UNIVERSITÉ DE SHERBROOKE

MÉMOIRE PRÉSENTÉ AU

PROGRAMME DE MAITRISE EN ADMINISTRATION

Par

Alexandre Tardif, Candidat à la M. Sc. en Stratégie de l'intelligence d'affaires

Olivier Caya, Directeur de recherche

Jessica Lévesque, Codirectrice de recherche

Lecteur
Jean Cadieux

Exploration de textes dans un corpus francophone de droit

Le cas SOQUIJ

Le 15 mai 2014

SOMMAIRE

L'intelligence d'affaires a mis en place des processus et des procédures permettant l'accès à une donnée unique. Des rapports, des requêtes et des analyses sont possibles sur cette structure. L'exploration de données a bénéficié de ces démarches et a fait naitre l'exploration de textes.

L'exploration de textes est peu employée par rapport à l'exploration de données et ce autant par la communauté scientifique que par le domaine privé. La syntaxe et la grammaire mathématique sont universelles tandis que la syntaxe et la grammaire linguistique sont régionales et plus complexes. Ces limitations ont restreints les recherches sur l'exploration des textes..

Ce mémoire s'intéresse à l'utilisation d'un outil d'exploration de textes dans le contexte juridique. Plus précisément, l'objectif de la présente recherche est d'utiliser l'outil pour en découvrir les défis et opportunités découlant de l'exploration des liens des textes et de la classification supervisée et non supervisée. Afin d'atteindre cet objectif, la présente recherche s'appuie sur le « design science » et la méthodologie « CRISP-DM », le tout dans le but de sélectionner un outil logiciel approprié à la recherche, d'effectuer l'exploration de textes et d'analyser les résultats.

Les principaux résultats qui émanent des analyses effectuées avec l'outil IBM PASW SPSS sont les suivants. Premièrement, une analyse des liens entre les textes permet de faire ressortir les concepts des différents domaines de droit. Deuxièmement, l'analyse « Two-Steps » fait ressortir 3 classes dans le corpus complet qui comprend 4 domaines. Enfin, les analyses de classifications supervisées ont eu un taux de succès entre 46 et 60 % sur les échantillons de validation.

Les modèles développés sont peu performants et selon moi ils ne peuvent pas être déployés à la SOQUIJ. La connaissance du domaine juridique est importante afin d'analyser et interpréter les textes propres à la SOQUIJ. Il en va de même afin de

créer un dictionnaire pour l'exploration de textes. Ce dictionnaire spécifique au droit manque pour l'obtention de résultats plus probants.

Plusieurs avenues sont intéressantes pour les recherches futures. Des plus intéressantes, notons la validation de l'impact de la création d'un dictionnaire pour réviser les différentes analyses et aussi d'étudier le résultat des 3 classes créées par le « Two-Steps ».

REMERCIEMENTS

J’ai passé plus de temps que j’aurais dû sur ce mémoire et pour moi il est important de remercier ceux qui me sont chers.

Julie, je sais que ça a été long et cela t’a stressé au plus haut point, merci et je t’aime.

Mes enfants pour leurs sourires, rires et pour l’avenir, merci et je vous aime.

Mes collègues de classe Phil, Geoff, Meh, Sylvie, Francis et Simon qui tous à un moment donné m’ont aidé sans s’en rendre compte, merci.

À mon beau-frère Stéphane, qui à lui seul m’a posé les meilleures questions, merci.

Aux autres membres de ma famille, auxquels quand je parle de ce que j’effectue ne comprennent rien du tout, mais qui quand même m’ont écouté, merci.

À l’amie de ma conjointe, Geneviève, pour m’avoir ouvert le chemin à la SOQUIJ, merci.

À M. Champagne et à la SOQUIJ pour l’intérêt envers l’avancée de la science, merci.

À mon collègue François qui m’a poussé à terminer, merci.

À mes professeurs Manon et Daniel, mon directeur Olivier et ma codirectrice Jessica pour votre dévouement, vos conseils et votre patience, merci.

À la vie et ses embuches et bien j’ai passé les plus belles années auprès de mes enfants, merci!

Table des matières

Liste des tables

Liste de figures

1 Introduction

Je suis né à l'ère analogique. Je regardais une télévision qui avait une antenne. Je devais me lever du divan pour changer de poste sur le téléviseur. J'écoutais la musique à la radio, sur une cassette ou bien sur mon tourne-disque. Je lisais uniquement sur des supports en papier.

La personne que je suis aujourd'hui a toujours la télévision, mais par satellite. J'ai maintenant 5 manettes pour contrôler différentes choses. J'écoute la musique satellite, sur un lecteur MP3 et maintenant rarement sur CD. Je lis des articles sur Internet, je consulte des fichiers PDF sur mon portable. Je suis encore « vieux jeu » et je lis encore des livres en papier. Je vis maintenant dans une ère numérique.

Cette ère numérique amène une nouvelle approche à l'information autant par les médiums, par les médias et par leurs utilisations. Les entreprises n'échappent pas à cette vague. Les dirigeants autrefois regardaient les chiffres de fin de mois sur un état financier préparé par le comptable doivent s'adapter au changement, tout comme moi. Les chiffres ne sont plus le seul élément factuel à la prise de décision. Le gestionnaire d'aujourd'hui doit aussi consulter les autres ressources dénommées numériques qui contiennent des vidéos (du son et des images), des images, des enregistrements audio ainsi que du texte. L'actif numérique est en augmentation, soit par le volume, la taille, le nombre et il devient difficile de tout lire, tout voir et de tout comprendre. En 2009,

selon Barker et al. (2009), 99% des documents produits dans une entreprise le sont de manière numérique.

J'ai travaillé dans une entreprise axée sur l'analyse des chiffres. Je constate l'explosion des nouvelles informations numériques disponibles aux gestionnaires. Je veux tout d'abord voir et expérimenter ce que l'exploration de textes peut faire dans un milieu d'affaires et par la suite, m'interroger sur l'implantation de ces pratiques et leurs bénéfices potentiels pour la prise de décision.

L'intelligence d'affaires (IA) se définit par l'ensemble de l'architecture technologique, les outils, les applications et les méthodologies pour la prise de décision (Turban (2007)). Foley & Guillemette (2010) eux définissent l'IA comme une combinaison de processus, de politique, de culture et de technologies pour la collecte, la manipulation, le stockage et l'analyse des données recueillies auprès de sources internes et externes, afin de communiquer l'information au bon moment, au bon endroit, et sous la bonne forme, de créer des connaissances et de supporter la prise de décisions. Ainsi, l'IA aide le gestionnaire à prendre une décision éclairée au bon moment.

Le « *datamining* » et le « *textmining* » se traduisent selon « *Le grand dictionnaire terminologique* » de l'Office québécois de la langue française comme étant l'exploration de données (DM) et l'exploration de textes (TM). Rexer Analytics (2011) sonde les tendances du marché auprès des utilisateurs d'exploration. Selon

cette firme, un des grands défis qui revient d'année en année est que les explorateurs doivent expliquer ce qu'est l'exploration de données aux autres.

L'exploration de données est un terme utilisé pour décrire la découverte de connaissances dans les bases de données (Turban (2007)). L'Office québécois de la langue française définit le DM comme étant une : *« technique de recherche et d'analyse de données, qui permet de dénicher des tendances ou des corrélations cachées parmi des masses de données, ou encore de détecter des informations stratégiques ou de découvrir de nouvelles connaissances en s'appuyant sur des méthodes de traitement statistique.* ». L'objectif du DM selon Spiegler (2003) se définit par la détection, l'interprétation et les prédictions de données par des modèles quantitatifs et qualitatifs. Les modèles conduisent à des informations et des connaissances. Le meilleur exemple est la légende urbaine concernant le grand détaillant qui, par une analyse statistique, lui a permis d'augmenter ses ventes de bière en les plaçant près des couches. Le DM est un processus qui utilise les statistiques, les mathématiques, l'intelligence artificielle et les techniques d'apprentissage automatique pour extraire et identifier l'information utile et pratique des connaissances à partir de grandes bases de données.

Le « Data Mining » (DM) est le précurseur du « Text Mining » (TM) et il possède une littérature plus abondante. Les bases, les processus et les autres caractéristiques du DM sont fréquemment repris par le TM. Une bonne connaissance du DM permet de comprendre plus aisément ce qui est fait en TM.

La différence saillante entre le TM et le DM est qu'elle repose sur des données textuelles plutôt que des données numériques. La source d'information du TM est constituée de documents tels que les courriels, sondages, rapports, documents HTML, etc. Lorsque ces documents sont volumineux, la tâche de les explorer est décourageante. Cette abondance de textes offre une opportunité d'accéder à des données pertinentes, à de nouvelles informations. Un enjeu est de rendre les masses de données utilisables.

Fan et al. (2006) soulignent que cette différence implique aussi de travailler avec des données non ou semi-structurées. La nature d'un texte fait qu'il n'est pas structuré et stocké comme un chiffre dans une base de données en colonne et ligne. Selon Fan et al. (2006), un problème de l'exploration de textes est que les applications et les ordinateurs ne manipulent pas la langue comme nous. Nous pouvons, les humains, distinguer et appliquer des modèles linguistiques, surmonter les obstacles de la grammaire et d'autres erreurs de langage.

Au cœur de l'analyse textuelle se trouve la notion de corpus, qui signifie selon l'Office québécois de la langue française de corpus est : « *Ensemble des sources orales et écrites relatives au domaine étudié et qui sont utilisées dans un travail terminologique.* » Dans le cadre du TM nous pouvons ajouter que le corpus est un recueil de documents concernant une même discipline. Le Web pourrait, par exemple, être jugé comme un corpus de texte géant. Les 17 millions de livres de la bibliothèque du Congrès américain représenteraient un corpus de 136 térabits. En comparaison, le jumeau du corpus en DM est l'entrepôt/base de données.

Cette introduction continue avec l'importance du principe de l'exploration, pour se conclure avec les objectifs de la recherche.

1.1 Importance du principe de l'exploration

La gestion d'entreprise est l'ensemble des activités d'organisation, d'optimisation d'harmonisation de ses ressources (c.-à-d. : « personnelles, financières et matérielles ») dans l'atteinte de ses objectifs. La collecte et l'utilisation de l'information des données textuelles et numériques ont toujours été présentes dans l'entreprise.

L'arrivée des systèmes informatisés a complexifié et bonifié l'utilisation des données de l'entreprise. L'exploration apporte de l'information aux gestionnaires car au lieu de seulement regarder le passé comptable et les analyses textuelles, il permet de comprendre le comportement de la clientèle et prédire l'avenir avec ses algorithmes prédictifs.

Les objectifs pour les analyses d'exploration sont très divers. Le sondage de Rexer Analytics (2010) décrit l'état actuel de l'utilisation de l'exploration de données en entreprise, voir tableau 1. Plusieurs objectifs reliés à la clientèle sont élevés sur la liste. Plus d'un tiers utilisent l'exploration pour améliorer la compréhension des clients.

Tableau 1 : Les principaux objectifs de l'analyse de données

Objectifs	Note globale
Améliorer la compréhension des clients	**37%**
Fidéliser les clients	**32%**
Améliorer les programmes de marketing direct	**29%**
Vente de produits / services aux clients existants	**29%**
Recherche sur le marché / analyse de sondage	**29%**
Acquisition de clients	**27%**
Gestion du risque et de crédit	**26%**
Améliorer la clientèle	**25%**
Détection ou la prévention des fraudes	**21%**
Prévision des ventes	**21%**
Avancement médical / biotech / génomique	**18%**
Optimisation des prix	**13%**
Amélioration de la fabrication	**10%**
Planification et optimisation des investissements	**10%**
Optimisation de site Web ou de moteur de recherche	**8%**
Détection criminelle ou terroriste	**6%**
Collections	**6%**
Optimisation de logiciel	**6%**
Compréhension du langage	**4%**
Levée de fonds	**3%**
Moyenne de buts/objectifs par répondant	**3.8**

Rexer (2010)

Le marché de l'exploration est présent, mais aucune donnée ne permet de connaître sa percée réelle dans le marché. Le gourou, selon le TDWI du « Text analytics », Seth Grimes ainsi que le chef de file en exploration de données et auteur du rapport Rexer Analytics (2010), Karl Rexer, m'ont confirmé l'absence de ce genre d'étude. (voir courriel en annexe 6.1). L'exploration de données est bien présente dans plusieurs secteurs d'activités, par exemple le marketing (41%), les finances (29%) (tableau 2) et ce, sur tous les continents (tableau 3) par exemple l'Amérique (45%) et l'Europe (36%) selon le rapport Rexer Analytics (2010).

