R}$ et finalement : $M = \frac{2*(P*R)}{P+R} = F1$

Nous constatons que la F1-mesure n'est que la moyenne harmonique de la précision et du rappel. L'usage de la moyenne harmonique ne se justifie pas par la nature du problème posé mais par ses propriétés. En particulier, la moyenne harmonique possède la qualité de baisser fortement lorsque l'un de ses paramètres est bas et d'être maximisée lorsque les deux paramètres sont proches entre eux tout en étant élevés. Cette propriété pénalise donc les algorithmes qui dégraderaient la précision ou le rappel, ou l'un exclusivement au détriment de l'autre. Prenons la F1-mesure et démontrons cette propriété : nous avons F1=2*P*R/(P+R), avec précision=P et rappel=R. Posons S=P+R et D=P-R. Notre problème devient donc : comment impacter la F-mesure pour qu'elle soit optimale quand on maximise S (donc la

précision ET le rappel) en minimisant D (en les conservant proches). Nous avons : $S^2 - D^2 = (S+D)(S-D) = (P+R+P-R)\times((P+R)-(P-R)) = 2P \times 2R = 4PR$

En conclusion, on a : $F1 = \frac{2*P*R}{P+R} = \frac{S^2 - D^2}{2S} = \frac{S}{2} - \frac{D^2}{2S}$; ceci explique pourquoi la F-mesure croît quand la somme (S) augmente, et décroît quand la différence (D) augmente. Nous pouvons observer cette propriété sur les graphiques suivants, où l'on compare la moyenne harmonique avec la moyenne géométrique et arithmétique :

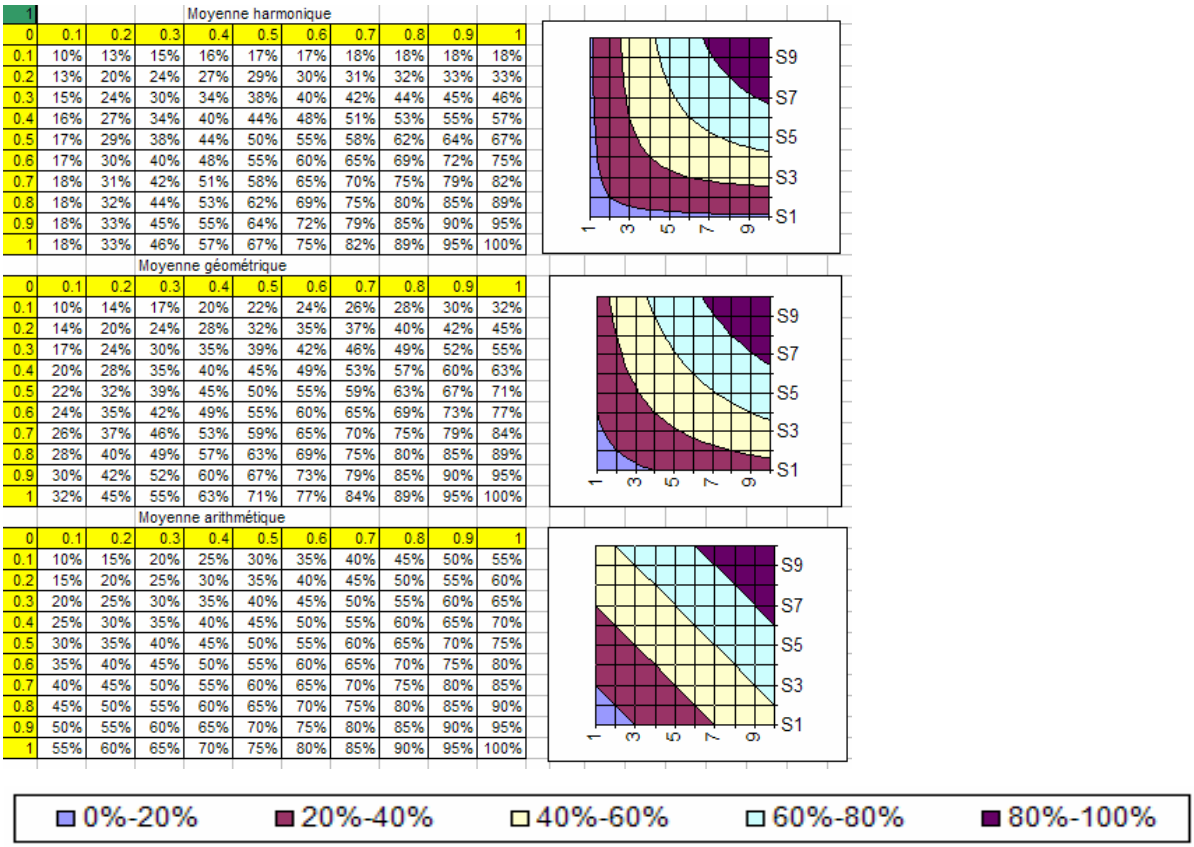


Figure 55. Comparaison des différentes moyennes

Nous observons que la surface basse (mesure d'évaluation comprise entre 0 et 0.20, donc mauvais score) est plus étirée vers les extrêmes, ce qui traduit bien le fait que si l'un des composants est faible, la mesure d'évaluation est faible.

La F_n-mesure possède une autre propriété : elle permet d'accorder plus de poids à la précision ou au rappel. Pour cela, on modifie le dénominateur original (P+R) en ((ß²*Précision)+Rappel). Le résultat de ce calcul est compris entre 0 et ß²+1. Pour que le résultat final reste dans l'intervalle [0,1], il faut donc multiplier le numérateur par (ß²+1). Cela nous donne la formule finale de la F-mesure. Nous pouvons observer la pertinence de cette propriété, en donnant 3 valeurs successives au paramètre ß² : 0.3, 1 , et 3.

			F-Mesure						Beta=	0.3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	10%	10%	11%	11%	11%	11%	11%	11%	11%	11%
0.2	18%	20%	21%	21%	21%	21%	21%	21%	21%	21%
0.3	26%	29%	30%	31%	31%	31%	31%	32%	32%	32%
0.4	32%	37%	39%	40%	41%	41%	41%	42%	42%	42%
0.5	38%	44%	47%	49%	50%	51%	51%	52%	52%	52%
0.6	42%	51%	55%	58%	59%	60%	61%	61%	62%	62%
0.7	47%	58%	63%	66%	68%	69%	70%	71%	71%	72%
0.8	51%	64%	70%	74%	76%	78%	79%	80%	81%	81%
0.9	54%	70%	77%	82%	84%	86%	88%	89%	90%	91%
1	57%	75%	84%	89%	92%	95%	97%	98%	99%	100%

			F-Mesure						Beta=	1
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	10%	13%	15%	16%	17%	17%	18%	18%	18%	18%
0.2	13%	20%	24%	27%	29%	30%	31%	32%	33%	33%
0.3	15%	24%	30%	34%	38%	40%	42%	44%	45%	46%
0.4	16%	27%	34%	40%	44%	48%	51%	53%	55%	57%
0.5	17%	29%	38%	44%	50%	55%	58%	62%	64%	67%
0.6	17%	30%	40%	48%	55%	60%	65%	69%	72%	75%
0.7	18%	31%	42%	51%	58%	65%	70%	75%	79%	82%
0.8	18%	32%	44%	53%	62%	69%	75%	80%	85%	89%
0.9	18%	33%	45%	55%	64%	72%	79%	85%	90%	95%
1	18%	33%	46%	57%	67%	75%	82%	89%	95%	100%

			F-Mesure						Beta=	3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	10%	18%	25%	31%	36%	40%	44%	47%	50%	53%
0.2	11%	20%	29%	36%	43%	50%	56%	62%	67%	71%
0.3	11%	21%	30%	39%	47%	55%	62%	69%	75%	81%
0.4	11%	21%	31%	40%	49%	57%	65%	73%	80%	87%
0.5	11%	21%	31%	41%	50%	59%	67%	75%	83%	91%
0.6	11%	21%	32%	41%	51%	60%	69%	77%	86%	94%
0.7	11%	22%	32%	42%	51%	61%	70%	79%	88%	96%
0.8	11%	22%	32%	42%	52%	62%	71%	80%	89%	98%
0.9	11%	22%	32%	42%	52%	62%	72%	81%	90%	99%
1	11%	22%	32%	43%	53%	63%	72%	82%	91%	100%

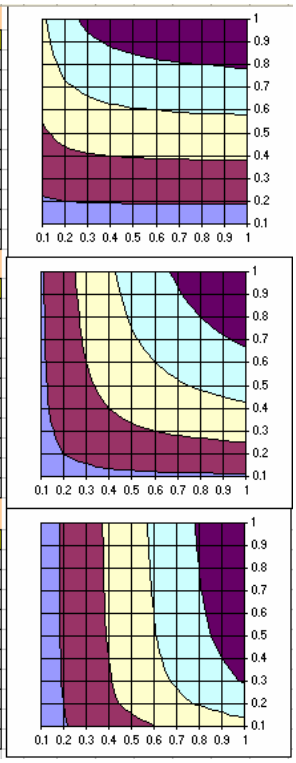


Figure 56. Le paramètre Beta de la F-mesure

4.5.2 Propositions de nouveaux indicateurs : vers la K-mesure

Nous avons tenté de définir la classification et de la dissocier de la catégorisation. Nous allons maintenant étudier quels pourraient être les indicateurs les plus riches en fonction de la nature des tâches évaluées.

4.5.2.1 Cas des catégorisations

Dans le cas d'une catégorisation, il n'y a aucun moyen d'évaluer une distance entre une proposition et la solution désirée. Nous n'avons d'autre possibilité que celle d'utiliser un système binaire, soit 0 ou 1 selon que le résultat est faux ou vrai. La F-mesure, bien que mathématiquement inférieure à la moyenne arithmétique, possède des propriétés très intéressantes exposées dans la section précédente. Tefko Saracevic [SAR 70], dans sa définition de l'évaluation, insistait sur l'importance du juge. Il est rejoint par [Voorhees 2003] et [Dimitrova 2004]. [Van Rijsbergen 79] y répond très partiellement en introduisant le paramètre β qui permet de privilégier la précision ou le rappel. Mais aucune variable ne permet à l'utilisateur de mettre un niveau d'exigence en terme de précision ou de rappel. Sur

le graphique suivant (chaque courbe représente un rappel variant de 0.1 à 1 et affiche la F-mesure résultante), nous constatons que le faible niveau de précision (ou rappel) abaisse la courbe, mais qu'au-delà de 0.5, le système devient pratiquement linéaire.

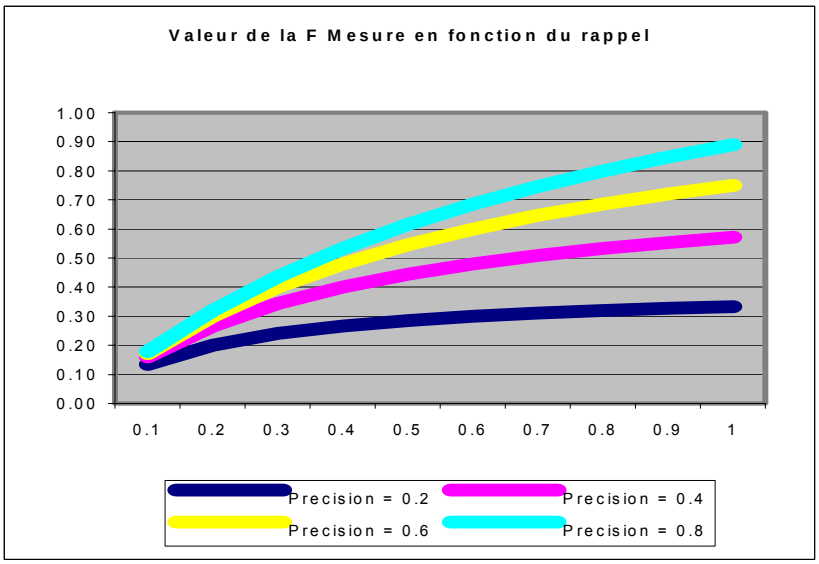


Figure 57. Evolution de la F-mesure en fonction du rappel

Or un utilisateur (le juge de Saracevic), peut très bien considérer une précision ou un rappel de 0.5 ou 0.6 comme méritant une mauvaise évaluation. La F-mesure ne permet pas de traduire ce qu'exprime le « juge ».

Le problème devient alors de trouver une fonction φ capable de minorer un résultat compris entre 0 et 1 et retournant un résultat compris dans ce même intervalle en fonction d'un niveau d'exigence du juge. La fonction puissance répond parfaitement à ce besoin. Nous avons donc intégré cette fonction φ dans la F-mesure, et introduisons la **K-Mesure** :

$$\textbf{K-Mesure} = (1+ß^2)*(\text{Précision}*\text{Rappel})^{\alpha} / ((ß^2*\text{Précision})+\text{Rappel})$$

Etudions maintenant les propriétés de la K-Mesure. En premier lieu, nous constatons que si $\alpha=1$, nous obtenons la F-mesure, dans sa version la plus complète. En posant $\alpha=1$ et $ß^2=1$, nous obtenons la F-mesure dans son acception la plus usuelle. La K-mesure est donc un sur-ensemble de la F-mesure. Ainsi, tout l'historique peut être conservé. Observons maintenant les propriétés de la K-mesure, en faisant varier le paramètre α, avec $ß^2=1$.

α =	1								Beta=	1
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	10%	13%	15%	16%	17%	17%	18%	18%	18%	18%
0.2	13%	20%	24%	27%	29%	30%	31%	32%	33%	33%
0.3	15%	24%	30%	34%	38%	40%	42%	44%	45%	46%
0.4	16%	27%	34%	40%	44%	48%	51%	53%	55%	57%
0.5	17%	29%	38%	44%	50%	55%	58%	62%	64%	67%
0.6	17%	30%	40%	48%	55%	60%	65%	69%	72%	75%
0.7	18%	31%	42%	51%	58%	65%	70%	75%	79%	82%
0.8	18%	32%	44%	53%	62%	69%	75%	80%	85%	89%
0.9	18%	33%	45%	55%	64%	72%	79%	85%	90%	95%
1	18%	33%	46%	57%	67%	75%	82%	89%	95%	100%

α =	1.2								Beta=	1
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	4%	6%	7%	8%	9%	10%	10%	11%	11%	11%
0.2	6%	11%	14%	16%	18%	20%	21%	22%	23%	24%
0.3	7%	14%	19%	22%	26%	28%	31%	33%	35%	36%
0.4	8%	16%	22%	28%	32%	36%	39%	42%	45%	48%
0.5	9%	18%	26%	32%	38%	43%	47%	51%	55%	58%
0.6	10%	20%	28%	36%	43%	49%	54%	59%	64%	68%
0.7	10%	21%	31%	39%	47%	54%	61%	66%	72%	77%
0.8	11%	22%	33%	42%	51%	59%	66%	73%	79%	85%
0.9	11%	23%	35%	45%	55%	64%	72%	79%	86%	93%
1	11%	24%	36%	48%	58%	68%	77%	85%	93%	100%

α =	1.6								Beta=	1
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	1%	1%	2%	2%	3%	3%	4%	4%	4%	5%
0.2	1%	3%	4%	6%	7%	8%	10%	11%	12%	13%
0.3	2%	4%	7%	10%	12%	14%	16%	19%	21%	22%
0.4	2%	6%	10%	13%	17%	20%	24%	27%	30%	33%
0.5	3%	7%	12%	17%	22%	26%	31%	36%	40%	44%
0.6	3%	8%	14%	20%	26%	33%	38%	44%	50%	55%
0.7	4%	10%	16%	24%	31%	38%	46%	53%	60%	66%
0.8	4%	11%	19%	27%	36%	44%	53%	61%	70%	78%
0.9	4%	12%	21%	30%	40%	50%	60%	70%	79%	89%
1	5%	13%	22%	33%	44%	55%	66%	78%	89%	100%

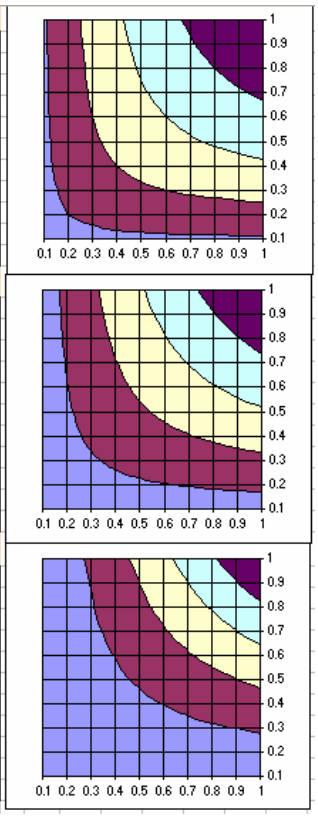


Figure 58. Variation de la K-mesure du paramètre Alpha

Nous constatons que plus le paramètre α augmente, plus le niveau d'exigence du juge grandit. Par exemple, avec P=R=0.4, la F-mesure vaut 0.4, alors que la K-mesure vaut 0.13 pour α =1.6

Ce résultat sera donc considéré comme mauvais alors que la F-mesure le considère comme moyen. Nous pouvons donc ainsi formaliser un niveau d'exigence du juge.

Observons à présent si la possibilité d'avantager la précision au détriment du rappel (ou inversement) est conservée, soit en faisant varier le paramètre β :

a=	1								Beta=	3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	10%	18%	25%	31%	36%	40%	44%	47%	50%	53%
0.2	11%	20%	29%	36%	43%	50%	56%	62%	67%	71%
0.3	11%	21%	30%	39%	47%	55%	62%	69%	75%	81%
0.4	11%	21%	31%	40%	49%	57%	65%	73%	80%	87%
0.5	11%	21%	31%	41%	50%	59%	67%	75%	83%	91%
0.6	11%	21%	32%	41%	51%	60%	69%	77%	86%	94%
0.7	11%	22%	32%	42%	51%	61%	70%	79%	88%	96%
0.8	11%	22%	32%	42%	52%	62%	71%	80%	89%	98%
0.9	11%	22%	32%	42%	52%	62%	72%	81%	90%	99%
1	11%	22%	32%	43%	53%	63%	72%	82%	91%	100%

a=	1.2								Beta=	3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	4%	8%	12%	16%	20%	23%	26%	28%	31%	33%
0.2	5%	11%	16%	22%	27%	33%	38%	43%	47%	52%
0.3	5%	12%	19%	25%	32%	39%	45%	52%	58%	64%
0.4	6%	13%	20%	28%	35%	43%	50%	58%	65%	72%
0.5	6%	13%	21%	30%	38%	46%	55%	63%	71%	79%
0.6	6%	14%	22%	31%	40%	49%	58%	67%	76%	85%
0.7	6%	15%	23%	32%	42%	51%	61%	70%	80%	89%
0.8	7%	15%	24%	34%	43%	53%	63%	73%	83%	93%
0.9	7%	15%	25%	35%	45%	55%	65%	76%	86%	97%
1	7%	16%	25%	35%	46%	56%	67%	78%	89%	100%

a=	1.6								Beta=	3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	1%	2%	3%	4%	6%	7%	9%	10%	12%	13%
0.2	1%	3%	5%	8%	11%	14%	17%	20%	24%	27%
0.3	1%	4%	7%	11%	15%	19%	24%	29%	34%	39%
0.4	2%	5%	9%	13%	19%	24%	30%	37%	43%	50%
0.5	2%	5%	10%	16%	22%	29%	36%	44%	52%	60%
0.6	2%	6%	11%	18%	25%	33%	41%	50%	59%	69%
0.7	2%	7%	12%	19%	27%	36%	46%	56%	66%	77%
0.8	2%	7%	14%	21%	30%	40%	50%	61%	73%	85%
0.9	3%	8%	15%	23%	32%	43%	54%	66%	79%	93%
1	3%	8%	16%	25%	35%	46%	58%	71%	85%	100%

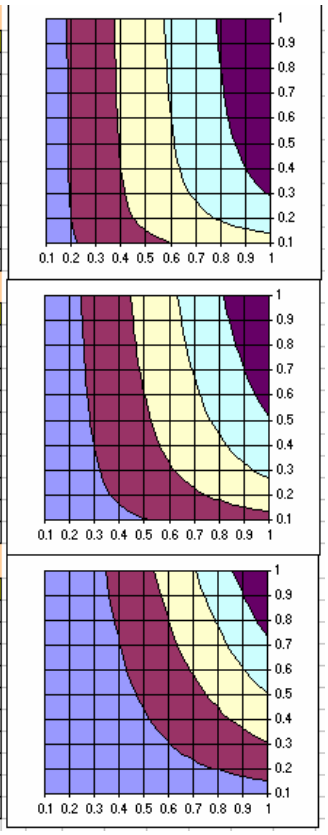


Figure 59. Variation du paramètre Beta de la K-mesure

Nous constatons qu'effectivement, cette possibilité est conservée. Nous obtenons la même propriété en faisant varier l'autre axe (soit des valeurs du paramètre β inférieures à 1). Exemple :

α =	1.4								Beta=	0.3
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	2%	2%	3%	3%	3%	3%	4%	4%	4%	4%
0.2	4%	6%	7%	8%	8%	9%	10%	10%	11%	11%
0.3	6%	9%	11%	13%	15%	16%	17%	18%	19%	20%
0.4	9%	13%	17%	19%	21%	23%	25%	26%	28%	29%
0.5	11%	18%	22%	26%	29%	31%	34%	36%	38%	40%
0.6	14%	22%	28%	33%	36%	40%	43%	46%	48%	51%
0.7	16%	26%	34%	40%	45%	49%	53%	56%	59%	62%
0.8	18%	31%	40%	47%	53%	58%	63%	67%	71%	74%
0.9	21%	35%	46%	54%	61%	68%	73%	78%	83%	87%
1	23%	39%	52%	62%	70%	77%	84%	90%	95%	100%

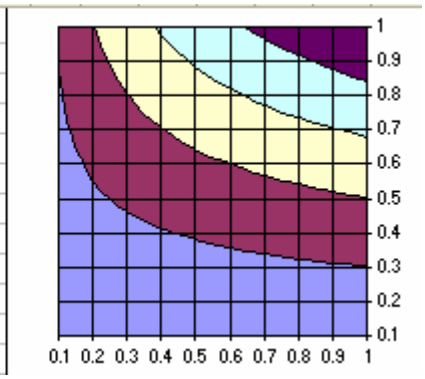


Figure 60. Variation des deux paramètres de la K-mesure

Enfin, nous pouvons donner au paramètre α une valeur comprise en 0.5 et 1, à condition que l'on ait β =1 (sinon, le résultat peut devenir supérieur à 1). La valeur ne peut en aucun cas être inférieure à 0.5 car les propriétés désirées ne seraient plus respectées. Une valeur remarquable est pourtant intéressante : α = 0.5 :

α =	0.5								Beta=	1
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	100%	94%	87%	80%	75%	70%	66%	63%	60%	57%
0.2	94%	100%	98%	94%	90%	87%	83%	80%	77%	75%
0.3	87%	98%	100%	99%	97%	94%	92%	89%	87%	84%
0.4	80%	94%	99%	100%	99%	98%	96%	94%	92%	90%
0.5	75%	90%	97%	99%	100%	100%	99%	97%	96%	94%
0.6	70%	87%	94%	98%	100%	100%	100%	99%	98%	97%
0.7	66%	83%	92%	96%	99%	100%	100%	100%	99%	98%
0.8	63%	80%	89%	94%	97%	99%	100%	100%	100%	99%
0.9	60%	77%	87%	92%	96%	98%	99%	100%	100%	100%
1	57%	75%	84%	90%	94%	97%	98%	99%	100%	100%

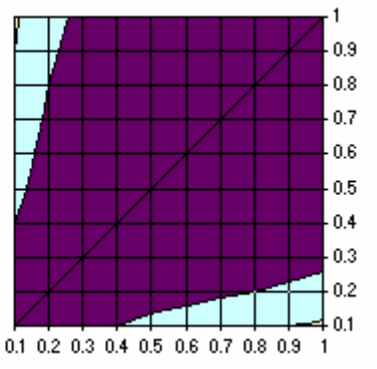


Figure 61. K-mesure avec alpha = 0.5

Nous observons que cette valeur permet de privilégier fortement l'indicateur lorsque l'on a précision P = rappel R, ce qui correspond à l'approche du 'Break Even Point'.

En conclusion, la K-mesure possède des propriétés qui la rendent intéressante :

- elle est un sur-ensemble de la F-mesure et en conserve les qualités,
- elle permet de représenter un niveau d'exigence du juge,
- elle permet de modéliser le Break Even Point.

Il s'agit donc d'une formule de convergence et d'un sur-ensemble des différentes approches utilisées à ce jour.

4.5.3 Cas des classifications

Une classification se distingue d'une catégorisation par le fait qu'il est possible de mettre en œuvre une mesure de distance entre les classes. Dans le cas d'une classification hiérarchique, la mesure de la précision et du rappel présente un paradoxe : plus la hiérarchie est fine, plus la mesure de l'évaluation sera mauvaise. Prenons le cas de la CIM 10 (classification internationale des maladies) : au niveau 4, nous disposons de 15000 codes

possibles, ce qui permet d'obtenir un niveau d'individualisation déjà très fin. Mais si l'on fait une estimation sur le niveau 5, il faut alors trouver le bon code parmi 51 000 possibilités. La probabilité de trouver le bon code diagnostic devient alors plus faible et la précision et le rappel seront également plus faibles.

En généralisant ce principe, on peut en déduire que la précision et le rappel convergent vers 0 lorsque le niveau de hiérarchie croît. Mais on peut alors considérer qu'une réponse donnée à un niveau très bas de hiérarchie avec juste une erreur sur le dernier niveau de hiérarchie est pratiquement correcte. Dans le cas d'une F-mesure, elle serait considérée comme fausse.

[Budanitsky 2001] a démontré que les meilleurs résultats étaient obtenus avec la mesure de Jiang et Conrath :

Soit $d=Dist_{jc}(c1 :c2) = 2 \log(p(lso(c1 : c2)))-(\log(p(c1))+\log(p(c2)))$

Avec lso(c1 : c2)= plus grand sous-groupe commun.

Si l'on nomme 'd' cette distance (avec d=1 si les classes sont distantes), la précision et le rappel peuvent alors se définir comme suit :

a = Nombre de classes pertinentes et proposées (donc correctement classifiées)

d(b)= proposé mais non pertinent, on prend la mesure de distance avec la solution la plus proche

d(c)= non proposé mais pertinent, on prend la mesure de distance avec la classe proposée la plus proche

Précision = a / (1- somme(d (b)))

Rappel = a / (1- somme(d (c)))

Il devient alors possible d'utiliser la K-mesure.

Cette proposition d'utiliser la hiérarchie pour affiner la qualité de l'évaluation augmentera mécaniquement les valeurs de précision et de rappel mais sera plus proche de la réalité. Or c'est bien l'objectif recherché par la mise en œuvre d'un indicateur d'évaluation.

4.5.4 Cas des recherches d'informations (IR)

La recherche d'informations se différencie des classifications et catégorisations par le très grand nombre de réponses possibles. L'application classique serait celle d'un moteur de recherche.

4.5.4.1 Approche sans juge

Dans le cas d'une recherche d'informations, la précision et le rappel ne sont pas indépendants. En effet, pour une requête donnée, il est facile d'obtenir un rappel à 1 : il suffit de proposer toutes les réponses possibles. Par contre, ce cas est peu intéressant car insuffisamment déterminant et nous obtiendrions une précision proche de zéro. Inversement, si un algorithme est correctement construit, la probabilité pour que la première réponse proposée soit correcte est plus importante que pour les suivantes. Nous obtiendrions alors une précision forte (tendant vers 1), alors que le rappel serait faible (tendant vers 0). En faisant varier le nombre de réponses, nous nous déplacerons d'un point extrême vers l'autre, c'est-à-dire du point (0,1) vers le point (1,0).

Si l'on souhaite maximiser le couple (précision, rappel) sans privilégier aucune de ces deux mesures, il convient alors de trouver le couple qui croise la seconde diagonale (y=-x+1). Mais cette hypothèse présente le défaut de laisser la possibilité de privilégier un couple de points marginal qui par exemple ne fonctionnerait très bien que pour une classe donnée (cas d'un pic graphique). Imaginons un moteur de recherche qui ne soit pertinent que pour une catégorie ou une classe donnée !

Pour résoudre ce problème, il faut donc avoir une approche plus globale. Si l'on représente la courbe d'évolution et que l'on souhaite maximiser l'ensemble des couples (précision, rappel), soit maximiser la surface comprise entre l'axe horizontal (y=0), l'axe vertical (x=0), et la courbe (ou à minimiser la surface au dessus de la courbe. Ceci revient à utiliser la formule de Salton, qui simplifie ce calcul en extrapolant 11 points de mesure. L'indicateur pertinent serait alors le calcul de cette surface ramené à la surface totale du carré. Si l'on souhaite privilégier la précision ou le rappel, il faudra alors représenter la courbe dans

un rectangle (m,n), m et n étant les poids affectés aux deux paramètres. Etant donné que la surface calculée est ramenée à la surface d'origine, l'indicateur conserve la même échelle (valeur comprise entre 0 et 1), et il est facile de comparer des algorithmes entre eux en fonction de la finalité.

Mais cette approche n'est valable que pour les approches booléennes : un document est bien retrouvé ou non. Ceci suppose l'existence d'un corpus de référence. La réalité est souvent plus complexe que cela, et les jugements plus mitigés : un document proposé est généralement partiellement pertinent. Or les formules existantes à ce jour ne permettent pas de formaliser cette approche. Une approche naturelle consisterait alors à mettre un poids sur la qualité de la réponse, variant de 0 à 1. Ce poids représenterait le pourcentage d'accord entre le résultat attendu et le résultat proposé. Mais le problème se complexifie lorsque l'on souhaite mettre un niveau d'exigence. Comme nous l'avons exposé, l'avis du juge est souvent un critère déterminant. Le niveau de précision ou rappel peut être important dans un domaine (en médecine par exemple), alors qu'il le serait moins dans une approche grand public.
Il convient donc de trouver une formulation adéquate.

4.5.4.2 Approche avec juge

Pour établir un indicateur, nous sommes partis de la formule de score donnée par [Voohrees 2003] $\frac{1}{Q}\sum_{i=1}^{Q}\frac{n}{i}$, où n représente le nombre de réponses exactes au rang i, et Q le nombre de questions. Ce qui donne le poids suivant de chaque réponse pour par exemple 20 réponses :

Rang	Contribution
1	17.99%
2	12.99%
3	10.49%
4	8.82%
5	7.57%
6	6.57%
7	5.74%
8	5.02%
9	4.40%
10	3.84%

Rang	Contribution
11	3.34%
12	2.89%
13	2.47%
14	2.09%
15	1.73%
16	1.40%
17	1.09%
18	0.79%
19	0.51%
20	0.25%

Figure 62. Poids de chaque rang pour l'évaluation

Soit le graphe suivant :

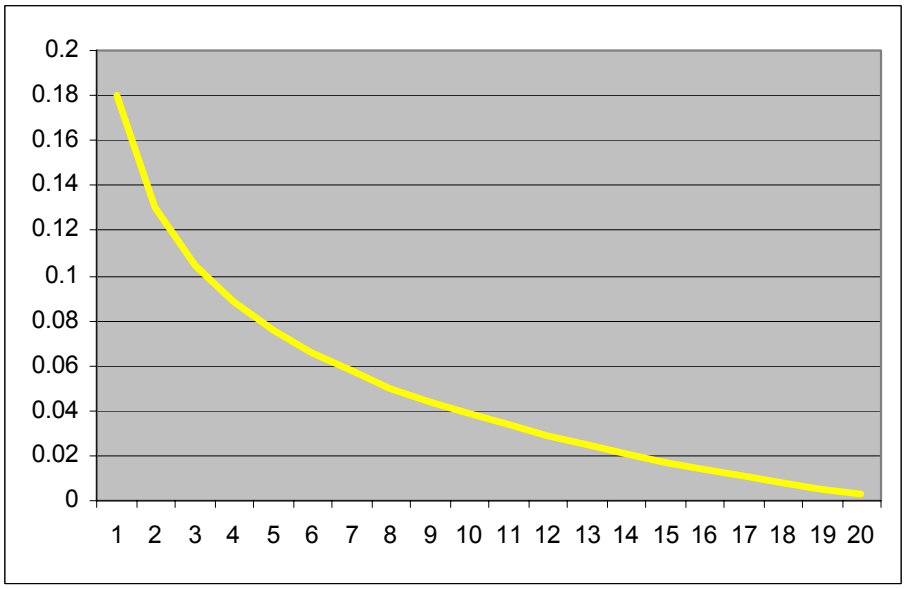


Figure 63. Poids de chaque réponse selon [Voohrees 2003]

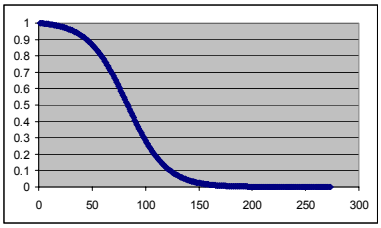
Pour pouvoir représenter un niveau d'exigence du juge (exemple : "je veux absolument que les réponses les plus pertinentes soient sur la première page de réponse, soit 30 réponses), il faut moduler cette approche booléenne et pratiquement linéaire. Une formule mathématique qui permet cela est la fonction sigmoïde : une réponse presque bonne est majorée, une réponse intermédiaire reste moyenne, une réponse plutôt mauvaise est minorée. Il fallait donc trouver une formule qui permette de paramétrer la représentation sigmoïdienne, en particulier pour tenir compte de ces critères et de l'effectif. Après des recherches fondées sur l'équation de la fonction sigmoïde, l'équation qui répond le mieux à ce besoin est la suivante :

$$w_i = \frac{1+e^{(-k*(1-l))}}{1+e^{(-k*(((N-i+1)/N)-l))}} = \frac{1+e^{(l-k)}}{1+e^{\left(-k\times\left(\frac{N-i+1}{N}\right)-l\right)}}$$

Équation 10. Poids pour la D-mesure

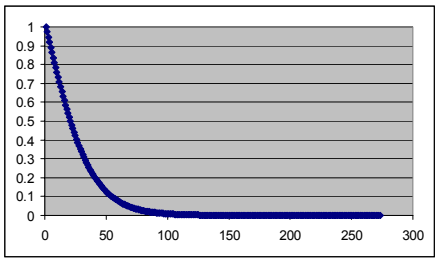
avec k et l, deux paramètres (par défaut k=15, l=0.7), N représente l'effectif total, i le rang de la réponse évaluée.

Observons les propriétés de cette équation en faisant varier les deux paramètres k et l (pour cette illustration, nous avons pris N=273)

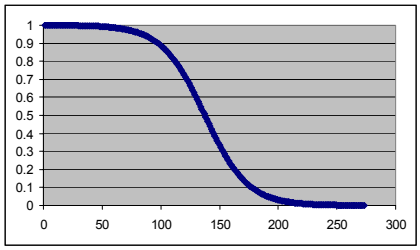


K=15, l=0.7

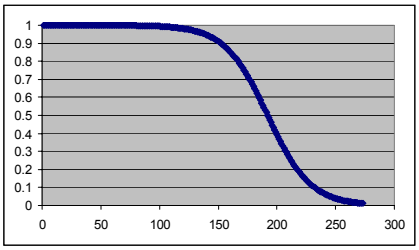
On observe que si la réponse attendue n'apparaît pas dans les 50 premières réponses proposées, le score est fortement minoré, voire quasi nul au delà de 150. Le paramètre l déplace le point d'inflexion, alors que le paramètre k modifie le niveau d'inclinaison :



K=15, l=1

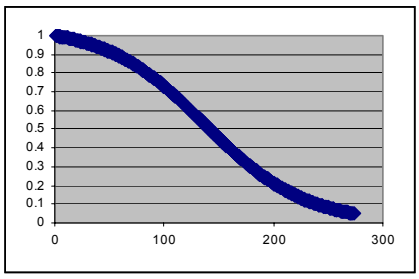


k=15, l=0.5

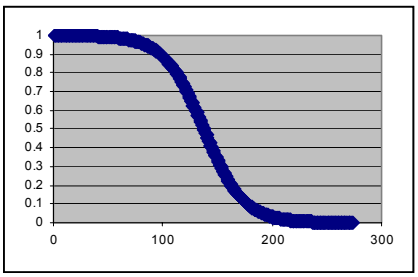


k=15, l=0.2

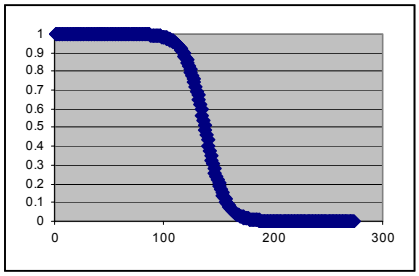
K=6, l=0.5



k=15, l=0.5



k=30, l=0.5



Les deux paramètres permettent donc de générer le niveau d'exigence souhaité. Ce score privilégie donc le fait de proposer en premier les bonnes réponses. Il suffit de multiplier le poids de la réponse par son niveau de pertinence par rapport au résultat attendu. Dans le cas d'un traitement automatique, nous pouvons utiliser une logique booléenne : coder 1 pour une bonne réponse, sinon 0. Dans le cas d'une évaluation humaine, le juge peut donner un pourcentage

L'indice global d'évaluation devient alors :

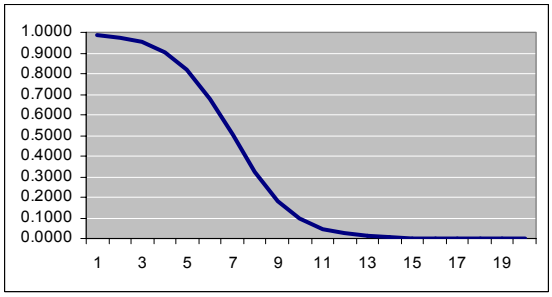
$$D-mesure = \frac{\sum_{i=1}^{N} \frac{Pertinence_i \times 1+e^{(l-k)}}{1+e^{(-k\times\left(\frac{N-i+1}{N}\right)-l))}}}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} \frac{\frac{1+e^{(l-k)}}{1+e^{(-k\times\left(\frac{N-i+1}{N}\right)-l))}} \times 1+e^{(l-k)}}{1+e^{(-k\times\left(\frac{N-i+1}{N}\right)-l))}}}{\sum_{i=1}^{N} w_i}$$

Équation 11. Calcul de la D-mesure

4.5.5 Expérimentation

Voici un exemple de ce que pourrait donner une évaluation selon la D-mesure. Nous avons choisi de valoriser fortement les 5 premières réponses, et de considérer comme faible une réponse correcte donnée au-delà de la 10ème position. Chaque réponse est valorisée par un niveau de pertinence : de 100% (réponse parfaite attendue) à 0% (réponse fausse), avec des valeurs intermédiaires. Le système présenté ci-après obtient un score de 64.4%

Rank i	Weight Wi	Pertinence	Valorized Weight
1	0.9890	100%	0.9890
2	0.9770	0%	0.0000
3	0.9526	100%	0.9526
4	0.9047	50%	0.4523
5	0.8176	70%	0.5723
6	0.6792	80%	0.5433
7	0.5000	100%	0.5000
8	0.3208	0%	0.0000
9	0.1824	0%	0.0000
10	0.0953	100%	0.0953
11	0.0474	100%	0.0474
12	0.0230	100%	0.0230
13	0.0110	100%	0.0110
14	0.0052	0%	0.0000
15	0.0025	100%	0.0025
16	0.0012	100%	0.0012
17	0.0006	75%	0.0004
18	0.0003	100%	0.0003
19	0.0001	30%	0.0000
20	0.0001	100%	0.0001



N =	20
k =	15
L=	0.7
Sum of weights =	6.510
Sum of valorized weights =	4.191
D-MEASURE =	**64.4%**

4.5.6 Conclusion

La F-mesure, créée il y a 25 ans, s'est imposée comme standard d'évaluation. Depuis, les besoins ont évolué, sans remise en cause de cet indicateur. L'analyse de la problématique nous a permis d'introduire une nouvelle mesure : la K-mesure, qui est un sur-ensemble de la F-mesure, capable d'intégrer des niveaux d'exigence de l'utilisateur. Nous avons également examiné comment il était possible d'utiliser la K-mesure pour les classifications.

Enfin, nous proposons une mesure spécifique (la D-mesure) adaptée aux problèmes de recherche d'information.

La K-mesure présente les avantages suivants : d'une part, elle représente une méta mesure de convergence entre la F-mesure de Van Rijsbergen et le break even point de Joachims. D'autre part, elle possède des propriétés mathématiques qui lui permettent de faire un indicateur de synthèse à partir de n'importe quelles autres mesures, dès lors que l'on souhaite les maximiser ensembles. Enfin, elle permet d'intégrer l'approche de Saracevic et de formaliser des niveaux d'exigence (les juges). Il s'agit donc d'une mesure qui est à la convergence de ses trois approches sans en dénaturer le sens.

CHAPITRE 5
LE PROJET RHEA, CONTEXTE DE CETTE THESE

5.1 De la cognition artificielle à la connaissance

La cognition, es qualité de science et de processus, est sans conteste utile au domaine médical. L'envergure du sujet est telle que l'on ne peut l'aborder que de manière sectorielle. Pour cette thèse, nous avons pris pour postulat de départ que la coopération entre informaticiens, médecins, et d'autres professionnels selon les cas, était la meilleure réponse possible pour favoriser l'émergence de nouvelles connaissances utiles, et donc passer de la cognition artificielle à la connaissance.

A cet effet, nous nous sommes associés à une équipe de médecins réanimateurs qui nous a signalé qu'un problème majeur était la détection et la prise en charge des évènements iatrogènes, en particulier des infections nosocomiales. Mais comment modéliser, extraire et calculer de la connaissance médicale ? L'orientation de cette thèse consiste à penser que des mécanismes inductifs exploités par des compétences complémentaires pourraient permettre de répondre à ces problèmes. Afin de confirmer cette hypothèse, il nous est apparu nécessaire de réaliser un datawarehouse élaboré à partir de l'expertise des médecins.

5.2 Définitions

Au sens étymologique, **nosocomial** vient du grec nosos qui signifie maladie et komein qui signifie soigner, mais nosocomium signifie maladie à l'hôpital. L'infection nosocomiale est habituellement définie par sa survenue au-delà de 48 heures après l'entrée à l'hôpital, ce qui signifie qu'elle n'est ni en incubation ni présente à l'admission [Lucet & Astagneau 1998]. Mais l'acception de ce terme est étendue aux soins hors hôpital. Nous considérons que l'infection nosocomiale est **l'infection acquise en se soignant**, ce qui inclut les soins hors établissements hospitaliers, chez un praticien, ou à domicile.

Les **évènements iatrogènes** sont occasionnés par le traitement médical, qu'il y ait ou non erreur de traitement. En grec, le mot signifie littéralement "provoqué par le médecin" (iatros signifie médecin, mais d'autres professionnels de la santé, comme les pharmaciens, peuvent également en être à l'origine).

5.3 Présentation de l'application RHÉA

Dans le cadre du projet labéllisé RNTS Rhéa, nous avons développé, en partenariat avec le laboratoire Prism et la société Oxymel le logiciel Rhéa pour permettre de saisir les données au lit du patient et de calculer en temps réel des indicateurs de gravité et de l'état de santé du patient. Ce logiciel offre de nombreux services complémentaires comme la codification automatique des actes, un requêteur, ou la génération du compte rendu hospitalier rédigé en langage naturel. In fine, les données saisie sont extraites, anonymisées, et transmises cryptées et scellées pour alimenter un entrepôt de données à travers une architecture 3-tiers. Le datawarehouse permet de faire des analyses statistiques et d'extraire des connaissances en utilisant une exploitation biostatistique et des algorithmes de datamining. Les connaissances générées sont réintroduites dans le logiciel pour permettre au praticien de disposer d'alertes sur l'évolution de l'état de santé de son patient.

5.3.1 Principes de fonctionnement

Avec pour objectif de lutter contre les infections nosocomiales, le projet RHÉA collecte une large collection de données sur les patients admis en réanimation, représentés par le 'case mix' des hôpitaux. A ce jour, le projet RHÉA est déployé dans 30 établissements en France et nous avons collecté des données portant sur plus de 3000 patients représentant l'équivalent de 35 000 journées d'hospitalisation. Pour chaque patient, environ 170 données sont collectées au quotidien.

A l'issue de cette saisie, les données de chaque établissement sont transmises et intégrées dans un entrepôt de données centralisé. Les données sont alors soumises à des processus de datamining et des études biostatistiques. Les connaissances générées permettent l'identification des niveaux de risques, la détection des bactériémies et la résistance aux antibiotiques. Les nouvelles connaissances sont alors réinjectées dans les établissements et permettent d'alerter les praticiens.

Les études biostatistiques concernent les domaines suivants :

- modèles d'aggravation des patients,
- règles d'alertes indiquant un risque pour un patient,
- évaluation des procédures visant à lutter contre les infections nosocomiales (durant les actes invasifs, suite à la prescription de médicaments, dues à l'organisation dans l'hôpital, etc.) en vue de les diffuser et de promouvoir les plus efficaces.

Les principes de base sont résumés en Figure 64 : le praticien saisit des données qui alimentent le datawarehouse. Des analyses sont réalisées et génèrent de nouvelles connaissances. Ces dernières sont réinjectées dans le logiciel de saisie. Nous souhaitons découvrir de nouvelles connaissances médicales depuis le datawarehouse. Pour cela, nous devons disposer de données et les praticiens doivent consacrer du temps à leur saisie. En contrepartie, il faut que l'application leur dégage le temps consacré à la saisie. Aussi, plusieurs outils permettent au praticien de gagner du temps : les systèmes d'alertes (incompatibilités entre deux molécules par exemple), restitutions automatiques, génération du codage administratif de la CCAM, contrôles, etc.

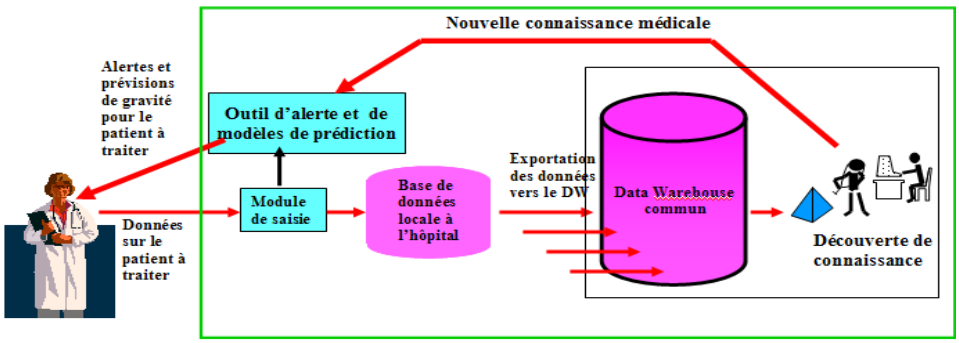


Figure 64: Principes généraux de RHÉA

5.3.2 Architecture

L'architecture générale (présentée en Figure 65) est basée sur 4 modules déployés sur une architecture n-tier. Ces 4 modules sont indépendants mais utilisent les mêmes données (ontologies, connaissances, données administratives …) :

- un module de saisie des données,

- un outil local destiné à réaliser des statistiques, requêtes personnalisées, ainsi qu'un outil de restitution, incluant un tableau de bord et des hypercubes OLAP,
- un projet d'analyse de données textuelles (comptes rendus hospitaliers),
- un projet de planification de soins infirmiers.