Tableau 2 : Secteurs d'activités de l'exploration de données

Domaine	%
CRM/Marketing	**41 %**
Finance	**29 %**
Académique	**25 %**
Assurance	**15 %**
Télécommunications	**15 %**
Vente au détail	**14 %**
Pharmaceutique	**13 %**
Technologie	**13 %**
Médical	**11 %**
Manufacturier	**10 %**
Internet	**10 %**
Gouvernement	**10 %**
Sans but lucratif	**6 %**
Tourisme/Divertissement/Sports	**4 %**
Militaire/Sécurité	**3 %**
Autres	**9 %**

Rexer Analytics (2010)

Tableau 3 : Origine des explorateurs de données

Origine	%
Amérique du Nord • USA 40 % • **Canada 4 %**	**45 %**
Europe • Allemagne 7 % • Royaume-Uni 5 % • **France 4 %** • Pologne 4 %	**36 %**
Asie Pacifique • Inde 4 % • Australie 3 % • Chine 2 %	**12 %**
Amérique centrale et du Sud • Colombie 2 % / Brésil 1 %	**4 %**
Moyen Orient & Afrique (3 %) • Israël 1 % / Turquie 1 %	**3 %**

Rexer Analytics (2010)

Les explorateurs prévoient que l'augmentation du nombre de projets en exploration de données et de l'optimisme des gestionnaires face à de tels projets est partagée dans

les divers milieux corporatifs. Les répondants au sondage de Rexer Analytics (2010) ont aussi partagé leurs idées sur les tendances futures dans l'exploration et elles sont énumérées dans le tableau 4.

Tableau 4 : Tendances futures en exploration de données

Tendances en DM	Nombre de répondants	%
Croissance de l'adoption du DM	50	25.9 %
TM	**32**	**16.6 %**
Analyse de réseaux sociaux	32	16.6 %
Automatisation	26	13.5 %
Informatique en nuage (infonuagique)	15	7.8 %
Visualisation de données	15	7.8 %
Obtenir des outils plus simples à utiliser	12	6.2 %
Utiliser une plus grande quantité de données	11	5.6 %

Rexer (2010)

Au sein des tendances futures, nous retrouvons à égalité au second rang le TM. Cette position est élevée s'explique par la difficulté à examiner et analyser toute une entreprise et l'ensemble de ses processus uniquement à l'aide de nombres. En effet, il y a jusqu'à 80 pour cent de l'information d'une entreprise n'est pas quantitative ou structurée de manière à être capturé dans une base de données relationnelle (Herschel et Jones (2005)). Le TM ajoute donc un élément précieux à la technologie de l'IA existante, car il permet de combler un manque à gagner en termes d'analyse des données corporatives.

Regardons maintenant l'utilisation en entreprise du TM. Le sondage de Rexer Analytics (2010) décrit l'usage du TM. Les répondants du sondage sont des utilisateurs du TM en tant qu'analyse, chef de service et autres. Le sondage exclut les

vendeurs. Trente pour cent des explorateurs de données font actuellement du TM et environ un tiers planifient le faire (figure 1).

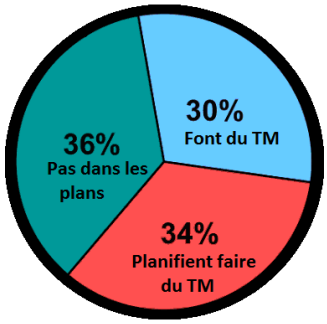


Figure 1 : Répartition de l'utilisation de l'exploration de textes selon Rexer (2010)

Les utilisateurs du TM extraient des thèmes clés (l'analyse des sentiments) à 59 %, utilisent des données additionnelles comme prédicateurs dans un modèle plus grand: 55 % et analysent les réseaux sociaux: 21 %**.**

En résumé, le sondage de Rexer Analytics (2010) souligne que le sujet de l'exploration prend de l'ampleur et ce, spécialement pour les données marketing anglophones. Le marketing est aussi présent dans les objectifs que le secteur d'activité. Le marché anglophone représente déjà plus de 50% de la part du marché (USA, Royaume-Uni, Inde, Australie, Canada, etc.). L'exploration est à son stade d'enfance car sa principale tendance est « la croissance de l'adoption ». Les autres marchés, langues et secteurs d'activités sont au stade de la petite enfance et cette recherche vient corroborer le besoin de comprendre ce qui se fait.

Les données textuelles (comme les données numériques auparavant) que l'on croyait être inclassables et inutilisables, gagnent en intérêt puisque leur valeur est comprise et que les logiciels existent pour leur donner un sens. Laurent (2008) mentionne aussi

qu'il est nécessaire de récolter l'information enfouie dans les données non structurées et peu gérées. Cette récolte sera pilotée par tous les secteurs d'exploitation marketing, ventes, ressources humaines, finances, opérations, etc. Ces secteurs d'exploitation regardent de plus en plus pour prendre des décisions à partir des connaissances contenues dans ces données.

1.2 Objectif de la recherche

Dans le cadre de cette recherche, mes objectifs sont très terre à terre. Le gestionnaire et le chercheur en moi se demande comment on fait pour effectuer du TM dans un milieu d'affaires. Je me demande aussi si un outil commercial parviendrait à extraire de l'information d'un corpus. Il existe des histoires à succès dans le domaine du TM, les applications sont présentes, mais comment une application commerciale réagit-elle avec un corpus francophone d'un domaine spécifique peu ou pas fréquenté par les explorateurs de textes? De plus, comment doit-on effectuer l'achet d'une application.

L'analyse des secteurs d'activités de l'exploration de textes, j'estime important d'effectuer une recherche dans le secteur d'activité juridique francophone, qui autant par sa nature de texte de loi et par la difficulté de la langue, est peu visité par les explorateurs de données.

Peu de présence d'exploration de données peut être répertoriée dans la littérature autant scientifique que commerciale. Selon notre connaissance de la littérature, seulement quatre utilisations sont recensées.

La première présence d'utilisation de l'exploration de textes est la « jurisprudence chiffrée » du groupe d'Éditions Lefebvre. L'outil d'exploration des données (jurisprudence chiffrée) collecte les montants versés/accordés par type de cause et par secteur dans les arrêts de cour d'appel française. (http://www.jurisprudence-chiffree.fr/).

La deuxième présence de l'utilisation de l'exploration de textes dans le domaine juridique est le classement automatique de textes (Pisetta et al. 2006). Dans cette recherche, deux corpus provenant du Bureau International du Travail (BIT) sont analysés. Les textes en relation avec les conventions n^{o} 87 et n^{o} 98 sont reclassés (et seulement ceux-ci). Les résultats sont intéressants car ils réussissent à bien classer les 65 textes avec deux types d'arbres (C4.5 et SVM) pour 2 catégories de classification (les conventions 87 et 98). Ce travail est assez limité autant pour le nombre de textes, de types d'arbre, et de catégories de textes.

La troisième présence est une application développée pour créer automatiquement une base de données de témoins experts à partir d'un corpus contenant des documents juridiques, médicaux, et de nouvelles (Dozier et Jackson 2005). L'application recueille et fusionne les informations provenant des diverses sources. Elle rend l'information sur les témoins experts plus accessible. Enfin, l'application extrait les décisions des Cour subséquentes, utilise les techniques d'apprentissage et lie chaque nouvelle affaire à des cas préexistants qui pourraient avoir un impact. Le but est de ressortir le concept de témoins experts et en faire une liste. Cette analyse est différente que celle que je désire effectuer.

La dernière, LexisNexis Canada (www.lexisnexis.ca) a lancé le 27 septembre 2011, la toute première encyclopédie du droit québécois et canadien qui contient plus de 350 auteurs. « *L'objectif poursuivi était clair dès le départ* », affirme Nicolas McDuff, directeur général de LexisNexis au Québec, « *développer de façon méthodique une encyclopédie du droit positif québécois et canadien qui couvrirait l'ensemble des domaines de pratique* ». Mais cette dernière est une encyclopédie numérique avec un moteur de recherche puissant. Cette encyclopédie est dans le même sens que ce mémoire, mais utilise un moteur de recherche.

Au Québec, pour obtenir de l'information juridique, nous pouvons nous tourner vers une société d'État, soit la SOQUIJ. Elle a le mandat de promouvoir la recherche, le traitement et le développement de l'information juridique en vue d'en améliorer la qualité et l'accessibilité au profit de la collectivité. La société d'état offre un bon terrain d'exploration afin d'approfondir ma recherche.

La société a aussi la mission de recueillir, d'analyser, de diffuser et de publier l'information juridique en provenance des tribunaux et des institutions, de présenter cette information sous la forme la plus complète, la plus à jour, la mieux organisée et la plus facile d'accès. Elle offre une expertise sans égale, des outils de recherche conviviaux, des contenus exhaustifs et un service à la clientèle des plus performants, au bénéfice de ses clients des milieux juridiques, des affaires et du travail ainsi que pour le public en général. Les revenus de la SOQUIJ proviennent entre autres de la vente de résumés de décisions de cour aux avocats par l'entremise de son site TOUT

AZIMUT. Cette avenue est intéressante pour la recherche. (Voir un profil plus complet en annexe 6.2)

La problématique et l'intérêt de la SOQUIJ envers un outil automatisé d'analyse de texte ne datent pas d'hier. Il existe une étude préalable des textes de la SOQUIJ. En 1995, Bertrand–Gastaldy ont eu des résultats peu probants. Selon eux, la sélection de jugements repose sur des opérations cognitives complexes mettant en jeu de nombreuses connaissances spécialisées du domaine juridique et du monde en général et c'est pourquoi, selon eux, les textes se prêtent inégalement à une aide informatique. Un défi de taille se dresse devant moi.

Alors, comment joindre ma curiosité scientifique aux besoins de la science? Comment transposer en objectifs et questions de recherche ce qui doit et peut être fait? Les questions de recherches sont:

1. Comment utiliser efficacement un outil de TM afin de faciliter la découverte de connaissances à partir d'un corpus juridique francophone à la SOQUIJ?
2. Quels sont les opportunités et les défis qui découlent de l'utilisation d'un outil de TM lors de la découverte de connaissances à partir d'un corpus juridique francophone ?
3. Comment choisir une application de TM et évaluer les impacts du choix sur les résultats ?

Dans le cadre de ce mémoire, cette introduction sera suivie du cadre conceptuel, de la méthodologie, des résultats, des discussions et de la conclusion.

2 Cadre conceptuel

Le cadre conceptuel présente les différents concepts propres à l'exploration de données, particulièrement de nature textuelle. Il permet d'élaborer les connaissances nécessaires de la base théorique pour effectuer la présente recherche. Tout d'abord, un survol de l'éthique en exploration est nécessaire, suivra l'exploration de données et pour terminer, l'exploration de textes.

2.1 L'éthique de l'exploration de données

L'éthique est toujours importante, même si le gestionnaire, l'analyste et le décideur se penchent sur des analyses effectuées sur des données secondaires. Ces acteurs doivent pensé et géré la provenance des données, les méthodes d'analyse, qui fait quoi et comment c'est fait, tout autant avec des chiffres que des lettres. La quantité de données disponibles représente un actif et qui dit actif, dit à être protégé. Sur ce point, Menon et Sarkar (2007) analysent comment éliminer / conserver certaines données confidentielles tout en minimisant les impacts sur l'information. Ce n'est que récemment que les chercheurs ont pu identifier des solutions exactes pour cacher ces connaissances sensibles. Des solutions exactes permettent de cacher des connaissances vulnérables sans aucun compromis essentiel. Gkoulalas-Divanis et Verykios (2009) ont réussi à faire du DM tout en préservant les connaissances sensibles dans de grandes bases de données transactionnelles. Selon Agrawal et al. (2007), les systèmes d'informations, dans le secteur de la santé, facilitent le transfert et le partage d'information ce qui tend aussi vers le potentiel d'abus de la vie privée.

L'éthique dans le cadre de l'exploration des données est centrée sur les données. Payne et Trumbach, (2009) identifient quatre problèmes d'éthique avec l'exploration de données. Le premier est la protection. Deuxièmement, des conclusions incorrectes peuvent être émises des données. Le troisième est que les données collectées peuvent être utilisées à d'autres fins que le but premier. Finalement le respect de la vie privée peut être violé. L'article de Wright et al. (2008) illustre bien ces quatre problèmes d'éthique avec l'exploration de données sans les nommer.

Les individus ont droit au respect de la vie privée et ceci est décrit par le groupe consultatif inter-agences en éthique de la recherche du gouvernement du Canada comme étant: « *Le respect de la vie privée est une valeur fondamentale, vue par beaucoup comme essentielle à la protection et à la promotion de la dignité humaine. En conséquence, l'accès aux renseignements personnels ainsi que le contrôle et la diffusion de telles informations ont une importance considérable pour l'éthique de la recherche.* .» Les points importants sont l'accès, le contrôle et la diffusion. C'est sur ce même sujet que Ann Cavoukian (1998), commissaire à la vie privée en Ontario, relève que les entreprises doivent protéger les données recueillies afin d'assurer la vie privée des gens. Au Québec, il n'y a pas de lignes directrices spécifiques à l'exploration des données, mais la loi s'applique à tous les renseignements personnels qu'ils soient individuels, nominatifs, en liste ou dans une base de données. La définition d'un renseignement personnel est: *« information qui permet d'identifier une personne* ». Donc, si un croisement, un tri, une sélection dans

une base de données permet d'identifier une personne, ce sont des renseignements personnels et la loi s'applique (plus de renseignements en annexe 6.3).