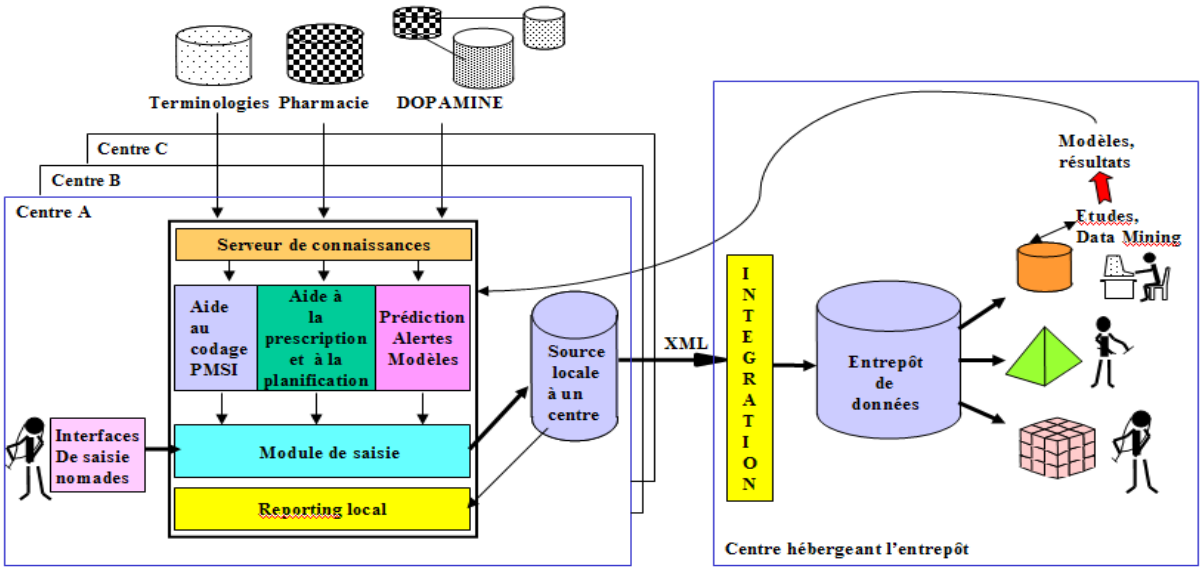


Figure 65. Architecture de RHÉA et ses différents modules

Du point de vue de l'architecture physique, nous avons implémenté une architecture 3-tiers :

- une base de données locale pour la saisie des données par le praticien,
- un serveur WEB et FTP pour les échanges,
- un serveur central pour constituer le datawarehouse et le gérer.

Tous les échanges sont fortement sécurisés.

5.3.3 Enjeux

Les organismes de santé publique estiment que les évènements iatrogènes affectent chaque année environ 2 millions de patients hospitalisés aux USA, pour un coût approximatif de 57,6 milliards de dollars en 2000. Si aucune amélioration n'était apportée, [Null 2003] considère que le nombre de décès pourrait atteindre 7,8 millions de patients au cours des dix prochaines années. En Europe, les infections nosocomiales touchent de 5 à 35% des patients admis dans les services de réanimation [Vincent 1996]. En France, on estime qu'elles augmentent la durée du séjour des patients de 4 jours et coûtent entre 0,73 et 1,8 milliard d'Euros par an [Vanesse 2006]. Selon les sources (ministère ou milieux associatifs), les infections nosocomiales sont responsables de 4 à 10 000 morts par an en France. [Timsit et al.

2005] ont évalué le coût des infections nosocomiales en soins intensifs à partir de l'analyse rétrospective d'une cohorte de 1 698 patients admis entre 1997 et 2000 dans six services de réanimation. Ils aboutissent à un coût moyen allant de 16 300 € pour la BPCO (Broncho Pneumopathie Chronique Obstructive) à 27 900 € pour les péritonites. Le coût moyen en soins intensifs était plus de deux fois supérieur chez les patients ayant une infection sévère par rapport à ceux sans infection sévère. **Le surcoût entre un patient ayant contracté une infection nosocomiale pendant son séjour en soins intensifs et un patient sans infection est de 25 500 €**. Ces études évaluent les coûts pour les seuls soins intensifs et ne prennent pas en compte le degré de sévérité de la maladie ni les conséquences pour les patients.

Pourtant, si nous savions évaluer le pronostic des patients de réanimation et les facteurs de risque d'acquisition des événements iatrogènes et des infections nosocomiales, nous pourrions espérer prévenir la survenue de ces événements et par là même améliorer la qualité des soins.

Peu d'auteurs se sont intéressés à l'utilité de la répétition des mesures des scores au cours du temps [Janssens et al 1998] et à la création de modèles qui incorporent des informations évolutives en cours de séjour. Pourtant, intuitivement, les médecins se servent tous les jours non seulement de l'état des patients, mais aussi de l'évolution de leur état pour déterminer la conduite diagnostique et thérapeutique à suivre.

Pour obtenir des pronostics valides, il faut que les études soient menées sur un échantillon représentatif de la population explorée, c'est-à-dire non seulement sur les patients suivis dans le centre local de réanimation, mais également sur les populations en provenance d'autres centres. Pour prédire le devenir des patients, différents scores de gravité sont calculés à l'admission ou en cours de séjour : MPM II [Lemeshow S 93], APACHE III [Knaus WA 91], IGS II [Le Gall 1993], score Trio, score de Knaus, score de Mac Cabe, score de charge en soins OMEGA. Bien que la capacité de discrimination des scores utilisés en réanimation soit honorable (aire sous la courbe ROC > 0.8), elle reste notoirement insuffisante pour utiliser le rapport mortalité prédite / mortalité observée comme indicateur fiable de performance. En particulier, le pronostic est influencé par la catégorie de patient traitée (regroupé sous le terme anglo-saxon de 'case mix' qui tient compte du diagnostic et des pathologies chroniques). L'écart entre la prédiction de survie et l'observation ainsi que le

manque de discrimination de ces scores peuvent avoir plusieurs causes : le 'case mix', la qualité des soins du service et de l'environnement hospitalier, le pays [Nouira 1998] et l'erreur aléatoire des modèles. Pour Knaus [Knaus.et al 1993], la mortalité hospitalière s'explique par les facteurs suivants :

- la gravité : 73,1 %
- le diagnostic : 13,6 %
- les maladies chroniques : 2,9 %
- l'âge : 3,1%

Enfin, la performance de ces modèles diminue avec la prolongation du séjour [Wunsch 2004].

5.3.4 Objectifs

Le but de RHÉA est d'élaborer un système automatisé, évolutif, et individuel de veille des patients de réanimation, permettant d'estimer en temps réel, un jour donné, la probabilité de survie et les probabilités de survenue de complications. Ce système permet une amélioration du pronostic des patients en tenant compte de leur évolution pendant le séjour en réanimation. L'objectif est également l'identification et la détermination des facteurs de risque et de pronostic des infections nosocomiales (bactériémies, infections sur cathéter, pneumonies nosocomiales, sinusites nosocomiales) et des événements iatrogènes. "*L'événement défavorable doit être distingué des effets secondaires, qui sont des complications imprévisibles et inévitables, et peuvent se produire malgré l'usage de la meilleure pratique*" [Eldar 2002].

Le point fort de RHÉA est d'une part d'allier dans une même suite logicielle les outils décisionnels et les outils de gestion, et, d'autre part, de posséder (et faire vivre) la base de données. Ainsi tous les résultats scientifiques issus du datamining (modèles d'aggravation et autres indicateurs) sont injectés dans la partie experte des outils d'aide. Lors de la saisie quotidienne des données sur un patient le praticien aura un feed-back par comparaison avec les tous derniers résultats issus des études statistiques. Réciproquement, dès qu'un praticien/chercheur a besoin de certaines données dans l'entrepôt pour son étude, les champs de saisie correspondants sont immédiatement ajoutés via le réseau dans une zone prévue à cet effet dans tous les logiciels de saisie de tous les centres, et ceci pour une durée suffisante à l'étude.

5.3.5 Nécessité d'une aide à la décision pour le praticien en réanimation

Tous les jours, les cliniciens doivent ajuster au mieux le traitement de leurs malades en fonction de leur gravité mais aussi du risque qu'ils développent certaines complications (liées à la pathologie sous-jacente, à des évènements iatrogènes, à des infections nosocomiales). Parmi ces complications, les infections nosocomiales viennent au premier plan. Cependant, la relation temporelle entre les facteurs de risque d'acquisition d'événements nosocomiaux, les événements nosocomiaux eux-mêmes, et leur morbidité n'a été que partiellement étudiée. Par exemple, actuellement, on ne sait pas de façon formelle si l'infection nosocomiale par son seul fait aggrave le pronostic des patients ou si l'évolution défavorable des patients conditionne l'apparition d'une infection nosocomiale qui n'en serait alors qu'une conséquence. De même, on ne sait pas si l'acquisition de germes multirésistants influence indépendamment le pronostic des patients de réanimation. Il existe donc très probablement une interaction entre les différents événements nosocomiaux, la gravité initiale, et l'évolution de cette gravité en cours de séjour ([Girou E et al 1998]).

Avec le développement des méthodes d'information et l'augmentation des informations cliniques et paracliniques obtenues quotidiennement chez chaque malade, nous disposons de possibilités nouvelles pour l'instant non exploitées. Il est devenu possible de développer des modèles basés sur les données quotidiennes. Il devient par conséquent possible d'envisager d'utiliser, en temps réel, au lit du malade, l'ensemble de l'information disponible en vue d'une aide décisionnelle la plus fiable possible. Notre outil permet une aide à l'identification plus précoce des patients susceptibles d'évoluer défavorablement et ainsi d'optimiser les ressources disponibles [Goldhill DR et al 1998]. Concernant la détection de l'infection nosocomiale, il peut identifier les patients à haut risque, détecter les épidémies de façon précoce, détecter les modifications d'incidence des infections nosocomiales ou des germes multirésistants. Le système d'information ainsi créé permet également :

- une automatisation du codage nécessaire au bon fonctionnement du service,
- d'évaluer partiellement la performance de chaque service de soins ([Teres D et al 1998], [Fallon et al 1997]),
- d'évaluer la qualité des soins ([Randolph A et al 1998]).

5.3.6 Etat de l'art

5.3.6.1 Un thème ancien mais une approche datawarehouse récente

Au cours du premier congrès mondial d'informatique médicale (1974) [Anderson 1974] qui s'est tenu à Oslo, le dossier médical informatisé faisait déjà l'objet d'une session. Trente trois ans plus tard, le sujet est toujours d'actualité et fait l'objet de nombreuses recherches. A titre d'illustration, lors du neuvième congrès mondial d'informatique médicale (1998) [Cesnik 1998] qui s'est tenu à Séoul, le dossier médical électronique faisait encore l'objet du premier chapitre des Actes du congrès. Le sujet est donc difficile et il est bien loin d'être épuisé [Huet 2000].

Le domaine de la santé représente en France environ 10% des emplois en l'an 2000 [Insee 2002] et développe de nombreux thèmes de recherches. Ceci pourrait laisser supposer que ce secteur est moteur dans les concepts du datawarehouse. Pourtant, lorsque l'on consulte les profils des membres de DM Review en 2001 (voir Figure 66), on constate que seuls 3.53% d'entre eux viennent du monde médical :

Industry Consulting/Systems Integrator	24.61%
Data Processing/Software Services	18.38%
Accounting/Banking/Finance	15.69%
Computer Hardware/Software	7.84%
Telecommunications	5.25%
Manufacturing/Chemicals	4.61%
Retail/Wholesale/Distribution/Food	4.50%
Education	3.74%
Government	3.63%
Health/Medical/Pharmaceutical	**3.53%**

Figure 66. Registered Member Profile in 2001 DM Review

Cet état de fait est surtout le résultat d'une complexité très importante. Dans le cadre d'un datawarehouse traditionnel, l'entreprise maîtrise les concepts. Dans le domaine médical, l'avancée de la recherche médicale et informatique, les nouvelles maladies, les protocoles de traitements ou modalités de prises en charge rendent plus difficiles la modélisation.

5.3.6.2 Qualité de soins

Evaluer la qualité des soins est une tâche complexe qui dépend de nombreux paramètres et facteurs. [Berenholtz 2002] a analysé les publications parues sur ce thème entre 1965 et 2000 et a sélectionné 66 études pour lister les indicateurs de qualité de soins dans les services de réanimation. Il a retenu 6 indicateurs de base : le taux de mortalité, le nombre de séjours supérieurs à 7 jours, la durée moyenne du séjour, le nombre moyen de jours sous ventilation mécanique, la prise en charge de la douleur, et la satisfaction du patient ou de sa famille.

5.3.6.3 Etat de l'art de l'échange de données patients

Une des principales difficultés de la représentation des informations médicales est liée aux différentes normes et standards. Les hôpitaux et services de santé divers collectent des données depuis des années. Face à la nécessité d'échanges électroniques, de nombreux standards différents se sont développés en fonction de la nature de l'information transmise (données médicales diagnostiques, de pharmacie, laboratoires, données comptables, échanges avec les organismes sociaux, …). Aussi, un important travail de recherche a favorisé l'élaboration de nombreux modèles de représentations et de méthodes [Thibault 2002].

Des études sont en cours pour la réalisation d'une norme internationale, tant pour le support (carte Vitale d'assuré social) que pour la nature des données transmises. En France, la normalisation est élaborée par l'AFNOR santé. En Europe, c'est le CEN/TC251 (http://www.centc251.org/) qui se charge de définir un standard. Cette structure de normalisation européenne est organisée en quatre groupes de travail. Le groupe 1 travaille sur les modèles d'informations, le groupe 2 sur la terminologie et les bases de connaissances, le groupe 3 sur la sécurité, la confidentialité et la qualité, et enfin le groupe 4 sur les technologies pour l'interopérabilité. Les Américains utilisent le standard HL7 (http://www.hl7.org/). Ce groupe américain de professionnels de la santé fondé en 1987 pour développer des standards d'échanges de données médicales et accrédité par l'American National Standards Institute (ANSI), base ses travaux sur le SGML. Enfin, les Australiens ont opté pour le standard openEHR (http://www.openehr.org/). A ce jour, ces différents modèles convergent et devraient déboucher sur une norme internationale. Tous ces standards sont

basés sur XML. Une attention particulière doit être portée sur la pré-norme ENV13606 (également appelée EHRCom) qui porte sur l'architecture du dossier médical et sur les messages liés à l'échange d'information. La mise en œuvre du Dossier Médical électronique est planifiée en 2007.

5.3.6.4 Etat de l'art des applications pour la réanimation

Plusieurs types de logiciels sont disponibles pour aider le praticien dans la gestion de leurs données patients, la recherche de connaissances, les restitutions, et parfois le calcul de scores. L'applicatif le plus proche de RHÉA est APPREA. Il s'agit d'un logiciel libre, qui calcule des scores de gravité (http://apprea.free.fr/). Le site http://www.masef.com/freewares2/realog.htm indique plusieurs composants logiciels téléchargeables, en français pour la plupart, mais sur des aspects précis (par exemple le monitoring des cathéters). Les logiciels et sites développés par les médecins semblent prometteurs. Une expérimentation positive d'outils d'alertes par rapport aux risques pour le patient est décrite par [Kucher 2005], mais malgré ses similarités avec RHÉA, ce logiciel n'est pas destiné aux services de réanimation dont les problématiques sont spécifiques. Sur le plan Européen, le plus important projet concernant les infections nosocomiales est le projet HELICS [Sueten 2004] qui associe plusieurs pays Européens en partenariat. Néanmoins, ce projet n'offre pas une approche médicalisée comme RHÉA et reste complémentaire à celui-ci. A notre connaissance, aucun autre logiciel n'est dédié aux services de réanimation.

5.3.7 Présentation de l'application

5.3.7.1 Les données numériques pertinentes

Un important travail a été réalisé par les médecins chercheurs pour déterminer les données pertinentes à saisir chaque jour pour chaque patient. Ce travail a fait l'objet de publications médicales. Au final, 160 variables sont collectées à l'arrivée d'un patient (données initiales), et 170 chaque jour (données évolutives). Il est également possible de stocker des données multimédia.

Les "données initiales" permettent de décrire le patient lorsqu'il arrive dans le service de réanimation (âge, sexe, fumeur, antécédents, motif du séjour, …). Ces données n'évoluent pas pendant le séjour. Elles permettent de calculer automatiquement des scores de gravité à l'admission tels que SAPS II, LODS, OMEGA, NEMS, Glasgow. Ces indicateurs sont utilisés par les médecins pour estimer le pronostic vital du patient.

Les "données évolutives" (voir un exemple Figure 67) sont les données classiques de monitoring (résultats d'analyses, poids, traitements, procédures de ventilation, actes, nutrition, infections nosocomiales, …).

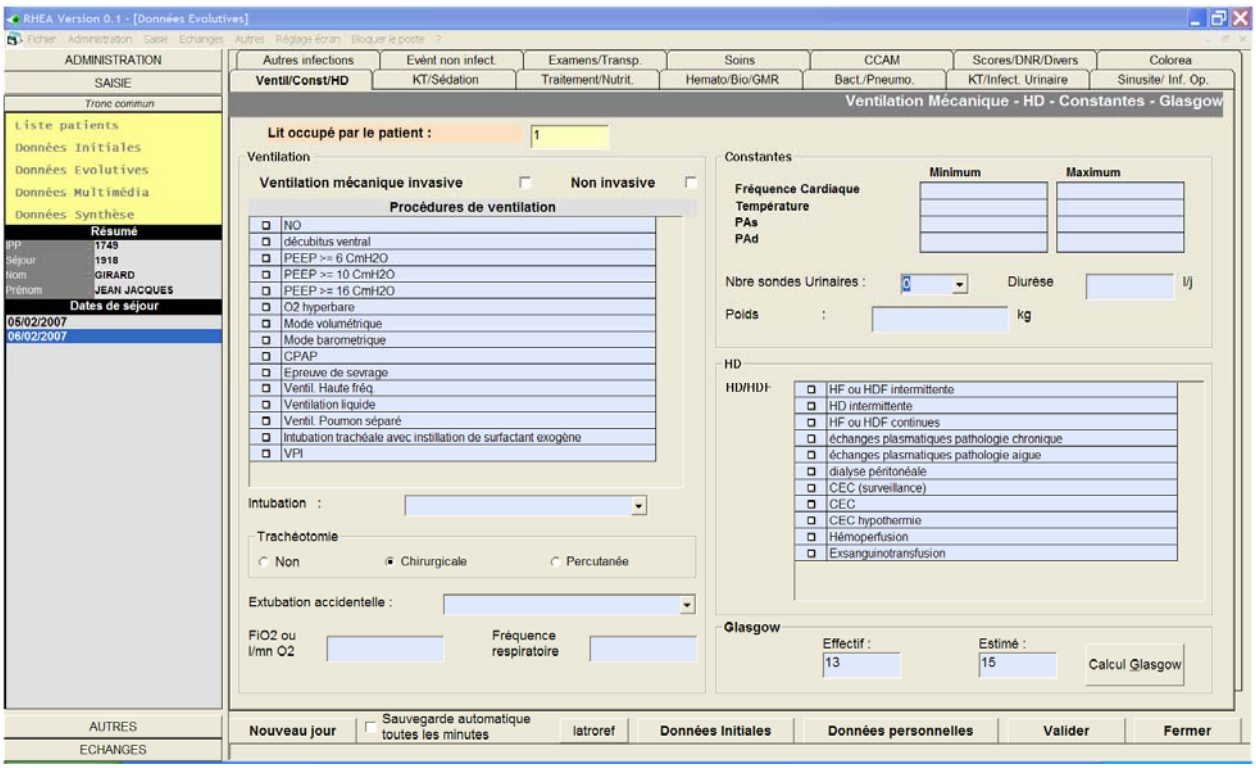


Figure 67. Données évolutives, onglet "procédures de ventilation"

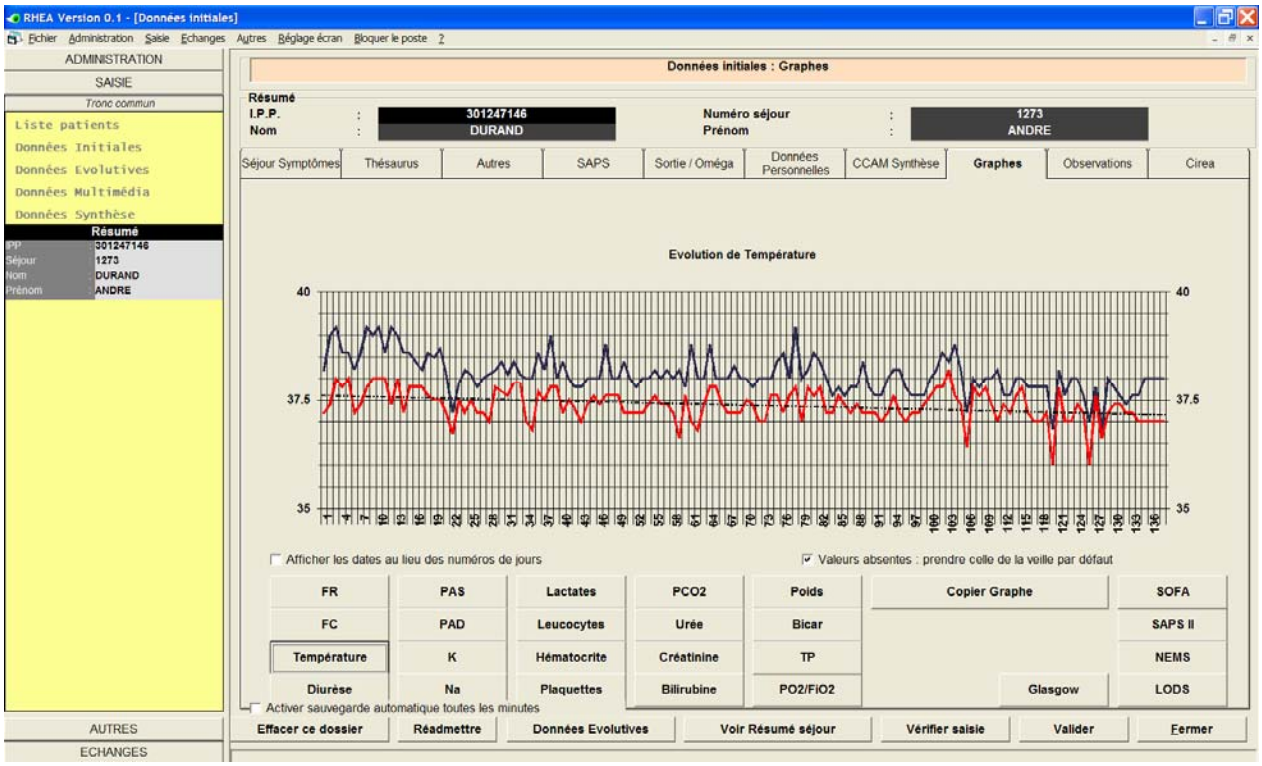


Figure 68. Données initiales, onglet "Graphes"

Environ 180 règles de contrôle permettent de garantir la qualité de la saisie. Le programme RHÉA génère également le compte rendu hospitalier en langage naturel avec une synthèse des évènements importants survenus durant le séjour, les codes CCAM (ontologie des actes médicaux), ...

Une originalité du projet est que l'utilisateur peut ajouter dynamiquement des champs qui correspondraient à l'un de ses centres d'intérêt. Cette fonctionnalité est également accessible à l'administrateur qui peut alors les propager sur tous les postes, par exemple dans le cadre d'une étude ponctuelle.

Une autre possibilité intéressante est d'assurer la continuité des données dans le temps. Par exemple, le champ actuel pour 'fumeur' est booléen : 'oui/non'. Si l'on souhaite détailler ce codage ('non', 'moins de 20/j', 'plus de 20/j'), nous serons capables de restituer l'information sur une période couvrant les deux codages, en prenant le plus large. Bien entendu, il faut que les sous ensembles soient disjoints.

5.3.7.2 Aspects éthiques et sécurité

Le traitement des données médicales est un sujet sensible pour lequel les aspects éthiques et sécurité sont très importants. Dans cette section, nous présentons les outils, techniques et méthodes utilisés pour couvrir ces aspects. Nous appelons "données personnelles" toute information concernant une personne physique identifiable ou potentiellement identifiable, directement ou indirectement.

Les fichiers et données sont sécurisés à plusieurs niveaux : anonymisation, codes de contrôle et cryptage. L'anonymisation consiste à supprimer toute information qui permettrait d'identifier, directement ou par inférence, un patient, un praticien ou un établissement. Le code de contrôle (ou résumé de message) consiste à calculer une clé unique à partir d'un fichier. Quand un fichier est reçu, le code est recalculé et doit correspondre à celui qui est transmis. Si ce n'est pas le cas, c'est que le fichier a été altéré pendant son transfert. Enfin, le cryptage consiste à coder le fichier, ce qui garantit que même si le fichier était intercepté, il ne pourrait être lu.

Anonymisation des données

Aucune information nominative, même indirectement ou par inférence, n'est transférée : les informations nominatives restent dans l'hôpital et sont confidentielles. Le projet RHÉA exporte les données vers le datawarehouse après anonymisation complète (praticien, patient, établissement). Par exemple, les dates des données évolutives sont remplacées par le numéro du jour d'hospitalisation. Si une requête retourne moins de 50 réponses, l'accord du médecin administrateur de l'entrepôt est indispensable. Enfin, il y a rejet si une requête retourne moins de 5 réponses.

Code de contrôle

Le code de contrôle permet de calculer un résumé unique à partir du fichier transmis. Si une modification quelconque, intentionnelle ou non, intervenait entre l'envoi et la réception, le code calculé changerait. Nous avons ainsi la garantie que le message reçu est bien celui qui a été expédié. Les algorithmes de hachage ne sont pas réversibles et ne permettent pas la correction d'erreur. Plusieurs algorithmes existent mais, selon les experts, le

meilleur est SHA-2 avec une clé de 256 bits. Cet algorithme est utilisé par le DOD (Department Of Defense), soit le ministère de la défense américain. Nous avons choisi cet algorithme.

Cryptage

Plusieurs algorithmes de cryptage existent. Nous avons choisi AES (Advanced Encryption Standard) avec une clé de 256 bits, qui remplace l'ancien DES (Data Encryption Standard) avec une clé de 56 bits. Le nombre de combinaisons possibles est très important. S'il existait aujourd'hui une machine capable de casser une clé DES en 1 seconde (cette machine n'existe pas à ce jour), il faudrait 149 000 milliards d'années pour casser la clé AES. Pour donner une idée plus concrète, on considère que l'univers date de 20 milliards d'années au maximum. Nous avons choisi cet algorithme pour sécuriser nos données, plus spécifiquement celui de Rijndael.

Aspects législatifs

L'application RHÉA respecte la législation française. Chaque établissement doit demander l'autorisation auprès de la CNIL (Commission Nationale Informatique et Libertés), en application de la loi 78/17 du 6 janvier. L'article 40-3 de cette loi, relative à la confidentialité des données transmises, impose de préserver le secret médical. Nous respectons également la législation de la communauté européenne (Directive 2001/20/EC, Directive 95/46/EC, Directive 83/570/EEC, Directive 98/44/EC, Directive 90/219/EEC, Directive 2001/18/EC).

5.4 Résultats obtenus avec la base RHÉA

La mise en œuvre de l'entrepôt de données a permis la réalisation de nombreuses publications médicales. L'une d'entre elles [Timsit 2005] a été utilisée pour réaliser le rapport sur les infections nosocomiales présenté au sénat [Vasselle 2006].

5.4.1 Rapport du sénat

Le rapport [Vasselle 2006] de l'office parlementaire d'évaluation des politiques de santé est un rapport sur la politique de lutte contre les infections nosocomiales, réalisé par M. Alain VASSELLE, Sénateur, en juin 2006. L'étude la plus importante citée dans ce rapport est issue d'une publication résultant de la base Rhéa.

5.4.2 Résultats scientifiques médicaux issus de la base RHÉA publiés dans des revues médicales

La base Rhéa a permis de réaliser des avancées scientifiques dans le domaine médical. Nous ne citerons ici que celles ayant fait l'objet d'articles parus dans des revues médicales :

[Timsit 2005] Epidemiology and economic evaluation of severe sepsis in France: age, severity, infection site, and place of acquisition (community, hospital, or intensive care unit) as determinants of workload and cost. Adrie Christophe, Alberti Corinne, Chaix-Couturier Carine, Azoulay Elie, De Lassence Arnaud, Cohen Yves, Meshaka Patrick, Cheval Christine, Thuong Marie, Troche Gilles, Garrouste-Orgeas Maite, Timsit Jean-Francois. Journal of critical care. 2005 Mar ; 20(1):46-58.

[Azoulay 2003] Élie Azoulay, Contact Information, Frédéric Pochard, Maité Garrouste-Orgeas, Delphine Moreau, Laurent Montesino, Christophe Adrie, Arnaud de Lassence, Yves Cohen, Jean-François Timsit and on the behalf of the Outcomerea Study Group : Decisions to forgo life-sustaining therapy in ICU patients independently predict hospital death". Intensive Care Med. 2003 Nov; 29(11):1895-901.

[Timsit 2002] Calibration and discrimination by daily Logistic Organ Dysfunction scoring comparatively with daily Sequential Organ Failure Assessment scoring for predicting

hospital mortality in critically ill patients. Timsit Jean-François, Fosse Jean-Philippe, Troche Gilles, De Lassence Arnaud, Alberti Corinne, Garrouste-Orgeas Maïté, Bornstain Caroline, Adrie Christophe, Cheval Christine, Chevret Sylvie. Crit Care Med. 2002 Sep;30(9):2003-13. PMID: 12352060

[Lassence 2002] Impact of unplanned extubation and reintubation after weaning on nosocomial pneumonia risk in the intensive care unit: a prospective multicenter study. de Lassence Arnaud, Alberti Corinne, Azoulay Elie, Le Miere Eric, Cheval Christine, Vincent Francois, Cohen Yves, Garrouste-Orgeas Maite, Adrie Christophe, Troche Gilles, Timsit Jean-Francois. Anesthesiology. 2002 Jul;97(1):148-56.

[Azoulay 2001] E Azoulay, P Moine, Y Cohen, JF Timsit : Nosocomial pneumopathy acquired under mechanical ventilation. Suspicion criteria, bacteriologic diagnosis, and remission criteria, Rev Pneumol Clin. 2001 Dec;57(6):380-90.

[Moine 2002] Mortality associated with late-onset pneumonia in the intensive care unit: results of a multi-center cohort study. Intensive Care Med. Moine Pierre, Timsit Jean-Francois, De Lassence Arnaud, Troche Gilles, Fosse Jean-Philippe, Alberti Corrine, Cohen Yves.Intensive Care Med. 2002 Feb;28(2):154-63. Epub 2002 Jan 16.

[Timsit 2001] Accuracy of a composite score using daily SAPS II and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 h. Timsit J F, Fosse J P, Troche G, De Lassence A, Alberti C, Garrouste-Orgeas M, Azoulay E, Chevret S, Moine P, Cohen Y. Intensive Care Med. 2001 Jun;27(6):1012-21.

[Clech 2006] Tracheostomy does not improve the outcome of patients requiring prolonged mechanical ventilation: A propensity analysis, in Crit Care Med, Nov 2006, Clec'h C, Alberti C, Vincent F, Garrouste-Orgeas M, De Lassence A, Toledano D, Azoulay E, Adrie C, Jamali S, Zaccaria I, Cohen Y, Timsit JF

[Timsit 2006] Attributable cost of methicillin resistance: an issue that is difficult to evaluate, Timsit JF. Crit Care Med. August 2006;10(4):157.

[Garrouste 2006] Excess risk of death from intensive care unit-acquired nosocomial bloodstream infections: a reappraisal. Garrouste-Orgeas Maite, Timsit Jean Francois, Tafflet Muriel, Misset Benoit, Zahar Jean-Ralph, Soufir Lilia, Lazard Thierry, Jamali

Samir, Mourvillier Bruno, Cohen Yves, De Lassence Arnaud, Azoulay Elie, Cheval Christine, Descorps-Declere Adrien, Costa de Beauregard Marie-Alliette, Carlet Jean. Clin Infect Dis. 2006 Jun 15;42(12):1818. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America.

[Azoulay 2006] Candida colonization of the respiratory tract and subsequent pseudomonas ventilator-associated pneumonia. Azoulay Elie, Timsit Jean-Francois, Tafflet Muriel, De Lassence Arnaud, Darmon Michael, Zahar Jean-Ralph, Garrouste-Orgeas Maite, Cohen Yves, Mourvillier Bruno, Schlemmer Benoit. Chest. 2006 Jan;129(1):110-7

[Timsit 2006b] Pneumothorax in the intensive care unit: incidence, risk factors, and outcome. Timsit Jean-Francois, Tafflet Muriel, De Lassence Arnaud, Azoulay Elie, Garrouste-Orgeas Maite, Cohen Yves, Jamali Samir, Vincent Francois, Alberti Corinne, Dreyfuss, Didier. Anesthesiology. 2006 Jan;104(1):5-13.

[Zahar 2006] Is methicillin resistance associated with a worse prognosis in Staphylococcus aureus ventilator-associated pneumonia? Zahar Jean-Ralph, Clec'h Christophe, Garrouste-Orgeas Maite, Timsit Jean Francois, Tafflet Muriel, Adrie Christophe, Jamali Samir, Mourvillier Bruno, De Lassence Arnaud, Azoulay Elie, Descorps-Declere Adrien, Costa de Beauregard Marie-Alliette, Schwebel Carole. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. Clin Infect Dis. 2006 May 1;42(9):1343. PMID: 16586396.

[Clech 2004] Efficacy of adequate early antibiotic therapy in ventilator-associated pneumonia: influence of disease severity. Clec'h Christophe, Timsit Jean Francois, De Lassence Arnaud, Azoulay Elie, Garrouste-Orgeas Maite, Alberti Corinne, Mourvillier Bruno, Troche Gilles, Tafflet Muriel, Tuil Olivier, Cohen Yves. Intensive Care Med. 2004 Jul;30(7):1327-33. Epub 2004 Jun 9.

[Bornstain 2004] Sedation, sucralfate, and antibiotic use are potential means for protection against early-onset ventilator-associated pneumonia. Bornstain C, Azoulay E, De Lassence A, Cohen Y., Costa MA, Mourvilliers B, Descorps Declere A, Garrouste Orgeas M, Thuong M, Schlemmer B, Timsit JF. Clinical infectious diseases. 2004 May 15;38(10):1409-11. PMID: 15156479.

[Azoulay 2004] Practices in non-neutropenic ICU patients with Candida-positive airway specimens. Azoulay E, De Lassence A, Cohen Y, Zahar JR, Garrouste Orgeas M, Adrie C, Moine P, Timsit JF. Intensive Care Med. 2004 Jul;30(7):1384-9. Epub 2004 May 18.

[Garrouste 2004] Body mass index. An additional prognostic factor in ICU patients.Garrouste-Orgeas Maite, Troche Gilles, Azoulay Elie, Caubel Antoine, de Lassence Arnaud, Cheval Christine, Montesino Laurent, Thuong Marie, Vincent Francois, Cohen Yves, Timsit Jean-Francois. Intensive Care Med. 2004 Mar;30(3):437-43. Epub 2004 Feb 6.

[Azoulay 2003] Determinants of postintensive care unit mortality: a prospective multicenter study. Azoulay Elie, Adrie Christophe, De Lassence Arnaud, Pochard Frederic, Moreau Delphine, Thiery Guillaume, Cheval Christine, Moine Pierre, Garrouste-Orgeas Maite, Alberti Corinne, Cohen Yves, Timsit Jean-Francois, Crit Care Med. 2003 Feb;31(2):428-32.

5.5 Conclusions sur RHÉA

RHÉA est un outil d'aide à la décision fondé sur un entrepôt de données qui permet d'aider le clinicien à prédire la survenue de complications nosocomiales ou d'aggravations des patients en réanimation.

L'entrepôt est alimenté par une extraction sélective et anonymisée à partir de bases de données locales aux centres de réanimation. Son exploitation biostatistique permet de mettre en évidence les corrélations entre les données, et ainsi de spécifier des outils de prédiction sur l'évolution de la gravité en intégrant les risques iatrogènes ou nosocomiaux.

Le logiciel de saisie des données dans chaque centre permet l'acquisition nomade des données au lit du patient, donne des prédictions d'aggravation sur ce patient. A ce jour, le projet RHEA est déployé dans 30 établissements en France et nous avons collecté des données portant sur plus de 3000 patients représentant l'équivalent de 35 000 journées d'hospitalisation. Des contacts sont en cours en Belgique et au Brésil.

La démarche a été de nature générique, tant dans la construction de l'entrepôt, que dans les outils décisionnels associés, afin de permettre une transposition des réalisations dans d'autres domaines d'activités médicales.

Parmi les fonctionnalités couvertes, nous trouvons :

- la génération automatique du résumé du séjour en langage naturel, avec synthèse des faits marquants s'il y a lieu,
- le calcul automatique des scores,
- la génération automatique des codes actes de la CCAM,
- la mise à jour automatique des référentiels,
- l'alimentation de l'entrepôt,
- le transcodage automatique de la biologie en fonction de l'unité de mesure saisie,
- l'affichage de toutes les mesures sous forme de graphique avec calcul de régression pour indiquer les tendances,

Ces informations concernent le patient à son arrivée dans le service (180 variables), la mesure de l'évolution quotidienne de son état de santé (170 variables par jour), les cathéters (à la demande) ou les données multimédia.

CONCLUSIONS

Nous proposons dans ce chapitre de faire le bilan de cette thèse, de rappeler le travail accompli et de montrer les résultats obtenus. Notre problématique initiale était double : d'une part, permettre de retrouver les codes diagnostics (parmi les 52000 codes de la CIM10) qui ont motivé le séjour d'un patient à l'hôpital à partir du compte rendu hospitalier rédigé en langage naturel, d'autre part mettre au point un logiciel permettant la réalisation d'un entrepôt de données (datawarehouse) favorisant l'émergence de nouvelles connaissances sur les infections nosocomiales et évènements iatrogènes.

L'application CIREA visant à retrouver les diagnostics nous a permis d'identifier et de résoudre de multiples problèmes. Nous avons dû dans un premier temps constituer un corpus d'apprentissage ainsi que des dictionnaires :

- base de comptes rendus pour l'analyse du texte,
- base de données Français (dictionnaires de termes et lemmes, liste de stopwords, thésaurus médicaux, liste des préfixes médicaux, …).

En parallèle, nous avons défini avec les praticiens un plan expérimental. La première étape d'analyse a mis en évidence que le taux de reproductivité du codage entre praticiens était très faible, de l'ordre de 18% à trois. Nous avons également mesuré le grand nombre de classes (codes CIM) par compte rendu. Ces constats ont transformé le problème de classification en une approche a trois dimensions en environnement flou. De ce fait, les algorithmes classiques que nous avons expérimenté ont donné de mauvais résultats. Nous avons repris une méthodologie classique d'analyse, en particulier la vectorisation des comptes rendus et mis au point un premier algorithme de désuffixation et d'enrichissement des termes (algorithme EDA) en fonction des préfixes, suffixes et affixes fortement présent dans la sémantique médicale. A partir des documents vectorisés, nous avons ensuite élaboré un nouvel algorithme de classification (algorithme CLO3) dont les résultats sont jugés très satisfaisants par les praticiens. Nous avons également réalisé de nombreuses analyses pour optimiser les résultats, l'apprentissage et les temps de réponse. Ce programme est implémenté dans le logiciel client que nous avons développé.

La faible reproductivité du codage a soulevé le problème de l'évaluation. En effet, comment évaluer correctement un algorithme lorsque nous sommes dans un environnement bruité, c'est à dire sachant qu'il existe une forte incertitude lorsque l'on compare un code prédit avec un code présent à l'origine ? De plus, le niveau d'exigence des praticiens étant élevé, il leur est apparu souhaitable d'avoir une faible variation de la qualité des réponses. En effet, un algorithme très booléen qui proposerait des réponses excellentes ou très mauvaises n'est pas souhaitable. Or la mesure d'évaluation standard, la F-mesure ne permettait de répondre à aucune de ces exigences. Pour compléter l'exposé du problème, il était également souhaitable de garder la comparaison possible avec la F-mesure car de nombreuses publications s'appuient sur celle-ci. Nous avons mis au point deux nouvelles mesures : la K-mesure et la D-mesure qui respectent toutes ces exigences et répondent à notre problème.

Nous avons également développé et implémenté le logiciel Rhéa qui permet de saisir des données médicalisées au lit du malade et les transmet en toute sécurité pour alimenter un entrepôt de données. Ce datawarehouse sert de support à des analyses de datamining et biomédicales pour extraire de nouvelles connaissances, en particulier pour viser à alerter le praticien en temps réel sur les risques nosocomiaux et évènements iatrogènes. L'exploitation de l'entrepôt de données est à l'origine de 17 publications dans des revues médicales par les médecins chercheurs utilisant le logiciel Rhéa. L'une d'entre elles est reprise dans un rapport qui a été rédigé par un sénateur et remis au sénat. Le logiciel est désormais opérationnel dans de nombreux services de réanimation.

BIBLIOGRAPHIE

[Abu 2005] Dictionnaire de la langue française créé par l'Association des Bibliophiles Universels, hébergé au CNAM, disponible sur http://abu.cnam.fr/DICO/

[Aca 1798] Dictionnaire de L'Académie française, 5th Edition (1798), Page 767 http://colet.uchicago.edu/cgi-bin/getobject_?p.16:32./projects/artflb/databases/artfl/dicos/acad1798/IMAGE/

[Ahonen 1998] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Verkamo. Applying data mining techniques for descriptive phrase extraction in digital document collections. In Advances in Digital Libraries (ADL'98), Santa Barbara, California, USA, April 1998, 1998.

[Akaike 1974] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723, 1974.

[Alberti Timsit 2000] C. Alberti, JF. Timsit, R. Porcher, S. Chevret. Contrôle de qualité des données - In Saulnier F, Bion J. Management en réanimation - évaluation, organisation et éthique; Elsevier 2000. PP/112-132

[Amini 2001] M.R. Amini. Apprentissage automatique et recherche de l'information: application à l'extraction d'informations de surface et au résumé de textes. Thèse pour obtenir le grade de docteur de l'université paris 6, soutenue le 13 juillet 2001

[Anderson 1974] Anderson A, Forsythe JM, (eds). Proc. MEDINFO74. Amsterdam : North-Holland PublComp. 1974

[Anton 1988] Anton, Système ACQUEDUC, 1988

[Aphinyanaphongs 2005] Aphinyanaphongs Y., Tsamardinos I., Statnikov, A., Hardin D. and Constantin F. Aliferis: "Text Categorization Models for High-Quality Article Retrieval in Internal Medicine"; doi:10.1197/jamia.M1641, J Am Med Inform Assoc. 2005:12:207-216

[Apté et al. 1994] C. Apté, F. Damerau, S. M. Weiss. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3), 233-251, 1994.

[Apté et al. 1998] C. Apté, F. Damerau, S. M. Weiss. Text mining with decision rules and decision trees. Proceedings of the Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web, 1998.

[Arampatzis et al. 2000] A. T. Arampatzis, Th. P. van der Weide, C. H. A. Koster, P. van Bommel. Term Selection for Filtering based on Distribution of Terms over Time. Proceedings of the Sixth Conference on Content-Based Multimedia Information (RIAO 2000), 1221-1237, 2000.

[Audibert 2002] Laurent Audibert, Etude des critères de désambiguïsation sémantique automatique : présentation et premiers résultats sur les cooccurrences, RÉCITAL 2002, Nancy, 24-27 juin 2002

[Audibert 2003] Laurent Audibert : Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences, in proceeding of TALN 2003, Batz-sur-Mer, 11-14 juin 2003

[Azoulay et al. 2003] Azoulay E, Adrie C, De Lassence A, Pochard F, Moreau D,Thiery G, Cheval C, Moine P, Garrouste-Orgeas M, Alberti C, Cohen Y, Timsit Jf -Determinants of Post-Intensive-Care-Unit Mortality: a prospective multicenter study – Crit Care Med 2003 – 31 ; 428-432

[Bachimont 1998] B. Bachimont. (1998). Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques. Document numérique, 2-3, disponible sur http://www.utc.fr/~bachimon/Publications_attachments/Bachimont-Biblios-AV.pdf

[Badran et al. 2001] M. Lebbah, C. Chabanon, S.Thiria, And F. Badran (2001). "Probabilistic Topological Map and Binary data" in ICANN2001.

[Badran et al. 2001b] M. Yacoub, N. Niang, F. Badran, And S. Thiria (2001) : "A New Hierarchical Clustering Method using Topological Map" ASMDA2001

[Badran et al. 2002] F. Badran, S. Thiria et M. Yacoub : "Cartes auto-organisatrices et classifications automatiques", Ouvrage collectif, Dreyfus et al Eyrolles, 2002.

[Baker 1975] J. A. Baker. The Dragon system – An overview. IEEE Transactions on Acoustic Speech Signal Processing, Vol. ASSP-23, N°1, pp.24--29, 1975.

[Bakis 1976] R. Bakis. Continuous speech recognition via centi-second acoustic states. In Proceedings ASA Meeting. 1976.

[Bartlett 1997] P.L. Bartlett. For valid generalization, the size of the weights is more important than the size of the network Neural Information Processing Systems, 9, Morgan Kaufmann, San Mateo CA, 1997.

[Baud 1992] R. Baud, A.M. Rassinoux, J.R. Scherrer - Natural language processing and medical records. In : Lun K, Degoulet P, Pierre T, Rienhoff O, (eds). Proc. MEDINFO92. Amsterdam : North-Holland Publ Comp. 1992 ; pp. 1362-1367

[Bauer 1999] E. Bauer, R. Kohavi. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants. Machine Learning : Proceedings of the Sixteenth International Conference, Kluwer Academic Publishers, N° 36, pp. 105-- 142, 1999.

[Baum 1966] L. E. Baum, T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. Annual Mathematical Statistics, vol. 37, pp.1554--1563, 1966.

[Baum 1967] L. E. Baum, J. A. Egon. An inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov process and to a Model for Ecology. Bull. Amer. Meterol. Soc., vol. 73, pp. 360--363, 1967.