Finalement, Wright et al. (2008) identifient un manque de sensibilisation du public envers ce que les entreprises possèdent comme information sur eux et des moyens qu'ils peuvent utiliser pour mieux les connaître. Le public a l'illusion que ses données sont en sécurité et que personne, ni aucune entreprise, peut les connaître parfaitement. Cette illusion persiste aujourd'hui. Le TM pourrait dans un avenir proche changer énormément la connaissance que les entreprises ont sur nous. Nos opinions et sentiments que nous exprimons sur les blogues et les médias sociaux vont venir grossir l'information que les entreprises ont déjà sur nous.

2.2 L'exploration de données

Le DM est un domaine émergent qui a beaucoup attiré l'attention dans un court laps de temps. L'origine de l'exploration de données et de l'intelligence artificielle est décrite par Cooper et Schindler (2000). Ceux-ci décrivent quatre grandes époques de l'évolution afin d'arriver à l'exploration actuelle :

- *la collection des données dans les années 60,*
- *l'accès aux données dans les années 80 avec les systèmes experts,*
- *la navigation dans les données dans les années 90 et des systèmes qui permettaient la découverte de règle,*
- *l'exploration de données dans les années 2000*.

Le DM est tel qu'il est aujourd'hui suite à la croissance exponentielle des bases de données dans toutes les facettes de la vie, l'idée étant d'extraire des informations de haut niveau depuis une abondance de données brutes. Les outils et les techniques sont développés en permanence pour faire face à des montagnes de données, généralement opérationnelles, afin d'obtenir quelques idées et/ou des connaissances. Suivant cette idée, Berry et Linoff (2004) mentionnent que les données sont présentes en quantité, les données sont accessibles dans les bases de données, les systèmes informatiques sont abordables et la pression de la compétition vient de partout. Les gestionnaires sont friands d'opportunités et flairent les menaces. Owen (1998) présente certains facteurs qui attirent les gestionnaires vers l'exploration de données. Il souligne la forte concurrence pour l'attention d'un client dans un marché de plus en plus saturé comme menace importante. La valeur latente de l'information des grandes bases de données, la possibilité d'obtenir une seule vision du client (consolidation des bases de données) et la réduction du coût de stockage et de traitement des données permettent de recueillir, d'accumuler et d'analyser plus de données, ce qui crée de nombreuses opportunités. Dès 1996 Fayyad et al. (1996) ont souligné le besoin urgent d'outils pour aider l'humain à extraire des informations utiles (connaissances) à partir des volumes en croissance rapide de données numériques. Pour sa part, Anthes (2009) indique que la technologie utilisée pour recueillir toutes les données a beaucoup, beaucoup surpassé la technologie permettant de les analyser et les comprendre. Davenport (2006) présente la prise de décision basée sur les faits comme étant une arme stratégique pour les entreprises de pointe. Ils utilisent les faits pour comprendre leur client, améliorer leur chaine d'approvisionnement tout en offrant un

bon service à la clientèle. Fayyad et al. (1996) indiquent que les entreprises utilisent les données pour obtenir un avantage compétitif, accroitre leur efficacité et offrir un meilleur service à la clientèle. Berry et Linoff (2004) soulignent sur ce dernier point ce que les entreprises de toutes tailles doivent copier la clé du succès des petites entreprises tournées vers le service : « *créer une relation d'un à un avec le client* ». L'information qu'une entreprise possède sur eux, les clients, lui confère un avantage, et ainsi l'information pourrait être considérée comme un actif. Beaucoup d'espoir, et peut-être lire pouvoir, active l'entreposage de données. Certains croient que des secrets enfouis peuvent être déterrés et mis à profit. C'est bien le but principal de l'exploration, mettre à profit les données.

Glymour et al. (1997) ont examiné les thèmes statistiques directement liés à l'exploration de données, ainsi que des possibilités de synergie entre les communautés (informatique, IA et statistique) pour de nouveaux progrès dans l'analyse des données. L'exploration de données comprend des tâches telles que l'extraction de connaissances, l'exploration de données et le traitement de modèles. Toutes ces activités sont effectuées automatiquement et permettent la découverte rapide de connaissance, même par des néophytes. L'exploration de données s'appuie sur des modèles mathématiques et Nemati et Barko (2001) les séparent en trois types : « simple, intermédiaire et complexe » afin de brosser un tableau de l'exploration de données dans le domaine des affaires. Nous retrouvons par exemple, les analyses sur les cubes dans le type simple. De plus, les règles d'affinités, des corrélations, des tendances se retrouvent dans le type intermédiaire, tandis que les modèles utilisant

l'apprentissage tel que les réseaux de neurones sont au sein des modèles complexes (Nemati et Barko, 2001). Le DM permet d'identifier des modèles valides, originaux et utiles, et des associations dans les données existantes afin d'ajouter à la connaissance organisationnelle.

Il existe divers intervenants dans le monde de la recherche en DM. Le monde de la recherche est divisé en deux types : « la recherche fondamentale et la recherche appliquée ». Les acteurs des deux domaines ont chacun leur propre point de vue. Les groupes les plus importants d'acteurs fondamentaux sont la communauté des chercheurs et des universitaires en DM par le nombre d'articles retrouvés. Le groupe d'intervenants appliqué comprend par exemple les chercheurs, mais aussi les gestionnaires et les experts du domaine qui ont leurs propres intérêts fondés sur l'utilité des résultats de recherche DM (Pechenizkiy et al. 2008).

2.2.1 La recherche fondamentale en exploration de données

La recherche fondamentale sur le DM a développé avec succès des techniques avancées, des algorithmes de DM et des règles formelles, par exemple Better et al. (2010), Ruthledge (2009), Menzies et al. (2007) et Ahmed et al. (2006).

Dès 1999, Chen et al. (1999) ont fait une analyse des techniques utilisées en DM, car il existait déjà plusieurs écrits techniques. Ahmed et al. (2006) ont examiné les moyens afin de partitionner les données lors de l'exploration de données avec les règles d'association lors de très grands ensembles. Basit et Jarzabek (2009) ont effectué une recherche pour éliminer les doublons de codes (autant des lignes que de

structures) afin d'alléger la programmation. Jukic et Nestorov (2006) tentent d'utiliser toutes les données d'un entrepôt de données pour effectuer des règles d'association. Ils n'utilisent pas une interface, ils le font directement sur l'entrepôt en utilisant les capacités de calculs de l'entrepôt. Liu et Tuzhilin (2008) suggèrent un processus de gestion concernant les modèles de DM. Un modèle est créé quand un algorithme est appliqué aux données. Il arrive donc un point où beaucoup de modèles sont utilisés, alors ils suggèrent de : « créer, analyser et maintenir » les modèles de DM. Je partage l'avis de Pechenizkiy et al. (2008) qui mentionnent qu'actuellement, l'accent de la plupart des recherches en DM est encore uniquement sur la technologie et peu sur la pratique. C'est lorsque l'on regarde la recherche fondamentale, mais regardons la recherche appliquée.

2.2.2 La recherche appliquée en exploration de données

La recherche appliquée s'intéresse à peu près à tout ce qui touche le DM. Afin de bien comprendre la littérature et le DM, il est important de comprendre le processus. Fayyad et al. (1996) décrivent bien un processus d'exploration de données pour l'acquisition de nouvelles connaissances. Le processus interactif et itératif implique plusieurs étapes. Fayyad et al. (1996) présentent neuf (9) étapes pour obtenir une connaissance avec le processus du DM. Ces étapes permettent de comprendre les caractéristiques du DM et en même temps la littérature.

Étape 1 : Développer des connaissances du domaine

La première étape est de comprendre le domaine d'exploration et d'identifier l'objectif du point de vue du client. Brisson et al. (2006) présentent la méthodologie EXCIS (« Extraction using a Conceptual Information System ») orientée en ontologie qui permet d'intégrer la connaissance des experts d'un domaine dans un processus d'exploration de données. Les résultats de l'étude de Sinha et Zhao (2008) (figure 2) soulignent la synergie créée entre la connaissance du domaine et le DM. Lorsque la connaissance est facilement disponible auprès d'experts ou d'autres sources, il est logique d'intégrer ces connaissances dans le processus de décision. L'objectif principal est d'améliorer la qualité de la connaissance extraite et de faciliter son interprétation par une connaissance du domaine.

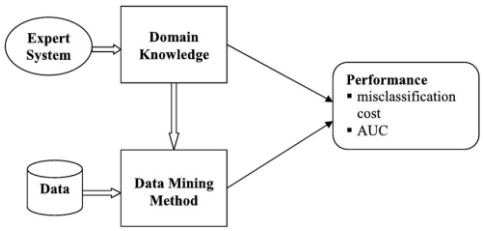


Figure 2 : Modèle de recherche de Sinha et Zhao (2008)

Étape 2 : Sélectionner, créer l'ensemble des données requises

La deuxième étape est de sélectionner, créer l'ensemble des données requises pour effectuer l'exploration. Berry et Linoff (2004) décrivent bien l'état de l'exploration de données. Tout d'abord, les données sont souvent enfouies dans les grandes bases de données sous-utilisées et ils les appellent les dépotoirs de données. Ces dépotoirs contiennent parfois des données sur plusieurs années. La littérature sélectionnée ne

m’a pas permis de trouver des indications sur la sélection des données, ni sur comment créer l’ensemble de données.

Étape 3 : Préparer et nettoyer les données

La troisième étape consiste à préparer et à nettoyer les données. À l’étape précédente, l’ensemble des données a été créé, mais cet ensemble a besoin d’être peaufiné. Berry et Linoff (2004) mentionnent que dans de nombreux cas, les données sont nettoyées et consolidées dans un entrepôt de données. Buck (2001) mentionne que la majeure partie des projets de DM que les organisations entreprennent aujourd'hui (2001) se concentre sur des données structurées qui ont été préparées, nettoyées et transformées. Mais nous pouvons aussi ajouter les opérations de base telles : « la suppression des observations aberrantes (outliers), du bruit, recueillir l'information nécessaire pour rendre compte de modèle ou de bruit, décider de la stratégie pour le traitement des champs manquants de données, etc. ».

Étape 4 : Réduction et la projection des données

La quatrième étape consiste à rendre utiles les données en fonction de l'objectif de la tâche. Avec la réduction, le nombre effectif de variables à l'étude peut être réduit pour améliorer l’efficacité du travail.

Étape 5 : Les méthodes d’exploration de données correspondent aux objectifs

La cinquième étape est de s'assurer que les méthodes de DM correspondent aux objectifs du processus (étape 1). Par exemple, est-ce que la segmentation va répondre aux questions posées? Selon Nemati et Barko (2001) le DMO *(DM Organisationnel)* utilise trois types de modèle pour répondre à ces questions. Ils sont modèles simples (par exemple, basé sur SQL d'interrogation de données, OLAP), les modèles intermédiaires (par exemple, la régression, arbres de décision, l'analyse par segmentation) et les modèles complexes (par exemple, les réseaux neuronaux).

Les modèles simples sont construits en utilisant les cubes OLAP (Online Analytical Processing) et/ou des requêtes SQL (Structured Query Language). OLAP présente des données en temps réel à l'utilisateur sous la forme d'un cube multidimensionnel. Les dimensions du cube sont les attributs de la requête (comme la clientèle, de temps et de produit), tandis que les données résumées (tel que le décompte, la somme et la moyenne) sont la granularité de la demande. Il est possible d'analyser des données via des agrégations, des sélections (découpe (dicing) et tranchage (slicing)), de drill-down, et drill-through. Les requêtes SQL peuvent donner des statistiques descriptives pour les sous-ensembles de données stockées dans des bases de données relationnelles.

Les modèles intermédiaires sont construits en utilisant des techniques d'analyse statistique. L'analyse de régression, les arbres de décisions (CART et CHAID), et l'analyse par segmentation sont des modèles intermédiaires qui tirent leur force des statistiques. Les règles d'association sont un outil clé du DM mais la régression est peut-être la méthode statistique la plus commune des études en sciences sociales.

Un exemple est « 98% des clients qui achètent des pneus et des accessoires automobiles optent également pour des services pour l'auto ». Lorsqu'il est appliqué à une base de données ou un entrepôt de données, une telle règle peut révéler de nouvelles informations cachées dans les données. Allentuck (2006) souligne la corrélation entre deux produits (fromage à la crème et nourriture pour chien), mais explique aussi qu'il n'y a pas un lien causal d'achat. Les facteurs externes, par exemple les promotions, peuvent expliquer une causalité entre deux produits corrélés. Allentuck (2006) mentionne qu'il est mieux de connaître les relations (corrélation) entre les produits que de gérer dans le noir sans le savoir. Il s'agit d'une technique dans les applications de marketing direct qui tentent d'inférer des modèles sur le comportement du consommateur. Buck (2001) relève plusieurs fonctionnalités du DM, notamment l'introduction de séquence dans la série temporelle. En introduisant la notion d'ordre dans l'analyse de tendance au sein des données complexifie l'exploration.

Les modèles complexes sont construits en utilisant l'intelligence artificielle. Les réseaux de neurones (RN) par exemple tirent leur force de méthodes non-statistiques. Les RN imitent les processus biologiques au sein de neurones humains à apprendre des expériences antérieures et de stocker ces connaissances pour une utilisation future. Même si ces techniques sont très puissantes à la modélisation prédictive, ils laissent beaucoup à désirer dans la catégorie des explications solution.

Pour conclure, un analyse se doit de comprendre son problème, cerner ses objectifs afin d'évaluer quel type d'analyse sera requise.

Étape 6 : Sélectionner le type d'analyse et de modèle

La sixième étape est la sélection du type d'analyse et de modèles à utiliser. Des objectifs cernés et une bonne compréhension de ce qui est fait, l'analyste va ressortir le type d'analyse il devra faire. Par exemple, l'utilisateur peut être plus intéressé à comprendre le modèle que ses capacités prédictives. Ainsi, à cette étape il faut sélectionner les modèles et les paramètres appropriés pour la recherche.

Étape 7 : Effectuer l'exploration de données

L'utilisateur peut de manière efficace faire l'exploration grâce aux étapes précédentes et ce, avec des outils (applications) sophistiqués, y compris les outils avancés de visualisation qui aident à extraire l'information enfouie dans des fichiers d'entreprise ou des documents d'archives publiques (Berry Linoff (2004). Au fil du temps, de meilleures pratiques ont émergé pour améliorer la qualité des projets d'exploration de données. Les méthodes les plus utilisées, selon le site KDnuggets (2002), tableau 5, sont SEMMA du SAS Institute et CRISP-DM qui détiennent 63% de l'utilisation par les répondants. Nous pouvons aussi rajouter le cycle vertueux (Virtuous Cycle) décrit par Berry et Linoff (2004) ainsi que le DMAIC Six-Sigma à ces deux méthodologies les plus utilisées (voir plus de renseignements sur les méthodologies en annexe 6.4).

Tableau 5 : Méthodologies utilisées pour effectuer l'exploration de données

Méthodologie	Résultat
CRISP-DM (96)	51 %
SEMMA (22)	12 %
Méthodologie développée par mon entreprise (13)	7 %

Méthodologie que j'ai développée (43)	23 %
Autre (8)	4 %
Aucune (7)	4 %

KDnuggets (2002)

Étape 8 : Analyser, interpréter et vulgariser les résultats explorés.

Cette étape implique l'analyse, l'interprétation et même une nouvelle itération des étapes précédentes (si le besoin est). L'explorateur est souvent un utilisateur final, selon Berry et Linoff (2004), habilité par les outils d'exploration de données et d'autres outils de requêtes pour obtenir réponse à ses questions *ad hoc* et les obtenir rapidement, avec peu ou pas de compétences en programmation. Les outils d'exploration de données sont facilement combinés avec des tableurs et autres outils de développement logiciel. Ainsi, les données extraites peuvent être analysées et traitées rapidement et facilement. Apte et al. (2002) soulignent que le DM est efficace, mais qu'il doit continuer à développer des techniques qui assistent l'usager dans la découverte de nouvelles connaissances. Trouver un filon consiste souvent pour l'explorateur à trouver un résultat inattendu et exige aux utilisateurs finaux de penser d'une manière créative. Sinha & Zhao (2008) eux mentionnent que les techniques de DM produisent de bonnes statistiques et trouvent des modèles dans de gros volumes de données, mais ils ne sont pas très intelligents dans l'interprétation de ces résultats, ce qui est crucial pour les transformer en connaissances intéressantes, compréhensibles et gérables.

Étape 9 : Agir sur les connaissances découvertes.

En IA il faut savoir comment utiliser les nouvelles connaissances, comment les intégrer, les documenter et transmettre aux parties intéressées au bon moment. Ce processus comprend également la vérification et la résolution des conflits potentiels avec les connaissances précédemment crues (ou extraits). Les cas à succès soulignent les utilisations commerciales les plus courantes du DM dans la finance, la distribution, et les secteurs des soins de santé (Turban, 2007). Le DM est utilisé pour réduire les comportements frauduleux, en particulier dans les réclamations d'assurance et de l'utilisation de cartes de crédit (Chan et al., 1999); pour identifier les habitudes d'achat des clients (Hoffman, 1999); pour récupérer des clients rentables (Hoffman, 1998); et pour identifier les règles d'associations à partir des données historiques et à l'aide de l'analyse du panier de consommation. L'exploration de données est déjà largement utilisée pour mieux cibler les clients, et avec le développement du commerce électronique, cela ne peut que devenir plus important avec le temps. Hair (2007) donne un aperçu de l'analyse prédictive et comment elle se répercute dans la création de connaissances en marketing. Il y voit l'utilité et même souligne qu'elle devrait être plus utilisée par les organisations et les chercheurs. Cui et al. (2006) utilisent le DM pour évaluer la réponse lors de campagne de marketing direct auprès de certains groupes. Liu et Shih (2005) se basent sur les groupes qui prennent les décisions d'achat selon le poids des critères RFM (recency, frequency et monetary) dans le but d'améliorer un système de recommandation. Cortez et al. (2009) utilisent le DM avec les données (composés chimiques) obtenues de vin portugais ainsi que des classements de ces derniers par les œnologues pour créer une classification supervisée ainsi que donner des pistes d'amélioration aux vins (quels

sont les attributs nécessaires lors de la classification). Bhandari et al. (1997) explique comment l'application AdvanceScout est utilisée par les entraineurs de la NBA pour la sélection des joueurs lors des parties. L'application donne les statistiques des joueurs selon les situations de jeu ainsi que les joueurs présents (adversaires et pairs). Wu et al. (2006) facilitent le travail des « intelligence workers » avec l'analyse de la navigation web des internautes par le DM (segmentation). Même dans le secteur minier, les données sont probablement l'actif le plus important que les sociétés minières possèdent (Bacon et Webb, 2009). Ces données pourraient (non structurées telles que des enquêtes de terrain minier), si accessibles, donner à l'entreprise une longueur d'avance sur son prochain projet – ainsi que donner au terme "exploration de données" une toute nouvelle signification.

2.3 L'exploration de textes

Laurent (2008) se posait la question suivante : « *How do I get meaning out of all the unstructured text that presently exists in my company ?* ». Sa réponse fut d'extraire les connaissances par le TM. L'exploration de textes est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques. Les disciplines impliquées sont donc la linguistique calculatoire, l'ingénierie du langage, l'apprentissage artificiel, les statistiques et bien sûr l'informatique. Par rapport au DM, le TM est par nature moins précis et complet;

«assez bon» est souvent la règle. Ainsi, selon Stoddler (2010), le TM peut être plus précieux quand il est utilisé en tandem avec des données structurées. C'est aussi ce que Carricano et De Lassence (2009) supportent. L'incorporation de textes aux modèles de DM représente bien un enjeu important afin d'obtenir de la connaissance pour la prise de décision. Le déploiement d'outils mixtes est compris de la part de la communauté académique et du monde de l'entreprise. Il faut établir les bénéfices liés à une modélisation issue de données complètes, nombres et mots, afin de prévenir une « myopie » de la décision.

Le TM est arrivé il y a environ 10 ans, selon King (2009) en 2009, mais il n'est pas encore largement utilisé. Le déploiement des outils TM représente un intérêt bien compris de la part de la communauté académique en systèmes d'information et du monde de l'entreprise (Carricano & De Lassence (2009)). Regardons ce que nous présente la littérature.

2.3.1 La recherche fondamentale sur l'exploration de textes

Le peu de littérature en recherche fondamentale du TM se concentre principalement sur les méthodes d'exploration et le type de classification.

Depuis les débuts l'exploration de textes les homonymes ont présenté un défi d'analyse. Schmid (1994) parvient à les classer avec des arbres de décisions lors d'analyse de textes.

Plus récemment, l'utilisation du langage naturel était présente en TM. L'office de la langue française définit le langage naturel comme étant un : « terme est utilisé en intelligence artificielle pour qualifier les langages humains écrits et parlés et les distinguer des langages de programmation artificiels. Un des buts majeurs des travaux en intelligence artificielle est la fabrication d'ordinateurs communiquant en langage naturel; l'apprentissage de ce type de communication avec la machine demanderait un effort minimum de la part d'un utilisateur novice. » Marcoux et Rizkallah (2009) suggèrent même que le langage naturel est le fondement sur lequel les outils de gestion et systèmes d'information devraient être développés.

Lee et al. (2010) examinent quatre procédés en traitement du langage naturel. Ils identifient les caractéristiques et les limites de : « (1) l'analyse sémantique latente, (2) l'analyse probabiliste latente sémantique, (3) l'allocation de Dirichlet latente et (4) le modèle de sujet corrélé ». Leurs particularités sont peu utiles car mon but n'est pas d'utiliser le langage naturel.

J'ai recensé peu d'article en recherche fondamentale. Regardons la recherche appliquée

2.3.2 La recherche appliquée sur l'exploration de textes

La littérature en recherche appliquée du TM est diversifiée comme le mentionnaient Cohen et Hunter (2008) (il y a au moins autant de motivations pour effectuer des travaux d'exploration de textes comme il y a de types de chercheurs).

Dans les débuts du TM, les outils d'analyse étaient moins poussés, mais l'intérêt pour la thématique était déjà présent. Swanson (1987) compare les bibliographies d'articles scientifiques sur un sujet très précis : « *comment certains changements dans le sang réduisent l'impact de la maladie Raynaud* ». La conclusion de l'article souligne que certains articles ne partagent pas du tout la même bibliographie malgré les connexions logiques entre eux.

Les bases de données, entrepôts de données, comptoirs de données et autres contiennent beaucoup de données, et ce tout d'abord sur la clientèle, et quelques fois dans les champs libres. Selon Carricano & De Lassence (2009), le traitement par le TM permet d'obtenir une meilleure connaissance de la clientèle et sur les projets que l'entreprise a effectués. Fréquemment les entreprises n'ont pas les ressources nécessaires pour examiner les rapports bilan des projets. Ces rapports contiennent la connaissance des projets, soit individuellement ou collectivement, ainsi que des informations importantes sur les bons/mauvais coups. Le TM permet de découvrir des modes, des associations et des tendances à partir de ces rapports (Choudhary et al. (2009)).

Les sondages, autant papier qu'électronique, contiennent des questions ouvertes. L'analyse de ces dernières a toujours été la bête noire des chercheurs. L'utilisation du TM par Abd-Elrahman et al. (2010) permet de comparer l'interprétation de réponses ouvertes aux questions fermées des étudiants sur la fiche d'évaluation des professeurs avec l'analyse automatisée.

Le volume et la complexité des documents textuels sur le web sont en augmentation, ce qui est décourageant à analyser manuellement. « *La technologie a créé le "problème" et la technologie offre des solutions…* » soutiennent Leong et al. (2004) dans leur étude sur l'utilisation des technologies de TM pour analyser des messages texte promotionnels en ligne de différents concurrents. Lu et al. (2010) analysent les réseaux sociaux de communauté de pirate informatique et Chung et al. (2005) par des recherches web, utilisent le TM pour faire une carte des connaissances d'un sujet particulier, le BI (figure 3). Dahl (2010) jette un regard sur les thèmes actuels de la recherche marketing publiée sociale en utilisant le TM sur le web pour analyser des articles publiés dans les 5 dernières années. Chou et al. (2008) proposent le TM pour détecter les mauvais sites Internet au travail (sites non reliés au travail). La catégorisation de textes est une approche prometteuse pour la détection de ces mauvais sites et ce, par la grande précision de la classification supervisée. Ils ont obtenu une précision de classification de plus de 99%. De plus, ils suggèrent d'étudier l'effet de l'incorporation de certaines fonctions liées aux images et aux contenus multimédias sur la performance de classification.

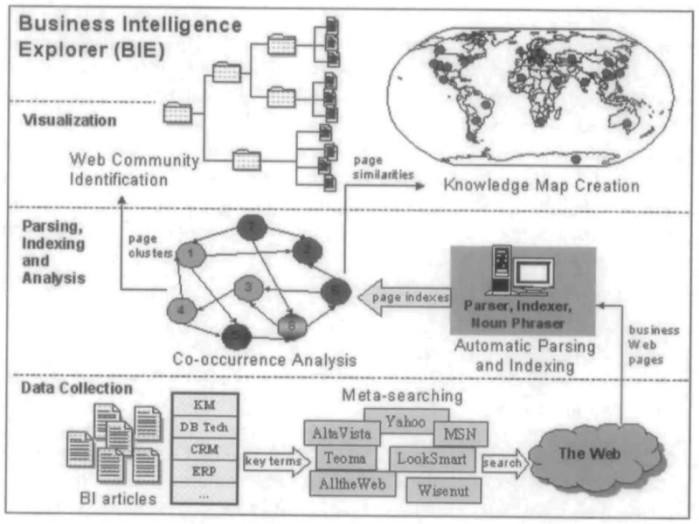


Figure 1. A Visual Framework for Knowledge Discovery on the Web

Figure 3 : Le cadre de travail de Chung et al. (2005)

Le TM ne se contente pas de "compter" le nombre de fois qu'un mot est utilisé, mais plutôt il repose sur l'identification des modèles dans lesquels le ou les termes sont utilisés, permettant ainsi aux chercheurs d'identifier les concepts de base et des structures dans le texte précédemment non structuré de manière plus contextuelle (Dahl (2010). Dahl (2010) a validé la cooccurrence dans un corpus de 272 résumés d'articles portant sur le marketing social. Sans étonnement les termes les plus cooccurrents sont : « marketing social », mais vient après la « santé publique ». Le TM fait partie de la famille du DM et reprend ces caractéristiques et définitions.