[Bayes 1763] T. Bayes (1763), « An Essay towards solving a Problem in the Doctrine of Chances », Philosophical Transactions of the Royal Society of London, 53.

[Beaulieu et al. 1997] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, P. Williams, P. Williams. Okapi at TREC-5. Proceedings of the Fifth Text REtrieval Conference (TREC-5), NIST Special Publication 500-238, 143-166, 1997.

[Bennett 2002] Paul N. Bennett, Susan T. Dumais, Eric Horvitz: Probabilistic Combination of Text Classifiers Using Reliability Indicators: Models and Results in proceedings of SIGIR'02, August 1115, 2002, Tampere, Finland. Copyright 2002 ACM 1581135610/02/0008

[Bensadon et al. 1997] M. Bensadon, T. J. William : de la gravité du codage – Collection Informatique et santé, Springer Verlag France, Volume 9, 1997.

[Benzécri 1973] J.P. Benzécri : La taxinomie, Vol. I ; L'analyse des correspondances, Vol II, Dunod, Paris, 1973.

[Berenholtz 2002] Berenholtz SM, Dorman T, Ngo K, Pronovost PJ. Qualitative review of intensive care unit quality indicators. J Crit Care 2002:17:12-5.

[Berners-Lee et al. 2001] T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web, Scientific American. 2001

[Berrut 1989] C. Berrut, P. Cinquin, Natural language understanding of medical reports. In : JR Scherrer, R Coté, S Mandil, (eds). Proc. Computerized natural medical language processing for knowledge engineering, Amsterdam: North-Holland Publ Comp 1989 , pp. 129-137.

[Berrut 1989] Berrut C, Cinquin P, Natural language understanding of medical reports. In: Scherrer JR, Coté R, Mandil S, (eds). Proc. Computerized natural medical language processing for knowledge engineering, Amsterdam: North-Holland Publ Comp 1989, pp. 129-137.

[Besançon 2001] R. Besançon, A. Rozenknop, J-C. Chappelier et M. Rajman : « Intégration probabiliste de sens dans la représentation de textes » - TALN 2001, Tours, 2-5 juillet 2001. Laboratoire d'Intelligence Artificielle, Département Informatique - École Polytechnique Fédérale de Lausanne

[Bigi 2000] B. Bigi, R. De Mori, M. El-Bèze, T. Spriet. A fuzzy decision strategy for topic identification and dynamic selection of language models. Special issue on Fuzzy Logic in Signal Processing, Signal Processing Journal, 80(6), 1085–1097. 2000

[Billsus 1999] D. Billsus and M. Pazzani. A hybrid user model for news story classification. In Proceedings of the Seventh International Conference on User Modeling (UM '99), Banff, Canada, 1999.

[Bishop 1992] C. M. Bishop. Exact Calculation of the Hessian Matrix for the multilayer perceptron. Neural Computation, 4(4), 494-501, 1992.

[Bishop 1995] C. M. Bishop. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995.

[Björck 1967] A. Björck. Solving linear least squares problems by Gram-Schmidt orthogonalization. Nordisk Tidshrift for Informationsbehadlung, 7, 1-21, 1967. Bibliographie 234

[Blanquet 1999] A. Blanquet, P. Zweigenbaum: A lexical method for assisted extraction and coding of ICD-10 diagnoses from text patient discharge summaries. In proceedings of TALN99.

[Bloehdorn 2004] Stephan Bloehdorn, Andreas Hotho: Boosting for Text Classification with Semantic Features, in proceedings of the Workshop on Text-based Information Retrieval (TIR-04) at the 27th German Conference on Artificial Intelligence. SEP 2004.

[Bloehdorn 2006] Stephan Bloehdorn, Philipp Cimiano, and Andreas Hotho : Learning Ontologies to Improve Text Clustering and Classification, Proceedings of the 29th Annual Conference of the German Classification Society (GfKl 2005), Magdeburg, Germany, March 9-11, 2005, volume 30 of Studies in Classification, Data Analysis, and Knowledge Organization, pp. 334-341. Springer, February 2006.

[Bodenreider 2000] Identifying proper name in parallel medical terminologies – MIE 2000 et TAL 2000.

[Bodenreider et Zweigenbaum 2000] O. Bodenreider, P. Zweigenbaum. Indentifying proper names in parallel medical terminologies, Medical Infobahn for Europe - Proceedings of MIE2000 and GMDS2000, Amsterdam, IOS Press, p. 443-447. 2000

[Bonnardel 1995] P. Bonnardel. Mémoire du diplôme d'études spécialisées de pharmacie industrielle et biomédicale : Test statistique Kappa - programmation informatique et applications pratiques. Soutenu le 25 juin 1996. http://kappa.chez-alice.fr/

[Bouaud 1994] Bouaud J., Bachimont B., Charlet J., and Zweigenbaum P., Acquisition an structuring of an ontology within conceptual graphs. In Proceedings of ICCS'94 Workshop on Knowledge Acquisition using Conceptual Graph Theory, pages 1–25, University of Maryland, College Park, MD, August 19 1994

[Boughanem et al. 2000] M. Boughanem, C. Julien, J. Mothe, C. Soule-Dupuy. Mercure at TREC-8: Adhoc, Web, CLIR and Filtering tasks. Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 431-444, 2000.

[Breiman 1994] L. Breiman. Heuristics of instability in model selection, Technical Report, université de Berkeley, 1994.

[Breiman 1996] L. Breiman. Bagging predictors, Machine Learning, vol. 4, pp. 123--140, 1996.

[Brossette 2002] Brossette S., Moser S. "Public administration and health care: data mining and hospital infection control surveillance" Handbook of data mining and knowledge discovery Pages: 928 - 935 Year of Publication: 2002 ISBN:0-19-511831-6 - Oxford University Press, Inc. New York, NY, USA

[Brown 2002] Brown S., Benneyan J., Theobald D., Sands, K., Hahn, M., Gail A. Potter-Bynoe, John M. Stelling, Thomas F. O'Brien, and Donald A. Goldmann: "Binary Cumulative Sums and Moving Averages in Nosocomial Infection Cluster Detection", in Emerging Infectious Diseases - Vol. 8, No. 12, December 2002

[Broyden 1970] C. G. Broyden, The Convergence of a Class of Double-Rank Minimization Algorithms 2: the New Algorithm. Journal Institute of Mathematics and its Applications, 6, 222-231, 1970.

[Brun 2002] A. Brun, K. Smaili, J-P. Haton. WSIM : une méthode de détection de thème fondée sur la similarité entre mots, in proceedings of TALN 2002, Nancy, 24-27 juin 2002

[Brun 2003] A. Brun, K. Smaili, J-P. Haton :"Nouvelle approche de la sélection de vocabulaire pour la détection de thème", TALN 2003, Batz-sur-Mer, 11–14 juin 2003, LORIA BP 239 54506 Vandœuvre-Lès-Nancy, France

[Buntine et Weigend 1991] W. Buntine, A.S Weigend. Bayesian back-propagation. Complex Systems, 5, 603-643, 1991.

[Burnside 2000] Burnside, Strasberg, and Rubin. Automated Indexing of Mammography Reports Using Linear Least Squares Fit. Stanford Medical Informatics, Stanford, CA

[Cai et Hofmann 2003] L. Cai, T. Hofmann. Text categorization by boosting automatically extracted concepts, in proceedings of SIGIR 2003, The 26th Annual International ACM SIGIR Conference, Toronto, Canada, July 28 to August 1, 2003.

[Carry 2002] M. Paternostre, P. Francq, J. Lamoral, D. Wartel et M. Saerens : "Carry, un algorithme de désuffixation pour le Français", Version électronique disponible sur http://www.galilei.ulb.ac.be, 2002

[Cedes et Métais 2002] F. Cedes et E. Métais : "Evolution d'entrepôts de données et détection de changements dans les documents semi-structurés", dans "Le Temps, l'Espace et l'Evolutif en Sciences du Traitement de l'Information", Prade H., Jeansoulin R., Garbay, C. eds., Cépaduès-Editions, septembre 2000.ISBN : 2854285352.

[Celeux 1992] G. Celeux, G. Govaert. A Classification EM algorithm for clustering and two stochastic versions. Computational Statistics and Data Analysis. Vol. 14, pp. 351-- 332, 1992.

[CES 2002] Avis du Conseil économique et social du 15/04/2002 sur Santé et nouvelles technologies de l'information – Jeannette Gros http://www.conseil-economique-et-social.fr/ces_dat2/2-3based/base.htm

[Cesnik 1998] Cesnik B, McCray AT, Scherrer JR, (eds). Proc. MEDINFO98. Amsterdam : IOS Press 1998

[Chapman 2002] W. Chapman, W. Bridewell, P. Hanbury,G. F. Cooper,and B. G. Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 34, 301–310 (2001). doi:10.1006/jbin.2001.1029. Received May 29, 2001; published online May 9, 2002

[Chen et al. 1989] S. Chen, S. A. Billings, W. Luo. Orthogonal least squares methods and their application to non-linear system identification. International Journal of Control, 50(5), 1873-1896, 1989.

[Choueka et Zampoli 1992] Y. Choueka et A. Zampoli. Responsa: An Operational Full-Text Retrieval System with Linguistic Components for Large Corpora: Computational Lexicology and Lexicography: a Volume in Honor of B. Quemada. Pisa: Giardini Press. 1992

[Church 1995] K. W. Church. One Term or Two? Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval 1995: 310–318. 1995

[Church et Hanks 1990] K.W. Church, P. Hanks: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16(1), 22:29. 1990

[Cittee 1999] J. Cittee : Évaluer les référentiels de codage de la morbidité en médecine générale de ville Santé publique 1999, Volume 11, no 4, pp. 503-512. www.sfsp-publichealth.org/Revue/Articles/50848_SFS_4_1999/IMP_CITEE_ps.pdf

[Cohen 1995] W. W. Cohen. Learning to classify english text with ilp methods. In Advances in Inductive Logic Programming (Ed. L. De Raedt), IOS Press, 1995.

[Cohen 1996] W. W. Cohen. Learning rules that classify e-mail. Proceedings of the 1996 AAAI Spring Symposium on Machine Learning in Information Access, 18-25, 1996.

[Cohen et Singer 1996] W. W. Cohen, Y. Singer. Context-sensitive methods for text categorization. Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR '96), 307-315, 1996.

[Cohn et al. 1996] D. Cohn, Z. Ghahramani, M. Jordan. Active learning with statistical models. Journal of Artificial Intelligence Research, 4, 129-145, 1996.

[Costar 1999] Optimizing Healthcare Research Data Warehouse Design through Past - COSTAR Query Analysis – 1999 - Shawn N. Murphy, MD, Ph.D., Mary M. Morgan, G. Octo Barnett, MD, and Henry C. Chueh, MD, M.S. - Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA. www.amia.org/pubs/symposia/D005802.PDF

[Costar 1999] Optimizing Healthcare Research Data Warehouse Design through Past - COSTAR Query Analysis – 1999 - Shawn N. Murphy, MD, Ph.D., Mary M. Morgan, G. Octo Barnett, MD, and Henry C. Chueh, MD, M.S. - Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA. www.amia.org/pubs/symposia/D005802.PDF

[Cover 1965] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers, 14, 326-334, 1965.

[Craven 1998] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the world wide web. In Proceedings of the Fifteenth National Conference on Artificial Intellligence (AAAI98), pages 509-516, 1998.

[Crimmins 1999] F. Crimmins, A. Smeaton, T. Dkaki, and J. Mothe. Tétrafusion: Information discovery on the internet. IEEE Intelligent Systems, 14(4):55-62, 1999.

[Cubréa 1998] P. Aegerter, B. Auvert, G. Buonamico, M. Sznajder, A. Beauchet, B. Guidet, JR Le Gall et CUB-Réa. Mise en œuvre et évaluation d'une base de données commune aux services de réanimation d'Ile-de-France. Rev Epidémiol Santé Pub 1998 ; 46 : 226-237. Et http:/

[Cybenko 1989] G. Cybenko. Approximation by Superpositions of Sigmoidal Function. Signals and Systems, 2, 303-314, 1989. Bibliographie 235

[Dalbin 2002] S. Dalbin : Indexation automatique et langage naturel - Journée d'étude du Jeudi 5 Décembre 2002 à Rouen, des professionnels de l'information et documentation - ADBS Normandie, AIVP et GIDE - Assistance & Techniques Documentaires - DESYBEL GIE

[David 2001] David E. Johnson, Frank J. Oles, Tong Zhang, Thilo Goetz :"A Decision-Tree-Based Symbolic Rule Induction System for Text "

[De Rosnay 1995] J. De Rosnay, "L'homme symbiotique, regards sur le 3ème millenaire", Seuil, 1995, Interview de Joël de Rosnay, Directeur de la Prospective et de l'Evaluation, Cité des Sciences et de l'Insdustrie, Paris. http://csiweb2.cite-sciences.fr/derosnay/articles/bogda.htm

[Dechilly 1999] T. Dechilly, G. Auffret, V. Brunie, & B. Bachimont. (1999). Représentation et structuration des connaissances pour les bibliothèques audiovisuelles.

[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), 391-407, 1990.

[Delassence et al. 2002] A. Delassence, C. Alberti, E. Azoulay, E. Le Mierre, C. Cheval, F. Vincent, Y. Cohen, M. Garrouste Orgeas, C. Adrie, G. Troche, J-F. Timsit for the Outcomerea study group – Impact of unplanned extubation and reintubation after weaning on nosocomial pneumonia risk in the intensive care unit – Anesthesiology 2002; 97 : 148-156

[Delichere 2002] M. Delichère, D. Memmi, Analyse Factorielle Neuronale pour Documents Textuels, in proceedings of TALN 2002, Nancy, 24-27juin 2002

[Dempster 1977] A. Dempster, N. Laird, D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Vol. B, N° 39, pp. 1-- 38, 1977.

[Dewey] http://dewey.yonsei.ac.kr/memexlee/links/categorization.htm

[Dice 1945] L.R. Dice: Measures of the Amount of Ecologic Association between Species. Journal of Ecology, 26, 297:302. 1945

[Doc 2006] Arrêté du 7 août 2006 relatif à la formation doctorale ; arrêtés des 6 janvier 2005 et 7 août 2006 relatifs à la cotutelle internationale de thèse ; arrêté du 3 septembre 1998 relatif à la charte des thèses, arrêté du 7 août 2006 relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou travaux présentés en soutenance en vue du doctorat. www.sup.adc.education.fr/bib/Acti/These/guidoct.rtf

[Dolisi 2005] http://georges.dolisi.free.fr/

[Dubois 1994] J. Dubois, L. Guespin, M. Giacomo, C. Marcellesi, J-B. Marcellesi, J-P. Mével, 1994, Dictionnaire de linguistique et des sciences du langage. Collection Trésors du Français, Larousse, Paris.

[Dubois 2004] Dubois J., Guespin L., Giacomo M., Marcellesi C., Marcellesi J.-B., Mével J.-P., 1994, "Dictionnaire de linguistique et des sciences du langage". Collection Trésors du Français, Larousse, Paris.

[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), 148-155, 1998.

[Dunning 1994] T. Dunning. Statistical identification of language, Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University. 1994

[E-datawarehouse 2001] E-Healthcare Data Warehousing. journal of ealthcare information management®, vol. 15, no. 2, Summer 2001 - www.himss.org/content/files/jhim/15-2/him15201.pdf

[Efron 1993] B. Efron, R. Tibshirani. An Introduction to the Bootstrap, Chapman & Hall, 1993.

[El-Beze 1995] M. El-Beze, B. Merialdo, B. Rozeron et A.-M. Derouault. Accentuation automatique de textes par des méthodes probabilistes, Technique et science informatiques, 13 (6), p. 797-815. 1995

[Eldar 2002] Eldar R, Understanding and Preventing Adverse Events, Croatian Medical Journal, 43(1), 2002.

[Eliassi 1999] J. S. . T. Eliassi-Rad. Intelligent agents for web-based tasks: An advice-taking approach. In Working Notes of the AAAI/ICML-98 Workshop on Learning for Text Categorization, Madison, WI, pages 588-589, 1999.

[Evans et Lefferts 1995] D. A. Evans, R. Lefferts. CLARIT-TREC experiments. Information Processing and Management, 31(3), 385-395.

[Eyheramendy 2003] S. Eyheramendy, David D. Lewis, David Madigan : On the Naive Bayes Model for Text Categorization (2003) - Ninth International Workshop on Artificial Intelligence and Statistics

[Fagon 1993] Fagon JY, Chastre J, Novara A, Medioni P, Gibert C - Characterization of intensive care unit patients using a model based on the presence or absence of organ dysfunctions and/or infection: the ODIN model. Intensive Care Med 1993; 19:137-144

[Fano 1961] R. M. Fano. Transmission of Information: A statistical Theory of Communications. MIT Press, Cambridge, MA, 1961

[Feldman 1995] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), pages 112-117, Montreal, Canada, 1995.

[Feldman 1995] Feldman R. and Dagan I. (1995). KDT – Knowledge Discovery in texts, Proceedings of the 1st Int. Conf. On Knowledge Discovery (KDD-95), Aug., 1995.

[Feldman 1998] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. Text mining at the term level. In Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, volume 1510 of Lecture Notes in Computer Science, pages 56-64. Springer, 1998.

[Féraud 1787] J-F. Féraud: « Dictionnaire critique de la langue française » (Marseille, Mossy 1787-1788). CLASSIFICATION (Page A458a). http://colet.uchicago.edu/cgi-bin/getobject_?p.0:918./projects/artflb/databases/artfl/dicos/feraud/IMAGE/

[Flechter 1970] R. Flechter. A New Approach to Variable Metric Algorithms. The Computer Journal, 13(3), 317-322, 1970.

[Frakes 1992] W. Frakes, R. Baeza-Yates (1992). Information Retrieval : Data Structures and Algorithms. Prentice Hall, Englewood Cliffs,NJ.

[Frakes 2003] William B. Frakes : "Strength and Similarity of Affix Removal Stemming Algorithms", SIGIR Forum 37(1): 26-30 (2003)

[Frank 1999] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-specific keyphrase extraction. In Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99, pages 668-673, 1999.

[Freitag 1999] D. Freitag and A. McCallum. Information extraction with hmms and shrinkage. In Proceedings of the AAAI- 99 Workshop on Machine Learning for Information Extraction, 1999.

[Freund 1990] Y. Freund. Boosting a weak learning algorithm by majority. Proceedings of the third Annual Workshop on Computational Learning Theory, pp. 202--216, 1990.

[Freund 1995] Y. Freund, R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Proceedings of the second European Conference on Computational Learning Theory. Springer-Verlag, pp. 23-37, 1995.

[Freund 1996] Y. Freund, R. E. Schapire. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kauffmann, pp. 148--156, 1996.

[Friedman 1993] C. Friedman, J. Cimino, S. Johnson, A conceptual model for clinical radiology reports In : Proc. 17th Annual SCAMC93. New York : Mc Graw Hill, 1993 ; pp. 829-833

[Friedman 1996] C. Friedman, PO Alderson, JH Austin, JJ Cimino, et SB Johnson. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-74.

[Funahashi 1989] K. Funahashi. On the Approximate Realization of Continous Mappings by Neural Networks. Neural Networks, 2, 183-192, 1989.

[Fürnkranz 1999] J. Fürnkranz. Exploiting structural information for text classification on the www. In Advances in Intelligent Data Analysis, Third International Symposium, IDA-99, pages 487-498, 1999.

[Gallinari et Cibas 1999] P. Gallinari, T. Cibas. Practical complexity control in multilayer perceptrons. Signal Processing, 74, 29-46, 1999.

[Gardarin 2002] Georges Gardarin : XML : des Bases de Données aux Services Web, 514 pages, Ed. Dunod, Novembre 2002.

[Geman et al. 1992] S. Geman, E. Bienenstock, R. Doursat. Neural Network and the Bias/Variance Dilemma. Neural Computation, 4, 1-58, 1992.

[Girou et al 1998] Girou E, Pujade G, Legrand P, Cizeau F, Brun-Buisson C. Selective screening of carriers for control of methicillin-resistant Staphylococcus aureus (MRSA) in high-risk hospitals areas with a high-level of endemic MRSA. Clin Infect Dis. 1998; 27: 543-550

[Goldfarb 1970] D. Goldfarb. A Family of Variable Metric Methods Derived by Variational Means. Mathematics of Computation, 24, 23-26, 1970.

[Goldhill DR et al 1998] Goldhill, David R. MA, MBBS, FRCA; Sumner, Anne: Outcome of intensive care patients in a group of British intensive care units. Clinical Investigations. Critical Care Medicine. 26(8):1337-1345, August 1998.

[Goldman 1999] R. Goldman and J. Widom. Approximate dataguides. In Proceedings of the Workshop on Query Processing for Semistructured Data and Non-Standard Data Formats, 1999.

[Goonarch 2005] http://pinpon29.free.fr/vocamedical/

[Grabar 1999] N Grabar, P. Zweigenbaum : Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In Pascal Amsili, rédacteur, Actes de TALN 1999 (Traitement automatique des langues naturelles), pages 175-184, Cargèse, juillet 1999. ATALA.

[Grabar 2000] N. Grabar, P. Zweigenbaum: Automatic acquisition of domain-specific morphological resources from thesauri. In Proceedings of RIAO

[Grappa] http://www.grappa.univ-lille3.fr/polys/apprentissage/sortie004.html

[Grenier 1993] Grenier : Décision médicale, Masson, 1993.

[Gross 1986] M. Gross. Lexicon-grammr, The representation of compound words, proceedings of COLING-86, Bonn, p. 1-6. 1986

[Grossman et Frieder 1998] D. Grossman, O. Frieder. Information Retrieval. Algorithms and Heuristics. Kluwer Academic Publishers, 1998.

[Grover 2002] C. Grover, E. Klein, M. Lapata and A. Lascaride : "XML-Based NLP Tools for Analysing and Annotating Medical Language" - University Of Edinburgh

[Grumbach 1999] S. Grumbach and G. Mecca. In search of the lost schema. In Database Theory - ICDT '99, 7th International Conference, pages 314-331, 1999.

[Guidet 1999] B. Guidet, B. Misset : Recommandations actuelles de codage d'un séjour de réanimation - http://www.sfar.org/thesaurus/recodage.pdf

[Gull 1988] S. F. Gull. Bayesian inductive inference and maximum entropy. G. J. Erickson, C.R. Smith eds. Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1: Foundations, 53-74, Dordrecht: Kluwer, 1988. Bibliographie 236

[Guttman 1941] L. Guttman: The Quantification of a Class of Attributes. In 'The prediction of personal adjusment, P.Horst ed., SSCR (New York), 1941

[Habert et Jacquemin 1993] B. Habert, C. Jacquemin. Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques, Traitement Automatique des Langues (TAL), n° 2, p. 5-41. 1993

[Hand 2001] Hand, D.J., Yu, K. Idiot's Bayes - not so stupid after all ?, in International Statistical review, 2001

[Harbeck et Ohler 1999] S. Harbeck et U. Ohler. "Multigrams for Language Identification", Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 99), Budapest, Hungary. 1999

[Hardt 1988] Hardt, S.L. On recognizing planned deception. In AAAI-88 Workshop on plan recognition, 1988.

[Harman 1991] D. Harman. How effective is suffixing? Journal of the American Society for Information Science, 42(1), 321-331, 1991.

[Hayashi 1956] C. Hayashi : Theory and Examples of Quantification (II) Proc. of the Institute of Stat. Math., 4 (2) p. 19-30, 1956

[Hayes 1990] Hayes PJ and Weinstein SP (1990) Construe/tis: A system for content-based indexing of a database of new stories. In: Second Annual Conference on Innovative Applications of Artificial Intelligence.