2.3.3 Comment faire de l'exploration de textes

La littérature en TM est moins abondante que celle en DM, mais elle reste très intéressante. Comment pouvons-nous l'appliquer à l'intelligence d'affaires? Afin de brosser une bonne image du TM, regardons un processus TM qui se veut la

combinaison ce que proposent Fan et al. (2006), Choudhary et al. (2009) et Tseng et al. (2007) (figure 4 et 5):

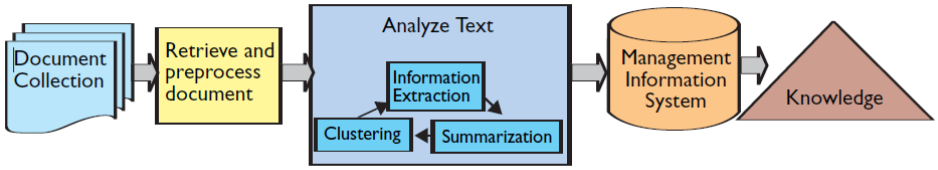


Figure 1. An example of text mining.

Figure 4 : Exemple de TM par Fan et al. (2006)

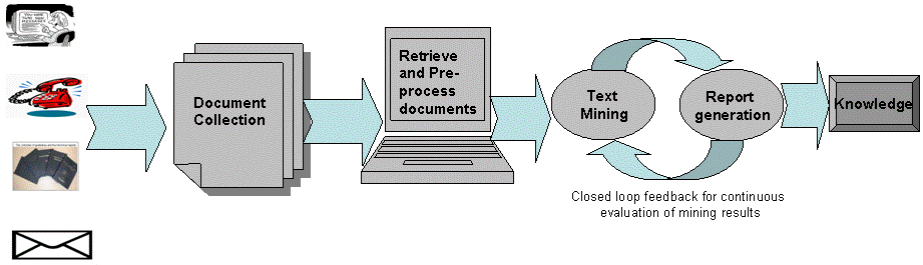


Figure 5 : Exemple de TM par Choudhary et al. (2009)

1. L'identification de l'étendue du projet
2. La sélection du corpus
3. Le prétraitement et l'acquisition des données
4. La modélisation TM
5. La connaissance

Cette synthèse des processus en TM de Fan et al. (2006), Choudhary et al. (2009) et Tseng et al. (2007) représente bien la littérature et est réutilisée dans les paragraphes suivants. Les 5 étapes de l'exécution d'une analyse de textes proposée ici haut sont reprises afin de préciser sur ce qui a été fait dans chacun des secteurs de l'exploration de textes.

2.3.3.1 Étape 1 : Définir l'étendue du projet

La première étape logique d'un projet est l'identification, la description et la planification de la portée, des tâches, des concepts à des fins d'analyse (Tseng et al. (2007), Cohen et Hunter (2008))

2.3.3.2 Étape 2 : Sélectionner le corpus

Le corpus est un recueil de documents concernant une même discipline regroupée pour un travail spécifique. Dans le cadre d'une analyse statistique, il est la base sur laquelle seront extraites les nouvelles informations. Le corpus doit être d'une taille suffisante pour permettre l'analyse, d'être du domaine de l'analyse désirée et ce, dans la bonne langue.

2.3.3.3 Étape 3 : Prétraiter et acquérir les données

Les scientifiques, sans formation linguistique, sont souvent surpris par l'éventail des possibilités linguistiques pour exprimer même les concepts les plus simples selon Cohen et Hunter (2008) et pour cela, il faut examiner les données manuellement.

La préparation des données est le préalable à tout travail sur le corpus. Il s'agit donc dans un premier temps de récupérer les documents qui peuvent être de formats très différents, de concevoir la structure de la base de données et enfin de stocker les documents dans la base de stockage (Carricano & De Lassence (2009)). Il s'agit de l'étape permettant d'effectuer une synchronisation massive du corpus vers les applications d'exploration de textes. Elle est basée sur des connecteurs servant à exporter ou importer les textes dans les applications, des transformateurs qui

manipulent les données (agrégations, filtres, conversions...), et des mises en correspondance (mappages).

2.3.3.4 Étape 4 : Modéliser

La première étape de la modélisation consiste à reconnaître les mots, les phrases, leurs rôles grammaticaux, leurs relations et leur sens. Cette étape est commune à tous les traitements. Par la suite, il est important d'effectuer la standardisation, d'éliminer les formats ou caractères qui pourraient fausser l'analyse et enfin réaliser la lemmatisation du corpus. La lemmatisation consiste à grouper les mêmes mots (journal/journaux) et épurer le vocabulaire des mots non informatifs Carricano & De Lassence (2009). Ainsi, l'exploration de textes consiste en l'indexation d'un ensemble de textes par rapport aux mots qu'ils contiennent. On peut ensuite interroger l'index ainsi créé pour connaître les similarités entre une requête et notre liste de textes. Un algorithme d'indexation peut se décrire comme suit :

- Indexer le texte avec les mots qui le composent
- Indexer les mots contenus par rapport aux textes les contenant

« *Il n'existe pas de règle permettant de définir à priori quel modèle va être le meilleur pour un problème précis. La solution empirique, rapide et efficace, est simplement de tous les tester et de voir quel est le modèle dit « **champion** »…* » Carricano & De Lassence (2009)

Il est important aussi de regarder certaines analyses qui permettent de créer des modèles statistiques. Tout d'abord, l'analyse de segmentation regroupe automatiquement un corpus en groupes distincts de documents similaires et discerne des thèmes généraux cachés. L'analyse par segmentation, par exemple, a été appliquée pour améliorer l'efficacité de la catégorisation de textes. La technique de segmentation utilisée par Wei et al. (2008) permet efficacement de regrouper des documents dans un corpus multilingue. Dans leur étude, Ponmuthuramalingam et Devi (2010) introduisent une notion de fréquence des mots à l'analyse de segmentation. Cet ajout permet d'améliorer la performance des algorithmes.

Une utilisation particulière de traitement de l'information non structurée peut déboucher sur une analyse du sentiment. Par exemple, ces documents montrent-ils que mon produit sera bien vu par les utilisateurs? Ou même, est-ce possible d'identifier des individus qui pourraient être à risque de suicide en analysant les sentiments (humeurs et émotion) du contenu de leurs blogues (Goh et Huang, 2009).

L'exploration de textes peut analyser les recherches d'informations sur un moteur de recherche de documents. Ainsi, l'exploration s'intéresse a priori aux types de requêtes possibles et aux indexations associées qu'à l'interprétation des textes. D'autres exemples d'applications sont la correction des fautes d'orthographe, la traduction, le dialogue personne-machine ou l'imitation d'un style d'écriture. L'exploration de textes se distingue du traitement automatique du langage naturel par son approche générale, massive, pratique et algorithmique en raison de sa filiation avec l'exploration de

données. Son approche est moins linguistique. De plus, l'exploration de textes ne s'intéresse pas au langage oral comme le fait la reconnaissance vocale.

L'interprétation de la modélisation permet de sélectionner un bon modèle parmi d'autres. Des exemples d'applications sont la classification de courriels en pourriel, c'est-à-dire les courriels non sollicités, ou pourriels. Il faut comparer les modèles pour en sélectionner un. Il s'agit non seulement d'identifier un bon modèle, un modèle qui minimise le taux d'erreur, mais qui soit surtout robuste, en d'autres termes dont la performance reste stable s'il est appliqué à des données différentes de celles sur lesquelles il a appris.

2.3.3.5 Étape 5 : Créer la connaissance

Les systèmes d'information de gestion, des systèmes informatisés, continuent de recueillir des données pertinentes, à la fois de l'intérieur et l'extérieur de l'organisation. Les analyses de textes effectuées sont ensuite traitées, intégrées et stockées dans une base de données centralisée (ou entrepôt de données) où elle est constamment mise à jour et rendue disponible à tous ceux qui ont le pouvoir d'y accéder, sous une forme qui convient à leur but. Les outils de TM offrent même maintenant la possibilité de visualiser les liens et les rapprochements de concepts (figure 6). Dans l'étude de Yang et al. (2008), les auteurs offrent un aperçu de l'interface graphique d'un outil de TM. Ci-jointe, figure 6, une visualisation d'agents chimiques et biologiques présentée au sein de brevets (droit d'interdiction par un tiers d'une invention).

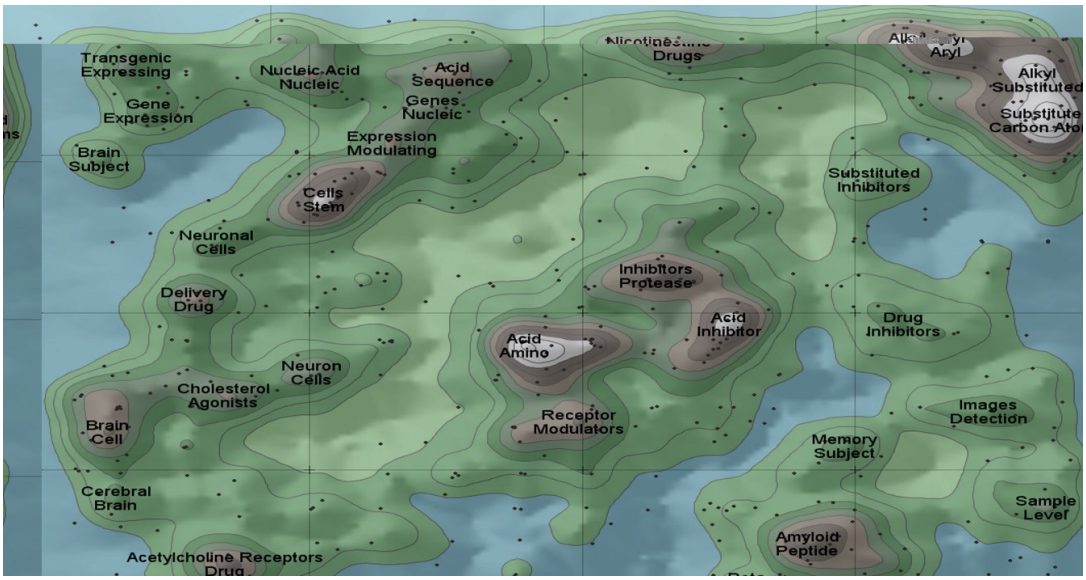


Figure 6 : Visualisation de concepts en TM par Yang et al. (2008)

Les systèmes d'information de gestion ont une approche organisée, afin que l'étude des besoins d'information de gestion de l'organisation, à tous ses niveaux, permette de rendre opérationnelles les décisions tactiques et stratégiques. L'objectif est de concevoir et d'appliquer des procédures, des processus et des routines qui fournissent des rapports détaillés de manière précise, cohérente et en temps opportun. Bref, que l'information puisse ultimement être transformée en connaissances. Par exemple, l'étude de Kloptchenko et al. (2004) combinait des méthodes de DM et TM pour analyser les données numériques et textuelles à partir des rapports financiers, afin de voir si la partie textuelle du rapport contient des indications sur la performance financière future. Elle a montré que les grappes d'analyse numérique et textuelle ne coïncident pas. La partie numérique d'un rapport reflète les performances passées d'une entreprise. La partie textuelle du rapport détient quelques messages à propos de la performance future. Quelle décision un investisseur doit-il prendre, investir ou non ?

L'objectif du cadre conceptuel était de comprendre les différents concepts liés à la recherche. Premièrement l'exploration de textes a été redéfinie par l'étude du processus et suivie par l'implication de l'exploration de données et de textes en intelligence d'affaires tout en démontrant les différentes études, sans oublier l'éthique de l'exploration qui est souvent sous-estimée.

La section suivante permettra d'élaborer une méthodologie adéquate afin d'effectuer cette recherche.

3 Méthodologie

Afin d'atteindre l'objectif de la recherche qui est d'appliquer l'exploration des textes avec un outil de TM commercial sur un corpus hors du commun francophone, il est important de suivre une méthodologie rigoureuse et pertinente. Les deux approches les plus fréquentes de recherche scientifique en système d'information sont l'approche béhavioriste et le « design science » (Österle et al. 2011). L'approche béhavioriste se concentre sur l'observation des caractéristiques des systèmes d'information et sur le comportement des usagers tandis que le « design science » porte davantage sur le développement d'un artéfact technologique. L'application de l'exploration des textes avec un outil de TM commercial afin d'améliorer l'organisation est à la base de la recherche et donc la création d'un artéfact technologique est attendue. Pour cette raison, la présente recherche se base sur les principes du« design science ».