[Hinton 1987] G.E. Hinton. Learning Translation Invariant Recognition in Massively Parallel Networks. Proceedings PARLE Conference on Parallel Architectures and Languages Europe, J. W. de Bakker, A. J. Nijman, P.C. Treleaven (Eds.), Springer-Verlag, Berlin, 1-13, 1987.

[Hofmann 1999] T. Hofmann. The cluster abstraction model: unsupervides learning of topic hierarchies from text data. In proceedings of 16th International Joint Conference on Artificail Intelligence IJCAI99, pages 682-687, 1999

[Hoi 2006] Steven Hoi, Rong Jin, Michael R. Lyu : Large-Scale Text Categorization by Batch Mode Active Learning, in proceedings of WWW2006, the 15th international World Wide Web conference, Edimburgh, May 23–26, Scotland, 2006

[Honkela 1997] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Websom - self-organizing maps of document collections. In Proc. of Workshop on Self-Organizing Maps 1997 (WSOM'97), pages 310-315, 1997.

[Horn 1989] L.R. Horn. A natural history of negation. Chicago, IL: Univ. of Chicago Press, 1989.

[Hornik et al. 1994] K. Hornik, M. Stinchcombe, H. White & P. Auer. Degree of Approximation Results for Feedforward Networks Approximating Unknown Mappings and their Derivatives. Neural Computation, 6(6), 1262-1275, 1994.

[Huet 2000] JFIM mai 2000 - Méta modélisation du dossier médical : Conception, Intérêt, Application. Bernard Huet, Bruno Lesueur, Jean-Yves Artigou, Gilles Blain

[Hull 1993] D. A. Hull. Using Statistical testing in the evaluation of the retrieval performance. Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR'93), 329-338, 1993.

[Hull 1994] D. A. Hull. Information retrieval using statistical classification. Ph.D. thesis, Stanford University, 1994.

[Hull 1996] D. A. Hull. Stemming Algorithms – A Case Study for Detailed Evaluation. Journal of the American Society for Information Science, 47(1), 70-84, 1996.

[Hull 1999] D. A. Hull. The TREC-7 Filtering Track: Description and Analysis. Proceedings of the Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242, 33-56, 1999.

[Hull et Robertson 2000] D. A. Hull, S. Robertson. The TREC-8 Filtering Track Final Report. Proceedings of the Eigth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 35-56, 2000.

[Insee 2002] INSEE : structure des emplois par domaine - novembre 2002 - http://www.insee.fr/fr/ffc/docs_ffc/iress11.pdf

[ISO-latin 1987] ISO 8859-1, Information processing — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No. 1. 1987

[Iwayama 1994] Iwayama, Makoto, Tokunaga, Takenobu: “A probabilistic model for text categorization based on a single random variable with multiple values”. TITCS 1994, ISSN : 0918-2802

[Iyer et al. 2000] R. D. Iyer, D. D. Lewis, R. E. Schapire, Y. Singer, A. Singhal. Boosting for Document Routing. Proceedings of the ninth International Conference on Information and Knowledge Management (CIKM 2000), 2000. Bibliographie 237

[Jacquemin 2000] Traitement automatique des langues pour l’accès au contenu des documents – I3 2000

[Jacquemin et Tzoukermann 1999] C. Jacquemin et E. Tzoukermann : NLP for Term Variant Extraction : Synergy between Morphology Lexicon and Syntax, Natural Language Information retrieval, T. Strzalkowski (ed.), Kluwer. 1999

[Jacquemin et Zweigenbaum 2000] C. Jacquemin et P. Zweigenbaum. Traitement automatique des langues pour l'accès au contenu des documents. In J. Le Maître, J. Charlet and C. Garbay, editors, Le document Multimédia en Sciences du Traitement de l'Information, CÉPADUÈS-Éditions, Toulouse, p. 71-110. 2000

[Jaillet 2004] Simon JAILLET : "Catégorisation automatique de documents. LIRMM, UMR 5506, 161 rue Ada, 34392 Montpellier Cedex 5 - France. jaillet@lirmm.fr, 2004. http://www.lirmm.fr/doctiss04/art/I02.pdf

[Jalam 2003] Radwan Jalam : "Apprentissage automatique et catégorisation de textes multilingues", thèse pour obtenir le grade de docteur en informatique présentée et soutenue publiquement par le 4 juin 2003. université lumière lyon2, préparée au sein du laboratoire ERIC, équipe de recherche en ingénierie des connaissances, sous la direction de jean-hugues Chauchat. Année 2003.

[Jelinek 1980] F. Jelinek, R. Mercer: Interpolated estimation of markov source parameters from sparse data. In Proceedings of Workshop Pattern Recognition in Practice, 1980, p. 381–397, Amsterdam.

[Jing et Tzoukermann 1999] H. Jing, E. Tzoukermann. Information Retrieval Based on Context Distance and Morphology. Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR '99), 90-96, 1999.

[Joachims 1997] Thorsten Joachims : A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization in Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997

[Joachims 1997b] Thorsten Joachims, D. Freitag, and T. Mitchell. Webwatcher: a tour guide for the world wide web. In Proceedings of IJCAI-97, 16th International Joint Conference on Artificial Intelligence, Pages 770-777, 1997

[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, 137-142, 1998.

[Jujols 1991] Jujols P, Aubas P, Baylon C et al. Morphosemantic Analysis and Translation of Medical Compound Terms. Meth Inform Med 1991; 30:30-5.

[Junker 1999] M. Junker, M. Sintek, and M. Rinck. Learning for text categorization and information extraction with ilp. In Proceedings of the Workshop on Learning Language in Logic, Bled, Slovenia, 1999, 1999.

[Kargupta 1997] H. Kargupta, I. Hamzaoglu, and B. Stafford. Distributed data mining using an agent based architecture. In Proceedings of Knowledge Discovery And Data Mining, pages 211-214. AAAI Press, 1997.

[Kerkri 1998] Elmostafa Kerkri, Catherine Quantin, Kokou Yetongnon, et Liliane Dusserre : les entrepôts de données, application au suivi épidémiologique – Collection Informatique et santé, Springer Verlag France, Volume 9, 1998.

[Kindermann et Leopold 2000] J. Kindermann, E. Leopold. Classification of Texts with Support Vector Machines. An estimation of the efficiency of Kernels and Data- Transformations. Proceedings of the 24th Annual Conference of the Gessellschaft für Klassifikation, 2000.

[Knaus 1985] Knaus WA, Drapper EA, and Wagner DP: APACHE II: a severity of disease classification system. Crit Care Med 1985; 12:975-977

[Koller et Sahami 1996] D. Koller, M. Sahami. Toward optimal feature selection. Proceedings of the Thirteenth International Conference on Machine Learning, 1996.

[Koller et Sahami 1997] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In Proceedings of the 14th International Conference on Machine Learning, Nashville, Tennessee, July 1997.

[Kononenko 1990] Kononenko, I. "Comparison of inductive and naive bayesian learning approaches to automatic knowledge acquisition". In B. Wielinga et al. (eds.) Current trends in knowledge acquisition, Amsterdam, IOS Press., 1990

[Kontostathis 2003] Kontostathis A., Galitsky L., Pottenger W. Roy, S. and Daniel J. Phelps. (2003) A Survey of Emerging Trend Detection in Textual Data Mining. In A Comprehensive Survey of Text Mining, Michael Berry, Ed., Springer-Verlag.

[Kou et Gardarin 2002] Huaizhong Kou, Georges Gardarin: Similarity Model and Term Association For Document Categorization. DEXA Workshops 2002: 256-260

[Kou et Gardarin 2002b] Huaizhong Kou, Georges Gardarin: Study of category score algorithms for k-NN classifier. SIGIR 2002: 393-394

[Krishnan 1987] T. Krishnan, S. C. Nardy. Discriminant analysis with a stochastic supervisor. Pattern Recognition, Vol. 20, pp. 379--384, 1987.

[Krogh et Hertz 1992] A. Krogh, J.A. Hertz. A Simple Weight Decay Can Improve Generalization. Advances in Neural Information Processing Systems, 4, J.E. Moody, S.J. Hanson and R.P. Lippmann, eds., Morgan Kauffmann Publishers, San Mateo CA, 950-957, 1992.

[Krovetz 1993] R. Krovetz. Viewing morphoogy as an inference process. Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR'93), 191-203, 1993.

[Kucher 2005] Kucher N., "electronic alerts to prevent venous thromboembolism among hopitalized patients" N Enl J Med 2005; 352:969-77.

[Kwok et al. 2000] K. L. Kwok, L. Grunfeld, M. Chan. TREC-8 Ad-Hoc, Query and Filtering Track Experiments using PIRCS. Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 217-228, 2000.

[Landau et al. 1993] M.-C. Landau, F. Sillion, F. Vichot. Exoseme: a Thematic Document Filtering System based on conceptual graphs. Proceedings of the International Conference on Conceptual Structures (ICCS'93), 171-187, 1993. Bibliographie 238

[Landis et Koch 1977] J.R. Landis, G.G. Koch: The Measurement of Observer Agreement for Categorical Data, Biometrics, 1977a, 33, 159-174.

[Laporte 2000] E. Laporte. Mots et niveau lexical, Ingénierie des langues, Hermes, p. 25-49. 2000

[Larkey 1996] Leah Larkey : Combining Classifiers in Text Categorization (1996) , in Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval

[Le Beux 2001] P. Le Beux : Systèmes de représentation du langage Médical – Faculté de Médecine – Support de cours - Université de Rennes 1

[Le Gall 1993] Le Gall JR, Klar J, Lemeshow S, Saulnier F, Alberti C, Artigas A, Teres D: The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. JAMA 1996; 276:802-10

[Lebart et Al. 1995] L. Lebart, A. Morineau, M. Piron : Statistique exploratoire multidimensionnelle, Dunod, Paris, 1995

[Ledoray 1992] V. Ledoray, B. Giusiano, M. Roux. A system for understanding medical reports: architecture and knowledge required. In Lun KC and al. Eds. Medinfo 92. Amsterdam: Elsevier science publ. 1992. 1389-94

[Ledoray 1992b] Ledoray V., Pellegrin L., Guisiano B., Roux M, Système de compréhension des comptes rendus médico-techniques : architecture, connaissances nécessaires et résultats, In : Degoulet P. et coll., (eds.) Informatique et santé. France : Springer-Verlag, 1992, pp. 111-125.

[Lemeshow 1994] Lemeshow S, Klar J, Terres D, Avrunin JS, Gehlbach SH, Rapoport J, Rue M: Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective multicenter study. Crit Care Med 1994;22:1351-1358

[Lewis 1992] D. D. Lewis. Representation and Learning in Information Retrieval. Ph.D. thesis, Graduate School of the University of Massachusetts, 1992.

[Lewis 1992b] D. D. Lewis: An Evaluation of Phrasal and Clustered Representation on a Text Categorization Task. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1992, p. 37–50.

[Lewis et al. 1996] D. D. Lewis, R. E. Schapire, J. P. Callan, R. Papka. Training algorithms for linear text classifiers. Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR '96), 298-306, 1996.

[Lewis et Ringuette 1994] D. D. Lewis, M. Ringuette. Comparison of two learning algorithms for text categorization. Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), 181-93, 1994.

[Li 2005] Li, Shenzhi and Pottenger, William M., (2005) DiHO: "A Distributed Higher-Order Association Rule Miner". In the Proceedings of the 24th ACM SIGMOD International Conference on Management of Data. Baltimore, MD, June.

[Likita 2004] http://membres.lycos.fr/likita/MO/prsu.htm

[Lohdi 2002] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins : Text Classification using String Kernels (2002)

[Lombrail et al. 1991] P. Lombrail, E. Minvielle, F. Kohler, D. Hève, D. Mayeux, M. Naiditch, M-L. Pibarot ML, M. Brémond. : problèmes posés par le codage de l'information médicale dans le cadre du programme de médicalisation des systèmes d'information hospitalier. Rev Epid S Pub 1991 ; 39(3) :285-95

[Lombrail et al.] Lombrail P., Minvielle E., Kohler F., Hève D., Mayeux D., Naiditch M, Pibarot ML, Brémond M. : problèmes posés par le codage de l'information médicale dans le cadre du programme de médicalisation des systèmes d'information hospitalier. Rev Epid S Pub 1991 ; 39(3) :285-95

[Loose 1998] R. Losee: "Text Retrieval and Filtering: Analytic Models of Performance", chapitre 3, 1998

[Lovins 1968] J. Lovins: Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11, 22-31, 1968.

[Lovins 1968] J.B. Lovins: "Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics", 11 (1-2), pp. 22-31, (1968).

[MacFarlane et Robertson 2000] A. MacFarlane, S. E. Robertson. PLIERS at TREC-8. Proceedings of the Eigth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 241-252, 2000.

[MacKay 1992a] D. J. C. MacKay. Bayesian interpolation. Neural Computation, 4(3), 415- 447, 1992.

[MacKay 1992b] D. J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. Neural Computation, 4(3), 448-472, 1992.

[MacKay 1992c] D. J. C. MacKay. The Evidence Framework Applied to Classification Networks. Neural Computation, 4(5), 698-714, 1992.

[MacKay 1999d] D. J. C. MacKay. Comparison of Approximate Methods for Handling Hyperparameters. Neural Computation, 11(5), 1035-1068, 1999.

[Manning 1999] C.D. Manning, H. Schütze: "Foundations of Statistical Natural Language Processing". MIT Press, 1999. Bibliographie 239

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

[Maoussc 96] http://www.med.univ-rennes1.fr/plaq/cristal/MAOUSSC.html

[McCallum et Nigam 1998a] A. McCallum, K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization, 41-48, 1998.

[McCallum et Nigam 1998b] A. McCallum, K. Nigam. Employing EM and Pool-Based Active Learning for Text Classification. Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98), 350-358, 1998.

[McDonough 1994] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish & J. Rohlicek : Approaches to Topic Identification On The Switchboard Corpus. In IEEE Transactions on Acoustics, Speech, and Signal Processing", 1994, p. 385–388.

[McLachlan 1992] G.J. McLachlan: Discriminant Analysis and Statistical Pattern Recognition. Wiley, New-York, 1992.

[McLachlan 1997] G. McLachlan, T. Krishnan. The EM Algorithm and Extensions. Wiley, New York, 1997.

[MedLEE 95] http://lucid.cpmc.columbia.edu/medlee/

[Meilland 2003] Jean-Claude Meilland, Patrice Bellot : "Extraction automatique de terminologie à partir de libellés textuels courts". Presses Universitaires de Rennes - janvier 2003 "Linguistique de corpus".

[Menelas 94] http://www.biomath.jussieu.fr/Menelas/

[MeSH 2007] http://ist.inserm.fr/basismesh/meshv07.html#

[Métais 2000] E. Métais et le groupe EVOLUTION "Les entrepôts de données", dans le volume "Informatique et systèmes d'information" de l'encyclopédie "Information, Commande, Communication", Hermès, 2000.

[Métais 2002] "Enhancing Information System Management with Natural Language Processing Techniques", Elisabeth Métais, revue DKE 41(1), 2002.

[Métais et Mayr 99] "Applications of Natural Language to Information Systems, Introduction to the Special Issue following NLDB'99", Métais E., Mayr H. dans la revue DKE, vol. 35, n°2, 2000.

[Michelet 1988] B. Michelet. L'analyse des associations. Thèse de doctorat, Université de Paris VII, UFR de Chimie, Paris, 26 Octobre 1988. Spécialité: Information Scientifique et Technique.

[Miller et al. 1990] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. Introduction to WordNet: An on-line lexical database. Journal of Lexicography, 3(4), 235- 244, 1990.

[Minoux 1983] M. Minoux. Programmation mathématique, théorie et algorithmes. Ed. Dunod, Tome 1, 1983.

[Misset 1997] Benoît Misset, Michel Naiditch, Fabienne Saulnier, Jean Philippe Fosse, Michel Pinsard, Alain Harari, Jean Pierre Blériot, Laure Comar, Bernard Garrigues, Catherine Guidon-Attali et Marie Claude Jars-Guincestre : construction d'une classification diagnostique pour le groupage médico-économique des patients de réanimation – Collection Informatique et santé, Springer Verlag France, Volume 9, 1997

[Misset 1999] B Misset, P Moine, B Garrigues, H Logerot-Lebrun, Commission d'Evaluation de la SRLF et Comité de Réanimation de la SFAR : « Recommandations pour le codage d'un patient de réanimation ». Société de Réanimation de Langue Française (SRLF) et Société Française d'Anesthésie et de Réanimation (SFAR)

[Misset 2000] Benoît Misset, Michel Pinsard, Fabienne Saulnier, Michel Naiditch : Le PMSI et les GHM français : classification alternative pour les séjours de réanimation. In F Saulnier et J Bion Editeurs : Management en réanimation : Evaluation, organisation et éthique. Elsevier, Paris, 2000 : pp 181-190.]).

[Misset 2005] Misset B, Metais E., Nakache D., Dumont S., A. De Lassence, M. Darmont, M. Garrouste Orgeas, B. Mourvillier, C. Adrie, S. Pease, M-A. Costa de Beauregard, C. Stocco, "Reproductibilité du codage", in 33ème congrès de la SRLF (Société de Réanimation de Langue Française), du 19 au 21 janvier 2005, Cnit Paris.

[Mitchell 1996] T. M. Mitchell: "Machine Learning", 1996, chapter 3. Mc Graw Hill.

[Mitchell 1997] T. M. Mitchell. Machine Learning. McGraw-Hill, New York, 1997.

[Moine et al. 2002] Moine P, Timsit JF, de Lassence A, Troche G, Fosse JP, Alberti C, Cohen Y, for the outcomerea study group – Mortality associated with late-onset pneumonia in ICU: Result of a multi-center cohort study – Intensive Care Med 2002;28:154-163

[Monari 1999] G. Monari. Sélection de modèles non linéaires par leave-one-out. Etude théorique et application des réseaux de neurones au procédé de soudage par points. Thèse de l'université Paris VI, 1999.

[Morin 1999] E. Morin : "Extraction de liens sémantiques entre termes à partir de corpus de textes techniques". Thèse de doctorat, Université de Nantes, Faculté des sciences et des techniques. 8 décembre 1999

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

[Muc7 1999] M. Kaufmann: "Proceedings of the Seventh Message Understanding Conference". 1999.

[Muslea 1998] I. Muslea, S. Minton, and C. Knoblock. Wrapper induction for semistructured, web-based information sources. In Proceedings of the Conference on Automatic Learning and Discovery CONALD-98, 1998.

[Nahm 2000] U. Y. Nahm and R. J. Mooney. Ua mutually beneficial integration of data mining and information extraction. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00), 2000.

[Nakache 2005a] Didier Nakache, Elisabeth Métais: "Indicators and methodology for evaluation" In KES 05, Melbourne, septembre 2005.

[Nakache 2005b] D. Nakache, E. Métais, J.F. Timsit : "Evaluation and NLP", in proceedings of DEXA 2005, database and expert system application, Copenhagen - Denmark august 2005, LNCS 3588, pages 626-632

[Nakache 2005c] D. Nakache, E. Métais, J.F. Timsit : "Evaluation : nouvelle approche avec juges" - XXIIIème congrès INFORSID, Grenoble, mai 2005.

[Namer 2000] F. Namer : "Flemm : Un analyseur Flexionnel du Français : base de règles", Traitement automatique des langues pour la recherche d'information, numéro spécial de la revue T.A.L, Paris (2000).

[Namer 2005] F. Namer : "Automatiser les définitions des termes médicaux : qu'est ce que le traitement automatique du langage apporte à la théorie morphologique ?", in JFIM 2005, Journées Francophones d'Informatique Médicale, Lille 12-13 mai 2005.

[Neal 1992] R. M. Neal. Bayesian Training of Backpropagation Networks by the Hybrid Monte Carlo Method. Technical Report CRG-TR-92-1, Department of Computer Science, University of Toronto, 1992.

[Neal 1994] R. M. Neal. Bayesian Learning for Neural Networks. Ph.D. thesis, University of Toronto, 1994.

[Neal 1996] R. M. Neal. Bayesian Learning for Neural Networks. New York : Springer- Verlag, 1996. Bibliographie 240

[Nesterov 1997] S. Nestorov, S. Abiteboul, and R. Motwani. Infering structure in semistructured data. SIGMOD Record, 26(4), 1997

[Ng et al. 2000] H. T. Ng, H. T. Ang, W. M. Soon. DSO at TREC-8: A Hybrid Algorithm for the Routing Task. Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 267-274, 2000.

[Nigam 1999] K. Nigam, A. K. McCallum, S. Thrun : "text classification from labeled and unlabeled documents using EM", 1999

[Nigam 1999b] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61-67, 1999.

[Nouira 1998] Nouira, S, Belghitti, M, Elastrous, S, et al (1998) Predictive value of severity scoring systems: a comparison of four models in Tunisian adult intensive care units. Crit Care Med 26,852-859

[Null 2003] Null G., Dean C., Feldman M., and Rasio D., Death by Medecine-2, http://www.garynull.com/documents/iatrogenic/deathbymedicine/deathbymedicine2.htm, Nov. 2003

[Oard et Wang 2000] D. W. Oard, J. Wang. TREC-8 Experiments at Maryland: CLIR, QA and Routing. Proceedings of the Eigth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 623-636, 2000.

[Ohman 1988] Ohmann, C., Yang, Q., Künneke, M., Stöltzing, H., Thon, K., and Lorenz, W.: "Bayes theorem and conditional dependance of symptoms: different models applied to data of upper gastrointestinal bleeding. Methods of information in medecine, 27, 73-83, 1988

[Pacak 1983] Pacak MG, Norton LM, Dunham GS. Morphosemantic Analysis of Compound Words Forms Denoting Surgical Procedures. Meth Inform Med 1983; 22: 99-105.

[Paice 1996] C. Paice: "Method for evaluation of stemming algorithms based on error counting", Journal of the American Society for Information Science, 47 (8), pp. 632-349 (1996).

[Parmentier 1998] F. Parmentier : Spécification d'une architecture émergente fondée sur le raisonnement par analogie : Application aux références bibliographiques - Thèse présentée et soutenue publiquement le 9 juin 1998 pour l'obtention du Doctorat en informatique de l'un

[Parry 2004] Parry D. "A fuzzy ontology for medical document retrieval" 2004, in proceedings of The Australasian Workshop on Data Mining and Web Intelligence (DMWI2004).

[Pertomed 2003] http://www.spim.jussieu.fr/article.php3?id_article=114

[Petard 2001] X. Petard : « Système collecticiels et médiation, Outillage du travail collaboratif par augmentation des communications »Stage de DEA (2001) de sciences cognitives au LIMSI-CNRS, Université de Paris-Sud XI, Orsay

[Petitpierre 1994] D. Petitpierre et G. Russel, G. (1994). ìMmorph - The Multext Morphology Program. Technical Report, ISSCO. http://issco-www.unige.ch/tools/

[Poibeau 2005] Thierry Poibeau : Sur le statut référentiel des entités nommées, in proceedings of Recital 2005, juin 2005, Dourdan, France

[Pol 2003] http://www.didier-pol.net/6RACINES.htm

[Porter 1980] M. F. Porter. An algorithm for suffix stripping. Program 14, 3, 130-1367, 1980.

[Pouliquen 2002] Bruno POULIQUEN : NOMINDEX : thèse pour obtenir le grade de Docteur de l'université Rennes I, Présentée et soutenue publiquement le 7 juin 2002 : Indexation de textes médicaux par extraction de concepts, et ses utilisations. Directeur de thèse : Pr. Pierre Le Beux.

[Pouliquen 2004] http://www.med.univ-rennes1.fr/~poulique/prefixes.html

[Press et al. 1992] W. H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery. Numerical Recipes in C : The Art of Scientific Computing. second Edition, Cambridge university Press, 1992.

[Quinlan 1996] J. R. Quinlan. Bagging, boosting, and C4.5. Proceedings of the Thirteenth National Conference on Artificial Intelligence, 725-730, 1996.

[Rajman 1998] M. Rajman, L. Lebart : « similarités pour données textuelles », 1998

[Rector 1992] AL Rector, WA Nowlan, et S. Kay Conceptual knowledge: the core of medical information systems. In: Lun KC, Degoulet P, Piemme T, et Rienhoff O, eds, Proc MEDINFO 92, Geneva. North Holland, 1992:1420-6.

[Rector 1997] AL Rector, S. Bechhover, CA Goble, et al. The GRAIL concept modelling language for medical terminology. Artif Intell Med 1997;9(2):139-71.

[Réhel 2005] Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés. Réhel, Simon. Thèse. Maître ès sciences (M.Sc.). Université Laval http://www.theses.ulaval.ca/2005/22376/22376.html

[Rennie 2001] Jason Rennie and Ryan Rifkin. Improving Multiclass Text Classification with the Support Vector Machine. Massachusetts Institute of Technology. AI Memo AIM-2001-026. 2001

[Rialle 2001] Rialle : Traitement du langage naturel et des informations textuelles en médecine : des premiers pas jusque XML – Faculté de médecine et CHU de Grenoble – Laboratoire TIMC-IMAG – 2001

[Richard et Lippman 1991] M. D. Richard, R. P. Lippmann. Neural Network classifiers estimate Bayesian a-posteriori probabilities. Neural Computation, 3(4), 461-483, 1991.

[Robertson et Sparck Jones 1976] S. Robertson, K. Sparck Jones. Relevance weighting of search terms. Journal of American Society for Information Science, 27(3), 129-146, 1976.

[Robertson et Walker 2000] S. Robertson, S. Walker. OKAPI/Keenbow at TREC-8. Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 151-162, 2000.

[Rocchio 1971] J.J. Rocchio. The SMART Retrieval System: Experiments in Automatic Document Processing, chapter 14, Relevance Feedback in Information Retrieval, p. 313–323. Gerard Salton (editor), Prentice-Hall Inc. : New Jersey, 1971

[Rocchio 1971b] J.J. Rocchio. Relevance Feedback in Information Retrieval. The SMART Information Retrieval System. G. Salton, Ed. Prentice Hall, Englewood Cliffs, NJ, 313-323, 1971. Bibliographie 241

[Roddick 2003] Roddick, J.F., Fule, P., Graco, W., "Exploratory Medical Knowledge Discovery, Experiences and Issues", in proceedings of SIG KDD 2003.

[Roddick 2003b] Roddick J.F., Hornsby K., de Vries D., "A Unifying Semantic Distance Model for Determining the Similarity of Attribute Values", Proceedings of the Twenty-Sixth Australasian Computer Science Conference (ACSC2003), Adelaide, Australia, 2003

[Romesburg 1990] HC Romesburg : "Cluster Analysis for Researchers", 1990, chapitre 10

[Rozenbaum 2002] L. Rozenbaum Qualité du recueil de données dans le S.I.H. - 6ième journée Nationale Phast 7 novembre 2002 – Paris - C.A.S.H de Nanterre.

[Rumelhart et al. 1986] D.E. Rumelhart, G.E. Hinton, R.J. Williams. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, MIT Press, Cambridge MA, 318-362, 1986.

[Russel 1921] B. Russel. The Analysis of Mind, London: George Allen & Unwin, 1921.

[Safran 1989] Safran C, Porter D, Lightfoot J, et al. ClinQuery: a system for online searching of data in a teaching hospital. Ann Intern Med - 1989;111(9):751-6.

[Sager 1995] Sager N., Lyman M., Nhan NT, Tick L., Medical language processing: applications to patient data representation and automatic encoding. Meth. Inform. Med. 1995.

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

[Sahami et al. 1998] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz. A Bayesian approach to filtering junk e-mail. AAAI 1998 Workshop on Text Categorization, 1998.

[Salton 1975] G. Salton, A. Wong, C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18, 613-620, 1975.

[Salton 1983] G. Salton et M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY. 1983

[Salton 1989] G. Salton. Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, Reading, USA, 1989.

[Salton et Buckley 1988] G. Salton, C. Buckley - Term Weighting Approaches, in Automatic Text Retrieval, Information Processing and Management, 24:5, 513-523. 1988

[Salton et Buckley 1990] G. Salton, C. Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), 288-287, 1990.

[Sandu Popa 2007] Iulian Sandu Popa, Karine Zeitouni, Georges Gardarin, Didier Nakache, Elisabeth Métais. "Text Categorization for Multi-label Documents and many Categories" - In CBMS2007, Maribor, Slovenia. 2007.

[Santoli 2003] F. Santoli, T. Maillard, P. Aegerter, A. Boumendil, G. Offenstadt, B. Guidet : qualité du codage en réanimation : évaluation sur 26 services et 258 séjours - XXXI ème Congrès 2003 de la SRLF- Paris 15 au 17 janvier 2003;

[Saporta 1990] G. Saporta. Probabilités, analyse des données et statistiques. Editions Technip, 1990.

[Satoshi 2003] Satoshi Suzuki : "Probabilistic Word Vector and Similarity based on Dictionaries" - http://www.kecl.ntt.co.jp/icl/mtg/members/satoshi/CICLing2003.pdf

[Schapire 1990] R. E. Schapire. The strength of weak learnability. Machine Learning, N° 5, Vol. 2, pp. 197--227, 1990.

[Schapire 1997] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee. Boosting the Margin: A new explanation for the effectiveness of voting methods. Machine Learning: Proceedings of the Fourteenth International Conference, Morgan Kauffmann, pp. 322--330, 1997.

[Schapire 1998] R. E. Schapire, Y. Singer, A. Singhal. Boosting and Rocchio Applied to Text Filtering. Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR '98), 215-223, 1998.

[Schapire 2000] R. E. Schapire, Y. Singer. BoosTexter: A boosting-based system for text classification. Machine Learning, Vol. 39, 2000.

[Schapire 2002] R. E. Schapire, M. Rochery, M. Rahim, N. Gupta : Incorporating Prior Knowledge into Boosting (2002)

[Schmidt 1994] H. Schmidt. Probabilistic part-of-speech tagging using decision trees. International Conference on New Methods in Language Processing, 44-49, 1994.

[Schneider 2004] K-M. Schneider : "On Word Frequency and Negative Evidence in Naive Bayes Text Classification", in proceedings of ESTAL2004, Adavances in Natural Language Processing, Alicante, Spain, LNAI Lecture Notes in Artificial Intelligence, Springer Verlag 2004, Volume 3230, pp 474-485.

[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), 229-238, 1995.

[Schütze et Pedersen 1995] H. Schütze, J. O. Pedersen. Information Retrieval Based on Word Senses. Fourth Annual Symposium on Document Analysis and Information Retrieval, 161- 175, 1995. Bibliographie 242

[Schwartz 1978] G. Schwartz. Estimating the dimension of a model. Annals of statistics, 6(2), 461-464, 1978.

[Scott 1999] S. Scott and S. Matwin. Feature engineering for text classification. In Proceedings of the 16th International Conference on Machine Learning ICML-99, 1999.

[Sebastiani 1999] Fabrizio Sebastiani: "A Tutorial on Automated Text Categorisation". Istituto di Elaborazione dell'Informazione, 1999. Proceedings of ASAI-99, Buenos Aires.

[Sebastiani 2002] Fabrizio Sebastiani: "Machine Learning in Automated Text Categorization". ACM Computing Surveys. Vol 34, pp 1-47. ISSN 0360-0300.

[Sebastiani 2004] Fabrizio Sebastiani: “Text Categorization” - In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK. Forth-coming.

[Serradura 2002] L. Serradura, M. Slimane, N. Vincent : "Classification semi-automatique de documents Web à l'aide des Chaînes de Markov Cachées". Publication de l’équipe RFAI, Colloque Inforsid 2002, Nantes (France), juin 2002, pp. 215-228

[Seydoux 2005] Florian Seydoux, Jean Cedric Chappelier : "indexation sémantique au moyen de coupes de redondance minimale dans une ontologie". In proceedings of TALN 2005, Juin 2005, Dourdan.

[Seymore 1998] K. Seymore, S. Chen, R. Rosenfeld. Nonlinear Interpolation of Topic Models for Language Model Adaptation. In Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 1998

[Shanno 1970] D.F. Shanno. Conjugate gradient methods with inexact searches. Mathematics of operations Research 3 (3), 244-256, 1978.

[Sheridan et Ballerini 1996] P. Sheridan et J.P. Ballerini. Experiments in Multi-lingual Information Retrieval using the SPIDER System. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 96, p.58–65. 1996

[Silberztein 2004] M. Silberztein: "NooJ : an oriented object approach" in Royauté Jean, Silberztein Max Eds, (2004). INTEX pour la Linguistique et le Traitement Automatique des Langues. Proceedings of the 4th and 5th INTEX workshop, Bordeaux, May 2001, Marseille, May 2002: Presses Universitaires de Franche-Comté (400 p).

[Simard 1996] M. Simard. Automatic Restoration of Accents in French Text, Centre d'innovation en technologies de l'information, Laval, Canada 1996 : 9 pages

[Singh 1998] L. Singh, B. Chen, R. Haight, P. Scheuermann, and K. Aoki. A robust system architecture for mining semi-structured data. In Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining, 1998, pages 329-333, 1998.

[Singhal 1996] A. Singhal. Pivoted Length Normalization. Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR'96), 21-29, 1996.

[Singhal 1998] A. Singhal. AT&T at TREC-6. Proceedings of the Sixth Text REtrieval Conference (TREC-6). NIST Special Publication 500-240, 215-226, 1998.

[Sjöberg 1994] J. Sjöberg, L. Ljung. Overtraining, Regularization and Searching for Minimum in Neural Networks. Technical Report LiTH-ISY-R-1567, Department of electrical Engineering, Linkoping Univeristy, 1994.

[Smadja et Al. 1996] F. Smadja, K.R. McKeown, et V. Hatzivassiloglou: Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics, 22(1), 1:38. 1996

[Soderland 1996] S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1-3):233-272, 1996.

[Sparck Jones 1972] Sparck Jones, K. "A statistical interpretation of term specificity and its application to retrieval". J. Documen. 28, 1 (March 1972),11-20.

[Sparck Jones 1999] K. Sparck Jones. Information retrieval and artificial intelligence. Artificial Intelligence, 114, 257-281, 1999.

[Stahlhut 1990] Stahlhut RW, McCallie DP, Waterman DM and Margulies DM. A Relational Model for Clinical Objective Results In: Miller RA, ed. Proceedings of the 14th annual SCAMC. Washington, D.C. IEEE Computer Society Press, 1990; 354-358.

[Stead 2005] Stead W., Kelly B., Kolodner R., "Achievable Steps Toward Building a National Health Information Infrastructure in the United States", First published November 23, 2004 as JAMIA PrePrint; doi:10.1197/jamia.M1685, J Am Med Inform Assoc. 2005;12:113-120. DOI 10.1197/jamia.M1685.

[Steinberger 2001] R. Steinberger: Cross-lingual Keyword Assignment. Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN'2001). Jaén, Spain. 2001

[Steinberger et al. 2002] R. Steinberger, B. Pouliquen, J. Hagman: Crosslingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc. In: A. Gelbukh (ed.) Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science for the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2002). Mexico-City, Mexico, 17-23 Février 2002. Springer, Heidelberg. 2002

[Stoica et Evans 2000] E. Stoica, D. A. Evans. Dynamic Term Selection in Learning a Query from Examples. Proceedings of the Sixth Conference on Content-Based Multimedia Information (RIAO 2000), 1703-1719, 2000.

[Stoppiglia 1997] H. Stoppiglia. Méthodes statistiques de sélection de modèles neuronaux; applications financières et bancaires. Thèse de l'université Paris VI, 1997.

[Stricker 2000] Mathieu Stricker : "Réseaux de neurones pour le traitement automatique du langage : conception et réalisatiion de filtres d'informations". Thèse de Doctorat de l'Université Pierre et Marie Curie - Paris VI (Décembre 2000).

[Stricker et al. 2000a] M. Stricker, F. Vichot, F. Wolinski, G. Dreyfus. Towards the Automatic Design of Efficient Custom Filters. Reconnaissance des formes et Intellignece Artificielle (RFIA 2000), 129-137, 2000.

[Stricker et al. 2000b] M. Stricker, F. Vichot, G. Dreyfus, F. Wolinski. Two Step Feature Selection for the TREC-8 Routing. Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 425-430, 2000.

[Stricker et al. 2001] M. Stricker, F. Vichot, G. Dreyfus, F. Wolinski. Training Context- Sensitive Neural Networks With Few Relevant Examples for the TREC-9 Routing. Proceedings of the Ninth Text REtrieval Conference (TREC-9). NIST Special Publication, 2001.

[Stricker et Haré 1998] M. Stricker, S. Haré. Réseaux de neurones et analyse marketing. Rapport Technique DTA, Informatique-CDC, 1998. Bibliographie 243

[Suetens 2004] Suetens C, Savey A, Labreuw J, Helics-ICU. Towards a European surveillance of nosocomial infections in intensive care unit: The HELICS project. Intensive Care Med 2001; 27: S204

[Teres et al 1998] D. Teres, T. Higgins, J. Steingrub et al. Defining a high-performance ICU system for the 21st century : a position paper. J Intensive Care Med 1998 ; 13 : 195-205.

[Thibault 2002] B. Thibault : Normes et Standards en Informatique de Santé - Echange de données et Communication - Etat des lieux et tendances par Bernard Thibault, 2002 ; http://www.infine.org

[Thibault 2002] Normes et Standards en Informatique de Santé - Echange de données et Communication - Etat des lieux et tendances par Bernard Thibault, 2002 ; http://www.infine.org

[Thodberg 1996] H. H. Thodberg. A Review of Bayesian Networks with an Application to Near Infrared Spectroscopy. IEE Transactions of Neural Networks, 7(1), 56-72, 1996.

[Timsit 2001] J.F. Timsit, J.P. Fosse, G. Troché, A. De Lassence, C. Alberti, M. Garrouste-Orgeas, E. Azoulay, S. Chevret, P. Moine, Y. Cohen - Accuracy of a composite score using daily SAPSII and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 hours- Intens Care med 2001; 27:1012-1021

[Timsit et al. 2002] J-F. Timsit, J-Ph. Fosse, G. Troche, A. De Lassence, C. Alberti, M. Garrouste Orgeas, C. Bornstain, C. Adrie, C. Cheval, S. Chevret for the Outcomerea study group – Calibration and discrimination by daily LOD scoring comparatively to daily SOFA scoring fo

[Timsit et al. 2002] Timsit JF, Fosse JPh, Troche G, De lassence A, Alberti C, Garrouste Orgeas M, Bornstain C, Adrie C, Cheval C, Chevret S for the Outcomerea study group – Calibration and discrimination by daily LOD scoring comparatively to daily SOFA scoring for predicting hospital mortality in critically ill patients – Crit Care Medicine 2002; 30:2003-2013

[Titterington 1987] D. Titterington. An alternative stochastic supervisor in discriminant analysis. Pattern Recognition, Vol. 22, N° 1, pp. 91--95, 1987.

[Todd 1994] Todd, B.S. and Stamper, R. "The relative accuracy of a variety of medical diagnostic programemes. In Methods of information in Medecine, 33, 402-416, 1994

[Toivonen 1999] H. Toivonen. On knowledge discovery in graph- structured data. In Workshop on Knowledge Discovery from Advanced Databases (KDAD'99), pages 26-31, 1999.

[Turenne 2001] Nicolas Turenne : "Etat de l'art de la classification automatique pour l'acquisition de connaissances à partir de textes." UMR INRA-INAPG – Biométrie et Intelligence Artificielle (BIA). INRIA, Technical report. 2001.

[Unicode 1997] The Unicode Consortium. The Unicode Standard, Version 2.0. Addison-Wesley. 1997

[Van Rijsbergen 79] K. Van Rijsbergen, "Information Retrieval", (2nd Ed.) Butterworths, London. www.dcs.gla.ac.uk/Keith/Preface.html.

[Vapnik 1982] V. N. Vapnik. Learning Dependencies based on Empirical Data. Springer, New York, 1982.

[Vapnik 1995] V.N. Vapnik. The Nature of Statistical Learning Theory. Springer, 1995

[Vichot et al. 1997] F. Vichot, F. Wolinski, J. Tomeh, S. Guennou, B. Dillet, S. Aidjan. High Precision Hypertext Navigation Based on NLP Automatic Extractions. Proceedings of Hypertext Information Retrieval Multimedia (HIM'97), 161-174, 1997.

[Vichot et al. 1999] F. Vichot, F. Wolinski, H.-C. Ferri, D. Urbani. Feeding a Financial Decision Support System with Textual Information. Journal of Intelligent and Robotic Systems, 26, 157-166, 1999.

[Vincent 1996] Vincent JL, Mendoca A, Cantraine F et al: The SOFA (Sepsis organ failure assesment) score to describe organ dysfunction/ failure. Intensive Care Med 1996; 22:707-71

[Vincent 1996] Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med 1996; 22:707–710.

[Vinot 2003] Romain Vinot, Natalia Grabar, Mathieu Valette : Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet, in proceedings of TALN 2003, Batz-sur-Mer, 11–14 juin 2003

[Vleduts 1987] Vleduts-Stokolov, Natasha. Concept recognition in an automatic text processing system for the life science. Journal of the American Society for Information Science. 38:269:287. 1987

[Voorhees 1993] E. M. Voorhees. Using WordNet to disambiguate words senses for text retrieval. Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR'93), 171-180, 1993.

[Voorhees et Harman 2000] E. M. Voorhees, D. K. Harman. Overview of the eighth Text REtrieval Conference (TREC-8). Proceedings of the Eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246, 1-24, 2000. Bibliographie 244

[VuMeF] http://www.vidal.fr/vumef/

[Wang 1999] H. L. K. Wang. Discovering association of structure from semistructured objects. To appear in IEEE Transactions on Knowledge and Data Engineering, 1999.

[Webb 1999] A. Webb. Statistical Pattern Recognition. Oxford University Press. 1999.

[Weiss 1999] S. M. Weiss, C. Apte, F. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. IEEE Intelligent Systems, 14(4):63-69, 1999.

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

[Wiener 1995] W. Wiener, J. Pedersen, and A. Weigend. A neural network approach to topic spotting. In Proceedings of the 4th Symposium on Document Analysis and Information Retrieval (SDAIR 95), pages 317- 332, 1995.

[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), 317-332, 1995.

[Wiki 2006] http://fr.wikipedia.org/wiki/Classification_automatique

[Williams 1995] P.M. Williams. Bayesian regularization and pruning using a Laplace prior. Neural Computation, 7(1), 117-143, 1995.

[Wilson 2004] Wilson AM, Thabane L, Holbrook A: "Application of data mining techniques in pharmacovigilance", British Journal of Clinical Pharmacology, 2004 - Volume 57 Issue 2 Page 127 - February 2004 - doi:10.1046/j.1365-2125.2003.01968.

[Witten 1999] I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan. Text mining: A new frontier for lossless compression. In Data Compression Conference 1999, pages 198- 207, 1999.

[Wolinski et al. 1995] F. Wolinski, F. Vichot, B. Dillet. Automatic Processing of Proper Names in Texts. Proceedings of the seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL'95), 23-30, 1995.

[Wolinski et al. 1998] F. Wolinski, F. Vichot, O. Grémont. Producing NLP-based On-line Contentware. Natural Language Processing & Industrial Applications, 253-259, 1998.

[Wolinski et al. 2000] F. Wolinski, F. Vichot, M. Stricker. Using Learning-Based Filters to Detect Rule-Based Filtering Obsolescence. Proceedings of the Sixth Conference on Content- Based Multimedia Information (RIAO 2000), 1208-1220, 2000.

[Wolinski et Vichot 2001] F. Wolinski, F. Vichot. Des multi agents pour produire des applications de contenu en ligne. Technique et Science Informatique, 2001.

[Wu 2003] Wu, T. and Pottenger, W. (2003) A Semi-supervised Algorithm for Pattern Discovery in Information Extraction from Textual Data. In the Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-03). Seoul, Korea, April/May.

[Wunsch 2004] Hannah Wunsch, Anthony R. Brady and Kathy Rowan: Impact of exclusion criteria on case mix, outcome, and length of stay for the severity of disease scoring methods in common use in critical care, in Journal of critical care, Volume 19, Issue 2, Pages 67-74 (June 2004)

[Xminer] http://www.francetech.gouv.fr/rntl/AAP2001/Fiches_Resume/XMINER.htm

[Yamada 2003] Yamada Y., Yokoi H., Takabayashi K : "Decision-tree Induction from Time-series Data Based on a Standard-example Split Test", in proceedings of ICML 2003.

[Yang 1997] Y. Yang. An evaluation of statistical approach to text categorization. Rapport interne Technichal Report CMU-CS-97-127, Carnegie Mellon University, 1997

[Yang 1999] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14(4):32 - 43, 1999.

[Yang 2003] Yimin Yang, Jian Zhang, Bryan Kisiel : A Scalability Analysis of Classifiers in Text Categorization, 2003, in proceedings of SIGIR 2003, The 26th Annual International ACM SIGIR Conference, Toronto, Canada, July 28 to August 1, 2003.

[Yang et Chute 1994] Y. Yang, C. G. Chute. An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems, 12(3), 252-277, 1994.

[Yang et Liu 1999] Y. Yang, X. Liu. A re-examination of text categorization methods. Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR '99), 42-49, 1999.

[Yang et Perdersen 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), 412-420, 1997. Bibliographie 245

[Yarowsky 1995] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 189-196, 1995.

[Yi 2006] Kwan Yi : Les défis de la catégorisation automatique utilisant les systèmes de classification de bibliothèque, Ecole des bibliothèques et sciences de l'information Université duKentucky Etats-Unis (Traduction : Annie Milhaud), in WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL 20-24 August 2006, Seoul, Korea

[Zaïane 1998] O. Zaïane and J. Han. Webml: Querying the world wide web for resources and knowledge. In Proc. ACM CIKM'98 Workshop on Web Information and Data Management (WIDM'98), pages 9-12, 1998.

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

[Zheng 2005] Zheng G., George O., Narasihman G. : "Neural Network Classifiers and Gene Selection Methods for Microarray Data on Human Lung Adenocarcinoma", in proceeding of Methods of Microarray Data Analysis IV, 2005.

[Zipf 1949] G. K. Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley, 1949. et http://users.info.unicaen.fr/~giguet/java/zipf.html

[Zweigenbaum 1992] P. Zweigenbaum, M. Cavazza, L. Doré, J. Bouaud, D. Sedlock, Natural language processing of patient discharge summaries : extraction prototype, In : Nothoven van Goor J., Christensen J.P., (eds.) Proc. AIM 92. Amsterdam : IOS Press, 1992, pp. 215-222

[Zweigenbaum 1995] Zweigenbaum P. et coll. MENELAS, The final report, Ménélas deliverable#17, Paris, 1995.

[Zweigenbaum 1997] P. Zweigenbaum : Construction d'une représentation sémantique en graphes conceptuels – TALN 1997.