3.1 « Design science »

L'objectif principal de la méthodologie « design science » selon Hevner (2004) est de créer et d'évaluer un artéfact technologique afin de résoudre un problème organisationnel. L'artéfact est représenté selon différentes formes : un construit, un modèle, une méthode et une instanciation (March et Smith, 1995). Selon Hevner (2004), l'instanciation montre que les construits, les modèles ou méthodes peuvent être mis en œuvre dans un système opérationnel. L'instanciation démontre la faisabilité par l'évaluation concrète de la pertinence d'un artéfact à son but et permet d'en apprendre davantage sur le monde réel. Les manifestions dans le monde réel peuvent être, selon Österle et al. (2011), des lignes directrices, des normes, des brevets, des logiciels (avec le code source ouvert), des modèles d'affaires, de nouvelles entreprises et bien plus encore.

Le modèle de recherche en système d'information de Hevner (2007) est décomposé en trois cycles: le cycle de pertinence, le cycle de rigueur et le cycle de conception. La pertinence relie le contexte environnemental du projet de recherche avec les activités du « design science ». La recherche doit avec l'artéfact être pertinente et améliorer, ou vouloir l'améliorer, une organisation. La rigueur relie les activités du « design science » avec la base de connaissances scientifiques. Le mot science est présent dans le terme « design science » ainsi, la rigueur scientifique doit être présente alors une recherche débute par une bonne revue de la littérature afin de connaître la situation initiale de la recherche. La conception itère entre les activités de base de l'élaboration et de l'évaluation de la conception de l'artéfact et les processus

de la recherche. Elle est le ciment; elle amalgame la rigueur, la pertinence, le développement de l'artéfact et des exigences pour obtenir les résultats.

De base, une recherche en « design science » doit avoir une problématique ou opportunité actuelle organisationnelle et contribuer à l'avancement des connaissances. Il aurait été simple de proposer des hypothèses logiques, de les tester et d'émettre les résultats, par exemple : « L'outil d'analyse de textes peut-il obtenir un niveau de performance de 90%? » Mais logiquement, l'exploration de données dans le cadre de prise de décision doit pouvoir permettre de faire avancer l'organisation et doit être plus large comme le permet le design science. Ainsi donc, cette recherche se base sur cette méthodologie établie au sein de la recherche académique en système d'information. Plus spécifiquement, voici comment les trois phases du modèle de Hevner s'appliquent à notre contexte de recherche :

La pertinence de la recherche est soulevée dans la section d'introduction. Le TM commercial est très rarement appliqué à l'analyse de données textuelles francophones comparativement à l'exploration de données quantitatives. Le corpus étant la base de la problématique, une fois qu'il sera identifié, nous pourrons alors demander aux gestionnaires les besoins de l'organisation.

La rigueur de la recherche est assurée par le cadre théorique et la méthodologie qui font état des connaissances nécessaires à la bonne conduite de la recherche. Finalement, l'artéfact technologique de l'instanciation sera détaillé dans les chapitres

suivants par ses contributions et ses limites, par la démonstration de la faisabilité de l'utilisation et par l'apprentissage d'un outil commercial.

3.2 Les méthodologies d'analyse de données textuelles

Comme mentionné plus tôt, les méthodologies en TM se ressemblent ainsi que les méthodologies des applications. Revoyons les étapes clés qui permettront d'atteindre les objectifs ainsi que la rigueur de la recherche.

3.2.1 Étape 1 : Présenter le cas

Le « design science » porte une attention particulière au fait que l'artéfact technologique doit répondre à une problématique organisationnelle. Ainsi, la définition des besoins d'affaires de l'organisation est importante et est la première étape du projet. Le site de la recherche est la Société québécoise d'information juridique. La SOQUIJ possède une base de données avec des résumés de décision de cour. Ces derniers sont le produit que vend la société pour son autofinancement. Ils sont accessibles via le site web par mot clé ainsi que par opérateur booléen. La banque de données contient toutes les décisions rendues par les différentes instances de nombreux domaines de droit que la société dessert.

Afin de promouvoir la recherche et le développement de l'information juridique, la SOQUIJ recherche toujours de nouvelles techniques, de nouvelles connaissances. Dans cette quête d'amélioration continue, elle se questionne sur le sujet de l'exploration de données textuelles pour son corpus de résumés de décisions de cour. La question soulevée lors des discussions avec la SOQUIJ est la suivante :

Est-ce que l'exploration de données textuelles par un outil commercial permettrait d'organiser différemment le corpus de résumés de verdicts et de les rendre plus simples d'accès tout en étant exhaustifs?

Les objectifs de la recherche sont de savoir comment un outil de TM peut être utilisé pour faciliter la découverte de connaissances sur un corpus ainsi que soulever les défis et bénéfices de l'utilisation de l'outil de TM. Une première analyse d'exploration peut être l'analyse des liens dans les textes et de ses concepts. Cette analyse permet d'identifier les relations entre les concepts des données textuelles sur la base de patrons connus (connus une fois que l'analyse est effectuée). Ces relations peuvent décrire les liens (ou associations) entre les résumés de décisions de cour. Par exemple, l'extraction de la décision d'une cause peut revêtir un intérêt marqué ou limité. Effectivement, une décision peut en être une parmi plusieurs (un cas de divorce parmi tant d'autres), sauf que ses faits peuvent être exactement les même critères recherchés par un parti.

Deuxièmement, il est prévu d'effectuer une analyse de l'organisation du corpus par l'entremise d'analyses de segmentation. Essentiellement, le corpus offert par la société pour fins d'analyse contient quatre catégories de textes (Pénal, Responsabilité, Procédure et Assurance). Ces catégories représentent des concepts de niveau supérieur dans le texte, comme un parent. Mais est-ce que les textes que les catégories contiennent peuvent être assez différents pour réobtenir les 4 catégories? Les textes sont des résumés, et se concentrent sur les faits de droit et ceux-ci pour un néophyte du droit peuvent se ressembler beaucoup.

Pour conclure, l'exploration de données textuelles ne peut pas rendre à elle seule un document plus simple d'accès lors d'une recherche, mais le résultat d'une analyse de classification supervisée peut permettre de rendre les résultats plus exhaustifs, car des liens cachés peuvent améliorer l'information que nous en avons. En procédant à une classification, on cherche à construire des ensembles homogènes de concepts, c'est-à-dire partageant un certain nombre de caractéristiques identiques. Puisque le corpus de départ est représentatif du domaine, les classes peuvent être assimilées à des thèmes de pratiques propres à ce domaine. Les termes utilisés en assurance devraient aider à classifier les documents dans le domaine assurantiel.

3.2.2 Étape 2 : Obtenir des données

La démarche de collecte de données n'a pas été très longue, ni fastidieuse. J'ai entrepris ma démarche en communiquant par courriel avec un contact, Me Geneviève Harpin. Celle-ci travaille à titre d'analyste en droit, sous la direction de l'information juridique de la SOQUIJ. Ce contact personnel a transféré ma demande, par courriel, à la conseillère d'affaires juridiques et responsable de l'accès et de la protection de l'information Me Hélène David, avocate. Me David m'a présenté les divers documents électroniques disponibles à la SOQUIJ et permis ainsi de mieux connaître les corpus disponibles. Me David a transféré ma demande au directeur de l'information juridique, Me Daniel Champagne. Nous nous sommes échangés des courriels avant de nous rencontrer le 10 février 2010. Cette rencontre exploratoire a permis d'élaborer sur les attentes de chacun des côtés sur un projet potentiel. Suite à

cette rencontre, Me Champagne m'a obtenu un accès au site « Tout Azimut[1] » pour mieux approfondir le sujet de recherche. Une deuxième rencontre fut faite pour que je puisse présenter le projet de mémoire en administration sur le sujet de l'exploration de textes avec un corpus juridique provenant de la SOQUIJ. Suite à cette rencontre, il y a eu entente sur le corpus et le type de format des données (HTML et document Word). La SOQUIJ m'a demandé de signer une entente de confidentialité concernant les données, ce qui fut fait et envoyé par télécopieur. Le corpus de textes fut reçu par courrier recommandé le 14 juillet 2010. Voici le résumé dans le tableau 6 de ce processus. Les principaux courriels se retrouvent en annexe 6.5.

Tableau 6 : Processus d'acquisition du corpus

Date	Type d'échange	Intervention	Interlocuteur
15/01/2010	Courriel	Demande d'aide auprès de mon contact à la SOQUIJ	Me G. Harpin
19/01/2010	Courriel	Évolution du dossier à la SOQUIJ	Me H. David
25/01/2010	Courriel	Évolution du dossier à la SOQUIJ	Me D. Champagne
05/02/2010	Courriel	Prise de rendez-vous	Me D. Champagne
10/02/2010	Rencontre	Rencontre exploratoire	Me D. Champagne
26/02/2010	Courriel	Accès à *ToutAzimut*	Me D. Champagne
10/05/2010	Courriel	Prise de rendez-vous	Me D. Champagne
25/05/2010	Rencontre	Présentation du projet de TM	Me D. Champagne
15/06/2010	Courriel	Entente sur le type de données	Me D. Champagne
28/06/2010	Courriel	Entente sur le corpus et le format des données	Mme J. Carré
09/07/2010	Courriel - Télécopie	Réception de l'entente de confidentialité SOQUIJ et envoi du document signé de ma part	Me D. Champagne
14/07/2010	Courrier recommandé	Réception du CD de données	

[1] Tout Azimut est le site transactionnel de recherche de résumé de verdict de cour de la SOQUIJ.

3.2.3 Étape 3 : Présenter le corpus

Le besoin d'affaires bien compris, bien explicité, il faut regarder le corpus qui permettra de résoudre la problématique. Le corpus contient beaucoup de données, d'informations et peut devenir très volumineux dans le cadre d'une analyse de TM. Le corpus choisi par SOQUIJ est tout ce qui a paru au JE (journal express) n^{os} 1 à 25 de 2009 et qui a été classé dans les quatre domaines de droit suivant : **Pénal, Responsabilité, Procédure et Assurance**, soit 544 documents. Ce nombre est suffisant pour obtenir une validité conceptuelle et une bonne généralisation. Les définitions des quatre domaines de droit sont les suivantes :

Pénal:

« Cette rubrique regroupe les jugements traitant du Code criminel et de toutes ses lois connexes, du Code de la sécurité routière et du Code de procédure pénale. L'appellation DROIT PÉNAL englobe toutes les poursuites de nature criminelle et de nature pénale. »

Responsabilité:

« Cette rubrique regroupe les jugements traitant des règles de la responsabilité civile extracontractuelle dans le cadre d'actions en dommages-intérêts. On y trouve les décisions expliquant les éléments généraux de la responsabilité (faute, préjudice et lien de causalité) ainsi que des cas d'application, qu'il s'agisse d'action en dommages-intérêts pour des atteintes d'ordre personnel (arrestation injustifiée,

diffamation, voies de fait) ou de responsabilité bancaire, du fait d'autrui, du fabricant, professionnelle, etc. Il y est notamment question des articles 1457 à 1481 C.C.Q. Les jugements concernant la responsabilité des municipalités sont aussi classés à la sous-rubrique MUNICIPAL (DROIT) — responsabilité. Par ailleurs, les jugements traitant de la responsabilité contractuelle sont plutôt classés sous la rubrique couvrant le contrat visé ou sous la rubrique CONTRAT. »

Procédure :

« Cette rubrique regroupe les jugements en procédure civile rendus en vertu du Code de procédure civile, des règles de pratique, du Tarif des honoraires judiciaires des avocats et de différentes lois statutaires. (ex : appel — permission d'appel — rejet de procédure — délai raisonnable — ...) »

Assurance:

«Cette rubrique regroupe les jugements traitant du droit des assurances, particulièrement les articles 2389 à 2628 C.C.Q. Elle couvre notamment les jugements traitant de la Convention d'indemnisation directe pour le règlement des sinistres automobiles et des lois suivantes: la Loi sur les assurances, la Loi sur l'assurance automobile et même que la Loi sur les intermédiaires de marché et la loi qui l'a remplacée, la Loi sur la distribution de produits et services financiers. »

Ces définitions du niveau de détail de l'indexation sont tirées du plan de classification de la SOQUIJ et permettent de mieux comprendre le corpus utilisé pour la présente recherche.

3.2.4 Étape 4 : Évaluer et choisir l'application

La sélection d'un outil peut faire la différence entre un projet réussi et échoué et cette règle s'applique aussi à l'exploration de textes. Dans le cadre de cette étude de type « design science », la sélection de l'outil fait partie de l'expérimentation puisqu'elle en influence le choix du type d'analyse qui sera effectuée sur le corpus et ainsi l'artéfact technologique.

Selon Pouponnot (2009) les applications se valent et la différence se situe au niveau de l'interface, avec une ergonomie plus conviviale, mais elles se distinguent par leur façon d'intégrer les modèles dans le processus de l'entreprise. Selon Spinakis et Chatzimakri (2005), les fonctionnalités offertes par chacun des logiciels de TM peuvent couvrir les besoins d'un projet d'analyse de textes. Ils ne peuvent suggérer un seul outil qui convient à tous les types de projets d'exploration de textes. L'exploration de données textuelles se pratique déjà en entreprise et donc certaines applications ont été développées pour des besoins spécifiques de certains clients. Puisqu'il serait coûteux, en temps et en argent, de sélectionner le mauvais outil, l'utilisation d'un cadre d'évaluation des outils d'exploration de données préexistant est sélectionnée.

Collier et al. (1999) fournissent une méthode pour le choix du meilleur logiciel versus un problème particulier. Son cadre est conçu pour accueillir différents environnements et problématiques. Ce cadre est repris et adapté au besoin de ce mémoire. L'expérience de Collier démontre qu'il n'y a pas un outil d'exploration qui

convient pour tous les cas. Le cadre se base sur l'utilisation de tests sur les applications. Je ne peux faire ces tests et je me base donc sur l'information fournie par les vendeurs, revendeurs et site web des applications pour effectuer les différentes matrices. Il va aussi donc de soi que toutes les matrices de Collier ne peuvent pas être effectuées avant l'achat. De plus, nous devons adapter Collier, car la méthode d'évaluation et de sélection est faite pour l'exploration de données et dans notre cas nous allons une étape plus loin, nous utilisons des données textuelles. Cette adaptation est effectuée avec les critères d'analyse des études de Spinakis et Chatzimakri (2005) et de Quatrain et al. (2004).

Le cœur du cadre d'évaluation est un outil de notation. L'expérience de Collier présente quatre catégories de critères d'évaluation des outils d'exploration de données**:** **la performance**, **la fonctionnalité, la convivialité et le travail de soutien**. Ces catégories forment la base du cadre d'évaluation mais suite à l'évaluation des critères, nous utiliserons que les critères de **fonctionnalité** et de **travail de soutien**. Je n'ai pas essayé les applications, il est donc impossible d'évaluer la performance et la convivialité. Plus de renseignements sur les critères en annexe 6.6.

La notation se fait par rapport à une référence, une application fétiche. En règle générale, l'évaluateur est prédisposé, selon ses heuristiques, vers une application pour une variété de raisons subjectives. Ce «favori» doit être sélectionné comme l'application de référence. Les autres applications sont ensuite évaluées par rapport à l'outil de référence pour chaque critère en utilisant la notation sur l'échelle discrète qui suit:

Performance relative	**Classement**
Bien pire que l'outil de référence	*1*
Pire que l'outil de référence	*2*
Identique à l'outil de référence	*3*
Mieux que l'outil de référence	*4*
Beaucoup mieux que l'outil de référence	*5*

L'application du cadre s'effectue en phases qui sont les suivantes:

1. Présélection des applications
2. Identification des critères de sélection supplémentaires
3. Poids des critères de sélection
4. Correcteur
5. Notation de l'évaluation
6. Évaluation et sélection de l'application

3.2.5 Étape 5 : Effectuer l'exploration de textes

Chacune des applications commerciales de TM vient avec sa méthodologie. Les différentes méthodologies sont présentées en annexe 6.4. Les différentes méthodologies se ressemblent et s'équivalent. Ainsi, la sélection de l'outil sélectionnera aussi la méthodologie qui en découle. Cette méthodologie sera encadrée par les besoins spécifiques de la recherche scientifique. Voici les principales étapes à suivre :

1. Préparation du corpus
    a. Sélectionner, créer l'ensemble des données requises
    b. Nettoyer les données et de prétraitement
    c. Réduire les données
2. Échantillonnage

a. Une table d'apprentissage pour construire les modèles
b. Une table de test pour tester les modèles sur des données différentes de celle sur lesquels ils ont été construits
c. Une table de validation pour optimiser les modèles

3. Les méthodes de TM correspondent aux objectifs?
   a. Sélectionner le type d'analyse et de modèle
4. Effectuer le TM
   a. Classification supervisée
   b. Analyse par segmentation
   c. D'autres types d'analyse (si offerts par l'outil)
5. Résultats
6. Analyser, interpréter et vulgariser les résultats explorés
   a. Refaire le processus si nécessaire
7. Agir sur les connaissances découvertes

3.3 Éthique de ce mémoire

Avant d'effectuer l'analyse de données textuelles, il faut glisser quelques mots sur l'éthique de cette recherche et pourquoi elle ne nécessite pas d'approbation par le comité d'éthique.

Un appel téléphonique auprès de Carole Coulombe, Coordonnatrice du CÉR lettres et sciences humaines, a été effectué concernant des questions sur le « Formulaire de soumission d'un projet de recherche pour évaluation par le comité d'éthique ». Selon cette dernière, le mémoire actuel ne nécessite pas l'évaluation d'un comité d'éthique

puisque les données sont d'origine publique et elle se base sur la proposition de la 2e édition de l'Énoncé de politique des trois Conseils : *« Éthique de la recherche avec des êtres humains (EPTC) décembre 2009* ». Nous retrouvons les articles soutenant ce point à l'article 2.2 en annexe 6.7.

Ce chapitre présente la méthodologie qui est utilisée dans le cadre de cette recherche. Tout d'abord, la base de recherche est le « design science » sur laquelle est juxtaposé un amalgame des processus de TM (provenant de trois auteurs) et l'utilisation de la méthodologie de TM de l'application qui sera sélectionnée plus tard. Le chapitre suivant est consacré à l'analyse de données textuelles d'un corpus provenant de la SOQUIJ.

4 Analyses et résultats

L'analyse des résultats doit suivre plusieurs étapes des grandes méthodologies d'exploration, du moins celle qui sera retenue. La première étape est de revoir le terrain, l'ensemble des données de la SOQUIJ et leurs problématiques, le présenter et bien comprendre les données obtenues de la SOQUIJ. Il faut évaluer et choisir l'application appropriée d'exploration de textes avec les connaissances, les besoins et les données. L'application choisie oriente ensuite le choix et l'utilisation d'une méthodologie spécifique. En respectant cette méthodologie, nous effectuerons l'exploration de textes en incluant des analyses statistiques, des résultats et des descriptifs. J'ai bien hâte de décrire l'artéfact technologique de la recherche et donc le travail d'exploration de textes dans le domaine juridique francophone, de son processus, de ses défis et bénéfices.

Tel que décrit dans la section précédente, le corpus choisi par SOQUIJ est tout ce qui a paru au Journal Express (JE), n^{os} 1 à 25 de 2009 et qui a été classé dans les quatre domaines de droit suivant : **Pénal, Responsabilité, Procédure et Assurance**, soit un total de 544 documents.

4.1 Évaluer et choisir l'outil d'analyse de textes

Comme mentionné précédemment à la section de méthodologie, l'évaluation et la sélection d'un outil sont cruciales pour un projet. L'utilisation de la méthodologie décrite par Collier en six (6) étapes est utilisée afin de sélectionner l'application appropriée.

4.1.1 Étape 1 : Présélectionner les applications

Plusieurs applications ont été sélectionnées par le biais du web, de forum de discussions et de suggestions des pairs. En tout, 19 applications d'exploration de données textuelles, dont 15 applications commerciales et 4 applications « open source » ont été analysées dans la présente étude (voir annexe 6.8 pour plus de renseignements sur les applications sélectionnées).

Ces applications ne correspondent pas toutes aux besoins et ainsi l'élimination de certaines est effectuée dû à différentes contraintes. Ces contraintes peuvent être organisationnelles, applicatives ou autres, telle que la non-représentation par le vendeur. Prenons par exemple, les 4 applications « open source ». Elles n'offrent peu ou pas de support et formation et sont éliminées puisque le support est un élément

important pour moi. Dans les applications commerciales, plusieurs ne rencontrent pas certaines contraintes importantes au contexte de la présente recherche:

- Un fournisseur d'application n'offrait pas le français comme langue;
- Un mentionne qu'il en coûte dans les 6-7 chiffres pour l'installation pour s'arrimer aux bases de données;
- Un autre mentionne qu'il est trop cher pour nous et ce, sans plus d'explication;
- Un doit fonctionner de concert avec un système de gestion intégré (ERP);
- Trois sont très spécialisés et ce, dans un segment de marché autre que le légal, par exemple les banques ou le web;
- Des huit applications restantes, quatre n'ont pas soumissionné.

Il ne restait donc que quatre possibilités d'outils qui sont, en ordre alphabétique : « IBM SPSS, SAS, Statsoft et Temis ».

4.1.2 Étape 2 : Identifier les critères de sélection supplémentaires

Il est important de m'assurer que le choix de l'outil réponde aux besoins particuliers de l'exploration de textes en français dans le but de cette recherche. Bien que le cadre d'évaluation de Collier fournisse la plupart des critères techniques de sélection, le principal objectif de cette étape est pour moi d'identifier les critères supplémentaires qui sont spécifiques à ce projet. Le coût de l'application pour une raison évidente doit être rajouté. De plus, il doit soutenir un système d'exploitation Windows 7 / 64 bits sur un ordinateur portable.

4.1.3 Étape 3 : Pondérer les critères de sélection

À cette étape il faut pondérer les catégories ainsi que les critères de sélection. Chacun des critères des sections de fonctionnalité, travail de soutien et critères supplémentaires identifiés à l'étape 2 se voit attribuer une note. Au cours de l'étape 3, les critères de chaque catégorie sont pondérés. (voir tableau 8 pour la pondération). J'ai distribué la pondération assez uniformément au travers des critères.

4.1.4 Étape 4 : Corriger les critères

Maintenant que les critères ont été pondérés par rapport aux besoins ciblés, les applications peuvent maintenant être comparées. Plutôt que de noter sur une échelle absolue, la notation se fait par rapport à une référence, une application fétiche. Ici, l'application référence que j'ai préférée, est le logiciel Luxid de la compagnie Temis. Cette application semble avoir un avantage avec ses antécédents dans le domaine juridique et son représentant parle français.

4.1.5 Étape 5 : Évaluer de notation

La méthodologie d'évaluation de Collier est conçue pour expliciter le processus de sélection mais, selon ce dernier, l'intuition de l'évaluateur, et donc le chercheur, ne doit pas être complètement ignorée. Les écarts entre les résultats et l'intuition sont généralement dus à des coefficients incorrects de critères. Si un tel écart existe, l'étape 5 consiste à examiner les pondérations attribuées aux critères de sélection et de les ajuster si nécessaire et les ajustements ont été effectués afin d'accorder de l'importance au coût, au « clustering » et à la catégorisation.

4.1.6 Étape 6 : Évaluer et sélectionner l'application

L'utilisation d'un tableau facilite la présentation, mais les calculs ont été effectués dans un tableur de manière à faciliter la compilation des pointages des différentes applications. Les fournisseurs qui participent sont « IBM SPSS, SAS, Statsoft et Temis ». Leurs produits respectifs sont : « IBM® SPSS® Modeler Premium, SAS® Text Analytics, STATISTICA® Data Miner Academic Bundle et Luxid® de Temis. Comme mentionné plus tôt, l'application de référence est Luxid® de Temis. L'application gagnante est donc IBM® SPSS® Modeler Premium telle que présentée dans le tableau 8.

Le critère qui a fait basculer le choix est le coût par licence, ce qui est présenté dans le tableau 7.

Tableau 7 : Coût d'une licence d'application d'exploration de textes

Application	Coût
Temis Luxid®	26 500,00 $
IBM® SPSS® Modeler Premium	350,00 $
SAS® Text Analytics	550,00 $
STATISTICA® Data Miner Academic Bundle	775,00 $

La sélection d'IBM SPSS Modeler Premium est aussi primée dans la littérature. Selon l'expérience de Crowsey et al. (2007), suite à l'exécution d'un échantillon du corpus par des suites logicielles et en comparant la facilité d'utilisation des logiciels, deux produits se sont démarqués: SAS et SPSS. SPSS s'est avéré être le plus rapide

lors de l'exécution des fichiers. SPSS est également un bon choix pour le traitement de grandes quantités de documents. Il est plus efficace si les données ont été prétraitées et ainsi transformées de non-structurées à semi-structurées. Ce qui conforte le choix de SPSS.

Tableau 8: Comparaison des applications d'exploration de textes

Critères	Poids	Temis	SPSS	SAS	Statistica
Fonctionnalité					
Algorithmique variée	0,04	3	3	3	3
Méthodologie prescrite	0,04	3	3	3	3
Validation du modèle	0,04	3	3	3	3
Flexibilité du type de données	0,04	3	2	3	3
Algorithme modifiable	0,04	3	3	3	3
Données d'échantillonnage	0,04	3	3	3	3
Rapport	0,04	3	3	3	3
Exportation du modèle	0,04	3	2	3	3
Total	**0,32**	**0,32**	**0,29**	**0,32**	**0,32**
Travail de soutien					
Nettoyage de données	0,03	3	3	3	3
Valeur de substitution	0,03	3	3	3	3
Filtrage des données	0,04	3	3	3	3
Attributs	0,04	3	3	3	3
Randomisation	0,03	3	3	3	3
Suppression d'enregistrement	0,04	3	3	3	3
Manipulation des vides	0,04	3	3	3	3
Manipulation des métadonnées	0,03	3	2	3	1
Résultat d'évaluation	0,04	3	3	3	3
Sous-Total	**0,32**	**0,32**	**0,31**	**0,32**	**0,30**
Autres critères					
Coût	0,15	3	5	4	4
Segmenter	0,02	3	1	1	1
Tokéniser	0,02	3	2	2	1
Analyser morphologiquement	0,02	3	2	2	1

Désambiguïser les mots	0,02	3	2	2	1
Interroger un dictionnaire	0,02	3	2	1	1
Reconnaître des expressions idiomatiques	0,02	3	1	1	1
Windows 7 64 bits	0,03	3	4	4	4
Clustering	0,05	3	3	3	3
Catégorisation	0,05	3	3	3	3
Sous-Total	**0,36**	**0,36**	**0.42**	**0.36**	**0,34**
Total	**1,00**	**1,00**	**1,02**	**1,00**	**0,96**

4.2 Effectuer l'exploration de textes

L'application retenue est IBM® SPSS® Modeler Premium et la méthodologie qui lui est associée est CRISP-DM. En annexe 6.4, les différentes méthodologies sont présentées. Elles sont très semblables et il est donc normal, simple et pratique d'appliquer la méthodologie que l'application et que les gens de l'industrie utilisent. Selon le sondage de KDnuggets « *What main methodology are you using for data mining*? (2002) » 51 % des répondants utilisent la méthodologie CRISP-DM. Le CRISP-DM est une méthode en six étapes qui aide à structurer la section d'exploration de textes (figure 7).