[Zweigenbaum 1999] P. Zweigenbaum : Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. Innovation Stratégique en Information de Santé, (2-3):27-47, 1999

[Zweigenbaum 1999] Zweigenbaum P. & Grabar, N.: Automatic acquisition of morphological knowledge for medical language processing, In Werner Horn, Yuval Shahar, Greger Lindberg, Steen Andreassen, & Jeremy Wyatt, Artificial Intelligence in Medicine, Lecture Notes in Artificial Intelligence. Springer-Verlag, 1999.

[Zweigenbaum 1999b] Zweigenbaum, P. & Grabar, N.: "Automatic acquisition of morphological knowledge for medical language processing", In Werner Horn, Yuval Shahar, Greger Lindberg, Steen Andreassen, & Jeremy Wyatt, Artificial Intelligence in Medicine, Lecture Notes in Artificial Intelligence. Springer-Verlag, 1999.

[Zweigenbaum 2001] P. Zweigenbaum : Projection de requêtes en langue naturelle sur les termes du MeSh : l'apport de connaissances morphologiques – IPM 2001

[Zweigenbaum 2002] P. Zweigenbaum et Natalia Grabar : « Accentuation de mots inconnus : application au thesaurus biomédical MeSH ». TALN 2002, Nancy, 24 au 27 juin 2002.

[Zweigenbaum 2002b] P. Zweigenbaum and N. Grabar: "Accenting unknown words: application to the French version of the MESH", in Workshop NLP in Biomedical Applications, pages 69-74, Cyprus, March. EFMI.

TABLES DES ILLUSTRATIONS

8.1 Table des figures

8.2 Table des équations

ANNEXES

9.1 Annexe 1 : le code Source pour EDA

```
Public Function EDA(ByVal Mot As String)
Dim Xold, Xm, Xcar As String
Dim ii, jj As Integer
Xold = ""
Xm = ""
Mot = Trim(LCase(Mot))

' phase 1 du traitement
' Suppression des accents et caractères diacritiques ou ligaturés

' Si le mot fait moins de 5 lettres, arrêt de la procédure
If Len(Mot) <= 5 Then
  If Right(Mot, 1) = "e" Then
    Mot = Left(Mot, Len(Mot) - 1)
  End If
  If Right(Mot, 1) = "s" Then
    Mot = Left(Mot, Len(Mot) - 1)
  End If
  Eda = Mot
  Exit Function
End If

' Remplacements phonétiques
Mot = Replace(Mot, "cqu", "k")
Mot = Replace(Mot, "ck", "k")
Mot = Replace(Mot, "qu", "k")
Mot = Replace(Mot, "c", "k")
Mot = Replace(Mot, "y", "i")

' Liste des règles
Dim T(1 To 37) As String
T(1) = "s"
T(2) = "e"
T(3) = "x"
T(4) = "ant"
T(5) = "al"
T(6) = "au"
T(7) = "tion"
T(8) = "sion"
T(9) = "er"
T(10) = "iv"
T(11) = "if"
T(12) = "abl"
T(13) = "ibl"
```

```
T(14) = "ment"
T(15) = "tele"
T(16) = "tel"
T(17) = "tos"
T(18) = "ik"
T(19) = "ton"
T(20) = "tos"
T(21) = "ent"
T(22) = "en"
T(23) = "tik"
T(24) = "toid"
T(25) = "o"
T(26) = "i"
T(27) = "s"
T(28) = "dien"
T(29) = "u"
T(30) = "e"
T(31) = "era"
T(32) = "a"
T(33) = "ank"
T(34) = "enk"
T(35) = "teur"
T(36) = "trice"
T(37) = "i"

For ii = 1 To 37
  jj = Len(T(ii))
  If Right(Mot, jj) = T(ii) Then
    Mot = Left(Mot, Len(Mot) - jj)
    If Len(Mot) <= 5 Then
      If Right(Mot, 1) = "e" Then
        Mot = Left(Mot, Len(Mot) - 1)
      End If
      If Right(Mot, 1) = "s" Then
        Mot = Left(Mot, Len(Mot) - 1)
      End If
      Eda = Mot
      Exit Function
    End If
  End If
Next ii

Eda = Mot
End Function
```

9.2 Annexe 2 : arborescence et nouveautés du MeSH 2007

Source [MeSH 2007]

[A] - Anatomie
- [A01] - régions du corps
- [A02] - appareil locomoteur
- [A03] - système digestif
- [A04] - appareil respiratoire
- [A05] - appareil urogénital
- [A06] - système endocrine
- [A07] - système cardiovasculaire
- [A08] - système nerveux
- [A09] - organes des sens
- [A10] - tissus
- [A11] - cellules
- [A12] - liquides et sécrétions biologiques
- [A13] - anatomie animale
- [A14] - système stomatognathique
- [A15] - systèmes sanguin et immunitaire
- [A16] - structures embryonnaires
- [A17] - système tégumentaire

[B] - Organismes (invertébrés, vertébrés, bactéries, virus, algues et champignons, plantes, archéobactéries)
- [B01] - animaux
- [B02] - algues
- [B03] - bactéries
- [B04] - virus
- [B05] - champignons
- [B06] - plantes
- [B07] - archéobactéries
- [B08] - mésomycétozoaires

[C] - Maladies
- [C01] - infections bactériennes et mycoses
- [C02] - maladies virales
- [C03] - maladies parasitaires
- [C04] - tumeurs
- [C05] - maladies de l'appareil locomoteur
- [C06] - maladie de l'appareil digestif
- [C07] - maladies du système stomatognathique
- [C08] - maladies de l'appareil respiratoire
- [C09] - maladies oto-rhino-laryngologiques
- [C10] - maladies du système nerveux
- [C11] - maladies de l'oeil
- [C12] - maladies urogénitales de l'homme
- [C13] - maladies de l'appareil urogénital féminin et complications de la grossesse
- [C14] - maladies cardiovasculaires

[C15] - hémopathies et maladies lymphatiques
[C16] - malformations et maladies congénitales, héréditaires et néonatales
[C17] - maladies de la peau et du tissu conjonctif
[C18] - maladies métaboliques et nutritionnelles
[C19] - maladies endocriniennes
[C20] - maladies du système immunitaire
[C21] - troubles liés à l'environnement
[C22] - maladies animales
[C23] - signes et symptômes, états pathologiques
[D] - Produits chimiques, biologiques et pharmaceutiques
[D01] - produits chimiques inorganiques
[D02] - composés chimiques organiques
[D03] - composés hétérocycliques
[D04] - composés polycycliques
[D05] - structures macromoléculaires
[D06] - hormones, substituts d'hormones, et antagonistes d'hormones
[D08] - enzymes et coenzymes
[D09] - glucides
[D10] - lipides
[D12] - protéines, peptides et acides aminés
[D13] - nucléosides et nucléotides
[D20] - mélanges complexes
[D23] - facteurs biologiques
[D25] - matériaux biomédicaux et dentaires
[D26] - préparations pharmaceutiques
[D27] - actions chimiques et utilisations
[E] - Équipements et techniques analytiques, diagnostiques et thérapeutiques
[E01] - diagnostic
[E02] - thérapeutique
[E03] - anesthésie et analgésie
[E04] - interventions chirurgicales
[E05] - techniques d'investigation
[E06] - dentisterie
[E07] - équipement et appareillage
[F] - Psychiatrie et psychologie
[F01] - comportement et mécanismes comportementaux
[F02] - processus et phénomènes psychologiques
[F03] - troubles mentaux
[F04] - disciplines et activités comportementales
[G] - Sciences biologiques
[G01] - sciences biologiques
[G02] - profession santé
[G03] - environnement et santé publique
[G04] - phénomène biologique, phénomène cellulaire, et immunité
[G05] - processus génétiques
[G06] - phénomènes biochimiques, métabolisme, nutrition
[G07] - processus physiologique
[G08] - physiologie appareil reproducteur et urinaire
[G09] - physiologie respiration et circulation
[G10] - physiologie digestive, buccale, cutanée

[G11] - physiologie musculaire et appareil locomoteur, neurologique et oculaire
[G12] - phénomènes chimiques et pharmacologiques
[G13] - phénomènes génétiques
[G14] - structures génétiques
[H] - Sciences physiques
[H01] - sciences naturelles
[I] - Anthropologie
[I01] - sciences sociales
[I02] - éducation
[I03] - activités humaines
[J] - Technologie aliments et boissons
[J01] - technologie, industrie, agriculture
[J02] - aliments et boissons
[K] - Arts et sciences humaines
[K01] - arts et sciences humaines
[L] - Sciences de l'information
[L01] - sciences de l'information
[M] - Individus
[M01] - individus
[N] - Santé (administration des soins)
[N01] - caractéristiques de la population
[N02] - infrastructures, main d'oeuvre et services de soins
[N03] - aspects économiques et organisationnels des soins
[N04] - administration services de soins
[N05] - qualité, accès, évaluation soins
[Z] - Emplacements géographiques
[Z01] - emplacements géographiques

Le MeSH 2007 comprend :

- 24 357 descripteurs
- 83 qualificatifs (Subheadings/Qualifiers)
- 164 331 concepts chimiques supplémentaires (SCRs)

avec

- 494 nouveaux descripteurs
- 99 descripteurs modifiés
- 22 descripteurs supprimés

Les 494 nouveaux descripteurs classés par catégorie :

Anatomie [A]	23
Organismes [B]	61
Maladies [C]	70
Produits chimiques et pharmaceutiques [D]	255
Équipements et techniques analytiques, diagnostiques et thérapeutiques [E]	35
Psychiatrie et Psychologie [F]	6
Sciences biologiques [G]	53
Sciences naturelles [H]	7
Anthropologie, enseignement, sociologie et phénomènes sociaux [I]	3
Technologie, industrie et agriculture [J]	7
Sciences de l'information [L]	1
Individus [M]	1
Santé [N]	1

* Certains descripteurs apparaissant dans plusieurs catégories, leur somme est supérieure à 494.

Organismes [B]
31 nouveaux descripteurs de virus ont été ajoutés et 27 ont été modifiés. De nombreux descripteurs qui concernent les virus à ADN spécifiques ont été ajoutés.

Produits chimiques et pharmaceutiques [D]
Environ la moitié des descripteurs 2007 appartiennent à la catégorie "Produits chimiques et pharmaceutiques" [D]. Parmi ces derniers, 136 appartiennent à la sous-catégorie "Protéines, peptides et acides aminés" [D12]. De nombreux descripteurs concernent les apolipoprotéines, les récepteurs de l'interleukine et des kératines spécifiques.

Types de publication et catégorie V du MeSH
Avec la mise à jour 2007, tous les termes de la catégorie V seront identifiés comme types de publication (PT) dans les notices MEDLINE et les références du catalogue de la NLM.
Auparavant, les termes qui vont suivre étaient identifiés en tant que termes MeSH (MH) dans les notices MEDLINE et les références du catalogue de la NLM :

1. Les termes spécifiques de V04 utilisés pour l'indexation des notices de journaux :
   Support of Research [V04]
   Research Support, Non-U.S. Gov't/ [V04.124]
   Research Support, U.S. Gov't, Non-P.H.S. [V04.249]
   Research Support, U.S. Gov't, P.H.S. [V04.500]
   Research Support, N.I.H., Extramural [V04.500.500]
   Research Support, N.I.H., Intramural [V04.500.750]
2. In Vitro/ In vitro [V03.500]
3. Étude comparative [V03.250]
4. English Abstract/ Résumé anglais [V01.260]

Ces modifications permettront aux notices PubMed et aux références du catalogue d'être en synchronisation avec les termes "types de publications" du MeSH.

En effet, les termes de la catégorie V ne sont pas des descripteurs définissant le sujet de l'article, ils indiquent que l'article suit les caractéristiques de publication concernée par le concept. Ainsi, une notice avec le concept "Comparative Study/ Étude comparative" signifie que cet article présente une étude comparative et non une discussion sur les études comparatives. De même, l'utilisation des termes en "Research support/Subvention" signifie que l'article présente une recherche qui a reçu une subvention et non une discussion sur le rôle du financement dans la recherche. Quant à "English abstract", son emploi signifie que l'article concerné est de langue non anglaise et qu'un résumé anglais est disponible en ligne, ou bien pour les notices plus anciennes qu'on le retrouvera ce résumé anglais dans le document primaire. Enfin, le concept "In vitro" est utilisé quand les processus physiologiques sont étudiés dans des tissus ou organes extraits d'un organisme supérieur (Humains, animal ou plantes).

La recherche dans PubMed prend en compte le passage de ces termes du champ MH au champ PT.

9.3 Annexe 3 : calculs de scores utilisés en réanimation

Calcul du score Omega

Les actes de réanimation sont enregistrés différemment selon leur catégorie:
1 - Catégorie 1
Les actes de cette catégorie sont enregistrés s'ils sont effectués au moins une fois au cours du séjour dans l'unité de réanimation. Ils ne sont enregistrés qu'une seule fois par séjour s'ils sont effectués plusieurs fois.

2 - Catégorie 2
Les actes de cette catégorie sont enregistrés chaque fois qu'ils sont effectues. Les points Oméga correspondants sont donc comptabilisés autant de fois que l'acte est effectué au cours d'un même séjour chez un même malade.

3 - Catégorie 3
Les actes de cette catégorie sont enregistrés chaque jour de leur réalisation. Les points Oméga correspondants s'obtiennent donc en multipliant la valeur de l'acte par sa durée de réalisation en jours.
La valeur du score d'activité Oméga est égale à la somme en points Oméga des actes des catégories 1, catégorie 2, et catégorie 3, recueillis sur la totalité du séjour du patient.

Catégorie 1 : Actes à relever une seule fois par séjour	
D100 Trachéotomie: mise en place ou surveillance	6
D101 Drains thoraciques/péricardiques: mise en place ou surveillance	6
D102 Entraînement à la ventilation à domicile	6
D103 Cathéter central ou Désilet : mise en place ou surveillance (pour le nouveau-né: cathéter Jonathan, Broviak, veineux, ombilical)	3
D104 Cathéter artériel pulmonaire: mise en place ou surveillance	6
D105 Cathéter artériel: mise en place ou surveillance (pour le nouveau-né: cathéter ombilical, radial, temporal)	3
D106 Sonde d'entraînement électrosystolique: mise en place ou surveillance	3
D107 Intubation: mise en place ou surveillance	6
D108 Contrepulsion aortique: mise en place ou surveillance	10
D109 Cardioversion	3
D110 Traitement d'un arrêt circulatoire	10
D111 Utilisation de drogues vasoactives	6

D112 Utilisation de fibrinolytiques	10
D113 Perfusion de dérivés sanguins, volume supérieur à 1/2 masse sanguine en 24H (pour le nouveau-né: volume supérieur à 40ml/kg)	10
D114 Lavage gastrique	1
D117 Alimentation parentérale: 35 calories/kg/jour pendant au moins 10 jours (pour le nouveau-né: 20 calories/kg/jour)	6
D118 Alimentation entérale: 35 calories/kg/jour pendant au moins 10 jours, à débit constant	3
D119 Réinjection d'ascite	10
D120 Tamponnement de varices oesophagiennes: mise en place ou surveillance	3
D121 Shunt artério-veineux: mise en place ou surveillance	10
D122 Sonde urétérale: mise en place ou surveillance	3
D123 Cathéter sus-pubien: mise en place ou surveillance	1
D124 Traction orthopédique complexe	6
D125 Bilans neurologiques rapprochés	1
D126 Drainage de liquide céphalo-rachidien	1
D127 Monitorage de la pression intracrânienne	3
D128 Sédation de plus de 24 heures	6
D130 Ponction-lavage péritonéale	3

Catégorie 2 : Actes à relever à chaque réalisation	
D150 Épuration extrarénale séquentielle ou épuration extra-corporelle séquentielle	10
D151 Plasmaphérèse (pour le nouveau-né: exsanguino-transfusion)	10
D152 Endoscopie bronchique, y compris lavage alvéolaire, dans le service	3
D153 Endoscopie digestive dans le service	3
D154 Oxygénothérapie hyperbare dans le service	10
D155 Préparation et accompagnement de transport hors du service de réanimation (exclu: transport par le S.M.U.R., transport pour intervention chirurgicale)	3
D156 Préparation de transport par le S.M.U.R. (exclu : transport pour intervention chirurgicale)	1

D157 Échographie dans le service	3
D158 Scintigraphie dans le service	6
D159 Angiographie dans le service	10
D160 Préparation, accompagnement ou accueil au retour de bloc opératoire (exclu: transport par le S.M.U.R.)	6

Catégorie 3 : Actes à relever chaque jour	
D170 Ventilation spontanée avec P.E.P. et C.P.A.P	10
D171 Ventilation mécanique	10
D172 Dialyse péritonéale continue ou hémofiltration continue	10
D115 Pansement chirurgical complexe	6
D116 Réinstillation digestive	6
D129 Isolement d'un malade en chambre stérile ou isolateur	10
D177 Enfant sous incubateur	1
D179 Surveillance continue de réanimation	4

Charge en soins (NEMS)

Ce score (appelé aussi TISS simplifie) permet de calculer très facilement la charge en soins journalière

Monitorage de base *signes vitaux horaires, bilans entrées-sorties*	9
Médicaments intraveineux (sauf drogues vaso-actives) *administrés en bolus ou en continu*	6
Ventilation mécanique *toute forme de ventilation assistée ou mécanique*	12
Autre prise en charge ventilatoire *VS sur tube ou toute autre méthode d'oxygénation sauf mécanique*	3
Une seule drogue vaso-active *quelque soit la drogue*	7
Plusieurs drogues vaso-actives *plus d'une drogue vaso-active, quelque soit la drogue et la dose*	12
épuration extra rénale *toutes les techniques*	6
Interventions spécifiques à l'intérieur de la réanimation *e.g. intubation endo-trachéale,* *pose d'une sonde d'entraînement electrosystolique,* *cardioversion, endoscopies,* *chirurgie urgente dans les 24 dernières heures,* *lavage gastrique...*	5
Interventions spécifiques en dehors de la réanimation *e.g. interventions chirurgicales, procédures diagnostiques*	6
Total NEMS	

CALCUL DE SCORE SAPSII ET DE PROBABILITE DE DECES

VARIABLE	26	13	12	11	9	7	6	5	4	3	2	0	1	2	3	4	6	7	8	9	10	12	15	16	17	18
Age (year)												<40						40–59				60–69	70–74	75–79		≥80
Heart Rate (Beat/mn)				<40							40–69	70 – 119				120–159		≥160								
Syst Blood Press		<70						70–99				100 – 199		≥200												
Body temp (°C)												<39°			≥ 39°											
PaO2/FiO2 (mmHg) VENT or CPAP				<100	100–199		≥200																			
Urinary Out. (L/day)			<0,500						0,500–0,999			≥1,000														
Blood Urea (mMol/L)												<10,0					10,0–29,9				≥30,0					
WBC Count (10E3/ml)			<1,0									10,–19,9			≥20,0											
Serum K (mEq/L)										<3,0		3,0–4,9			≥5,0											
Serum Na (mEq/L)								<125				125–144	≥145													
Serum HCO3 (mEq/L)							<15			15–19		≥20														
Bilirubin (µmol/L)												<68,4				68,4–102,5				≥102,6						
Glasgow	<6	6–8				9–10		11–13				14–15														
Chronic disease:																				Mét.Can	Hém. Mal				AIDS	
Type of admission												Elec.					Med		S. Em.							

La probabilité de décès se calcule ainsi : Probabilité = Exp(X) / (1 + Exp(X)) , avec X = -7.7631 + (0.0737 * SAPSII) + 0.9971 * (Log(SAPSII + 1) / Log(2.718282))

CALCUL DU SCORE TRIO

To compute de probability of hospital mortality:

1. compute the logit:

   Logit= (-4.44) + 0.5543(Transfer)+0.1536(LOD)+0.0388(SAPS II)+ 0.8507(Chronic illness)+0.4161(SAPS2-SAPS3 alteration)+0.6940(LOD2-LOD3 alteration).

2. calculate the probability of hospital mortality (P(death))

$P(death)=(e^{Logit})/(1+e^{Logit})$ where e=2.7182818 (the base of the natural logarithm).

SCORE LOD

	LOD Points						
	Increasing Severity/ Decreasing Values			Organ dysfunction Free		Increasing Severity/ Increasing Values	
Organ System Measures	**5**	**3**	**1**	**0**	**1**	**3**	**5**
Neurologic							
Glasgow Coma Score	3-5	6-8	9-13	14-15	...	...	...
Cardiovascular							
Heart rate, beats/min	<30	...	...	30-139	≥ 140	...	...
	or			and	or		
Systolic Blood pressure, mm Hg	<40	40-69	70-89	90-239	240-269	≥ 270	...
Renal							
Serum urea, mmol/L (g/L)	...	...	...	<6 (<0.36)	6-9.9 (0.36-0.59)	10-19.9 (0.60-1.19)	≥ 20 (≥1.20)
or							
Serum urea nitrogen, mmol/L (mg/dL)	...	...	...	<6 (<17)	6-9.9 (17-<28)	10-19.9 (28-<56)	≥ 20 (≥ 56)
				and	or	or	
Creatinine, µmol/L (mg/dL)	...	...	...	<106 (<1.20)	106-140 (1.20-1.59)	≥ 141 (≥ 1.60)	...
				and		or	
Urine Output, L/d	< 0.5	0.5-0.74	...	0.75-9.99	...	≥ 10	...
Pulmonary							
PaO_2 (mm Hg)/FIO_2 on MV or CPAP		< 150	≥ 150	No ventilation	...	...	...
				no CPAP			
(PaO_2[kPa]/FIO_2)		(<19.9)	(≥ 19.9)	no IPAP	...	...	...
Hematologic							
White blood cell count, x 10^9/L	...	<1.0	1.0-2.4	2.5-49.9	≥ 50.0	...	...
			or	and			
Platelets, x10^9/L	...	...	<50	≥ 50			
Hepatic							
Bilirubin, µmol/L (mg/dL)	...	...	...	<34.2(<2.0)	≥ 34.2 (≥ 2.0)	...	...
				and	or		
Prothrombin time, s above standard (% of standard)	...	...	(<25 %)	≤ 3 (≥ 25 %)	>3	...	...

SCORE SOFA

SOFA (Sequential Organ Failure Assessment) Score

	0	1	2	3	4
Respiration Pao2/fio2(mmhg)	>400	<= 400	<= 300	<= 200 et ventilation mécanique	<= 100 et ventilation mécanique
Hemostase Plaquettes(Giga/L)	>150	<= 150	<= 100	<= 50	<= 20
Foie Bilirubine (mg/dL) (µmol/L)	<1.2 <20	1.2-1.9 20-32	2-5.9 33-101	6-11.9 102-204	>12 >204
Cardiovasculaire		Pression artérielle moyenne<70 mmHg	Dopa<= 5 gamma Dobu	Dopa >= 5 adrenaline <= 0.1 noradrénaline <= 0.1 gamma/kg/mn	Dopa >15 adrenaline > 0.1 noradrénaline> 0.1 gamma/kg/mn
Neurologique Glasgow	15	13-14	10-12	6-9	<6
Rein Creatinine (µmol/l) Débit urinaire	<110	110-170	171-299	300-440 ou <500ml/jour	>440 ou <200 ml/jour

Relation défaillance d'organe >= 3 le jour de l'admission et mortalité

	respiratoire	hemostase	Foie	Cardiovasculaire	Neurologique	Renal
Défaillance isolée	20.7%	16.7%	14.3%	27.9%	24%	23%
respiratoire		60.3%	59%	55.4%	48.1%	57.4%
hemostase			65.6%	69.2%	73.8%	72.3%
Foie				71.2%	67.6%	73.8%
Cardiovasculaire					64.7%	74.3%
Neurologique						66.7%