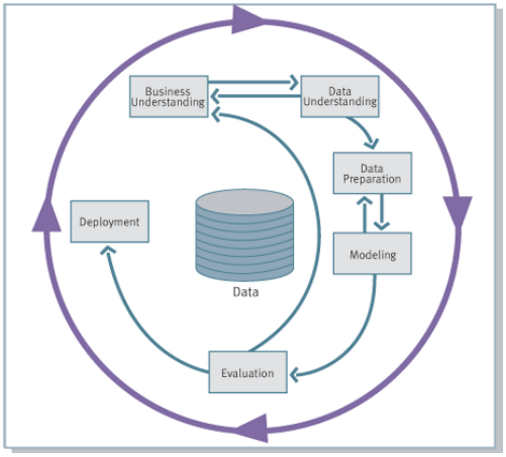


Figure 7 : Shearer, C. (2000). Modèle CRISP-DM

4.2.1 Comprendre les données et le problème d'affaires

Nous avons déjà observé auparavant la problématique de la SOQUIJ et ses données. La première étape selon la méthodologie CRISP-DM est de comprendre le métier. Comprendre le métier selon SPSS c'est de bien définir le problème. De cette définition, il faut définir quels sont les objectifs de l'analyse statistique vis-à-vis le problème d'affaire, ce qui a été bien explicité. Regardons ce corpus un peu plus attentivement.

4.2.2 Préparer le corpus (Préparer les données)

Les données collectées doivent être préparées en fonction des algorithmes qui seront utilisés. Le corpus a été gracieusement offert par la SOQUIJ, il a été fourni sur un média CD qui contenait les 544 documents. La première activité, extraire le corpus du CD et le charger dans un dossier accessible à l'application.

Il y a peu de nettoyage à faire dans le cadre du TM. Dans le cadre de la recherche, j'ai fait le choix des textes en format HTML au lieu des documents Word. Les documents HTML étaient déjà scindés en unité et ainsi ils sont plus simples à utiliser pour la segmentation de l'échantillonnage.

Les 544 documents ont été visualisés afin d'extraire les documents anglophones. Ainsi, l'outil de visualisation de Microsoft Explorer a permis de retirer 40 documents anglophones. Durant cette visualisation, certains documents, plus courts, sous forme de procès-verbaux ont été identifiés. Ils sont aussi exclus de l'analyse puisque la

majorité d'entre eux ne contiennent aucune donnée intéressante et seulement les minutes de rencontres de dossier de cour. Des 504 documents restants, 45 procès-verbaux sont retirés de l'analyse et il restera que 459 documents pour l'analyse statistique. Comme mentionné plus tôt ils proviennent de 4 domaines de droit. Les 459 documents HTML ne sont pas distinctement classifiés dans les quatre domaines. Ainsi, j'ai créé un fichier Excel pour classifier chacun des documents au sein d'une catégorie des domaines **Pénal, Responsabilité, Procédure et Assurance**. Les définitions tirées du plan de classification de la SOQUIJ présentées plus tôt permettent cette classification. Aucune validation externe de cette classification n'a été effectuée et cela peu erronée certains résultats. Ce fichier avec la classification permettra à certains algorithmes un apprentissage de classification. Les données sont maintenant prêtes à être traitées. SPSS appelle l'activité subséquente : « Modéliser ». Elle inclut l'échantillonnage, la sélection des analyses, l'analyse et l'obtention des résultats. L'activité de modélisation est normalement itérative avant d'arriver aux résultats escomptés et présentés.

4.2.3 Effectuer l'exploration de textes (Modéliser)

Afin de simplifier la compréhension, seules les étapes critiques et essentielles de la modélisation seront décrites. Comme mentionné dans la méthodologie, l'exploration de textes analysés sera celle de liens, des concepts, suivis par l'analyse de segmentation pour conclure sur l'analyse de classification supervisée. Un supplément sur les algorithmes utilisés est en annexe 6.9.

La préparation d'un dictionnaire efficace peut prendre beaucoup de temps. Puisque ma spécialité n'est pas le droit mais bien l'administration, je préfère ne pas créer un dictionnaire qui devra être vérifié par un spécialiste du domaine. Je suis conscient que l'absence de ce dernier influence les résultats. Selon Turban et al. (2007), une des erreurs de l'exploration de données est de laisser insuffisamment de temps pour la préparation des données.

La méthodologie CRISP-DM est itérative et normalement, cette étape est une étape d'action et d'application, elle ne comprend pas l'analyse des données obtenues. Afin de simplifier la lecture du mémoire, l'analyse est faite immédiatement à la suite de la modélisation puisque les différentes itérations d'optimisation sont déjà effectuées.

4.2.3.1 Échantillonnage

L'application permet de partitionner, ce qui permet de séparer les données en deux (apprentissage et test) ou trois (apprentissage, test et validation) échantillons. L'approche retenue dépendant de l'analyse à effectuer, l'utilisation des partitions sera ou ne sera pas utilisée. L'utilisation de l'échantillonnage réduit la taille du corpus à étudier et comme par exemple le corpus du domaine de droit « assurance » ne contient que 51 documents, il n'est pas avisé de le réduire pour analyser les liens du texte de ce domaine. À l'opposé, lors de l'analyse de classification supervisée sur le corpus complet de 459 documents, il est nécessaire de valider les modèles et d'utiliser les partitions. Donc, lors des différentes analyses, un commentaire sera présent sur

l'utilisation ou non de la partition. Un document peut contenir à lui seul plus d'une centaine de pages et ainsi des milliers de mots et ainsi aussi ralentir le traitement.

4.2.3.2 Analyse des liens du texte

Cette analyse de statistique descriptive permet d'identifier les relations entre les concepts dans les données de texte. Les relations et les associations sont identifiées et extraites par la découverte de modèles au sein des données texte. L'analyse des liens peut se faire à différents niveaux, soit par exemple au niveau des domaines de droit individuellement, soit au niveau du corpus complet.

La première analyse, celle des liens du texte sur un domaine de droit permet de mieux comprendre le domaine et ainsi, en aval, aider l'analyse de segmentation de ce domaine.

Une première sélection est effectuée pour utiliser les données d'un seul domaine. Les données sélectionnées proviennent du domaine de droit « assurance » (DomaineA = C / DomaineR = Assurance). Il n'y a pas d'échantillonnage effectué puisque le corpus est limité et que le besoin est d'obtenir les concepts d'un domaine pour un usage ultérieur de segmentation. La sélection du corpus parmi les quatre est basée sur les connaissances générales du chercheur qui a déjà travaillé dans le domaine assuranciel. Deux méthodes s'offrent à l'explorateur dans l'application SPSS : « interactive et directe ». La méthode interactive (flux TLA) permet à l'explorateur de reclasser les concepts extraits par le nœud précédent à l'intérieur de catégories prédéterminées par l'application ou bien de catégories créées par ce dernier. Ces

nouvelles catégories peuvent alors dans la session interactive être utilisées pour visualiser leurs interrelations. Par exemple, dans le domaine de droit « assurance » les catégories : « assureur, assuré, demandeur, défendeur et tribunal » pourraient être créées et contenir des concepts extraits par le nœud d'analyse de textes. Cette analyse requiert une connaissance du domaine poussé afin de bien regrouper sous les bonnes catégories les concepts extraits.

Une comparaison entre la méthode avec méthode directe sans le nœud d'analyse des liens du texte et avec a été effectuée. Sur les 500 concepts extraits, 71 étaient différents et ce, puisque le nœud d'analyse permet une sélection du type de concepts extraits, ainsi 68 concepts autres « unknown » ont été extraits par la méthode directe (6 currency, 44 date, 7 location, 6 organization, 5 person). Les trois concepts « unknown » après une analyse se retrouvent extraits différemment, par exemple : « location crédit ford » est un concept pour l'un et « location » et « crédit ford » deux concepts pour l'autre. Une sélection peut être effectuée pour éliminer les concepts de type : « date, location, nom de personne, etc. ». Ces concepts n'ont pas de valeurs ajoutées dans le cadre de ce mémoire puisque je m'intéresse au domaine en général, mais une utilisation temporelle, géo spatiale et autre des données pourrait avoir une valeur ajoutée aux documents pour certains clients. Le seul type conservé est « unknown ». Ce type de concept existe puisque le dictionnaire de base de l'application n'a pas de classification pour les termes francophones et est ainsi classifié comme « unknown ». L'avenue de la méthode directe, qui extrait et groupe

automatiquement les concepts, est donc retenue en ayant en amont un nœud d'analyse des liens du texte et de sélection afin d'affiner la recherche.

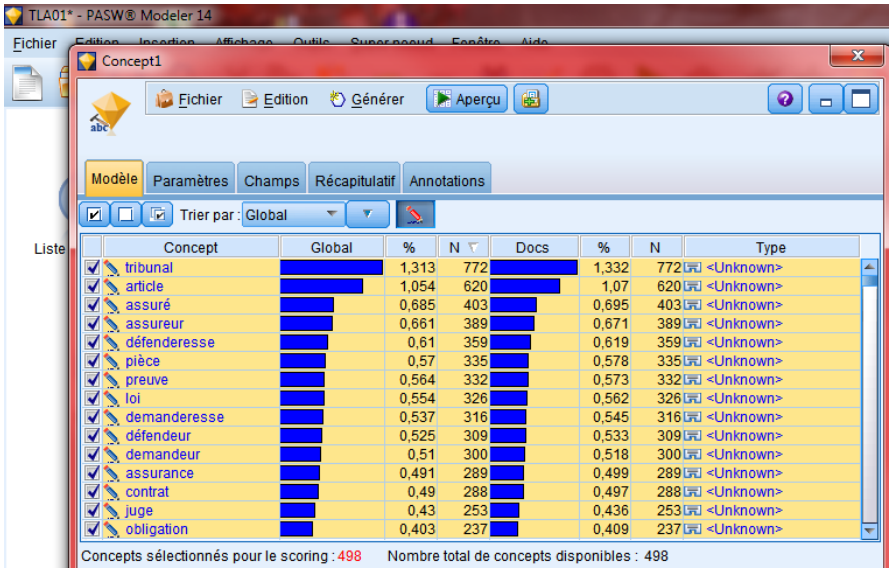


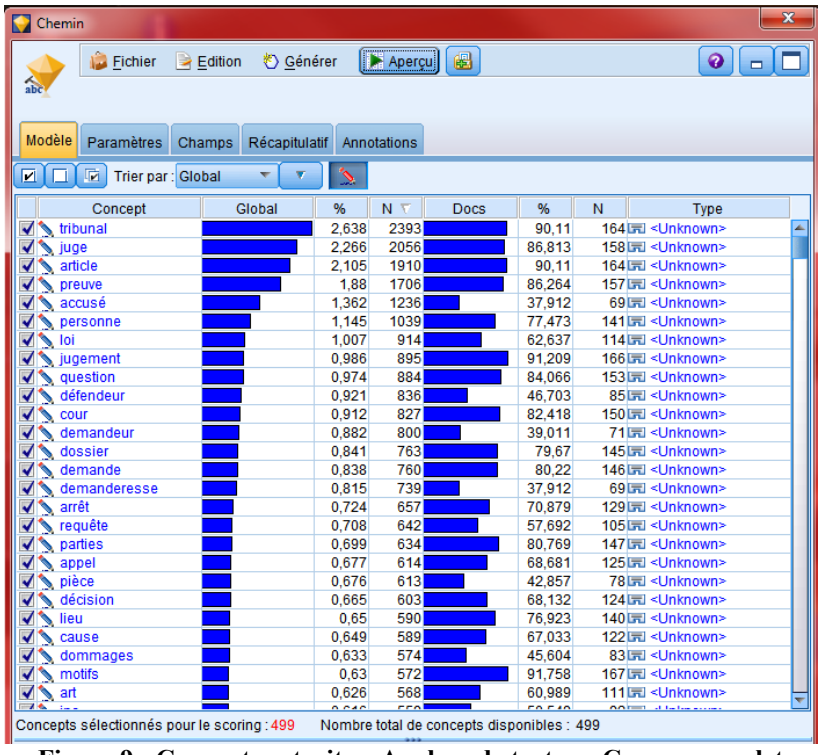
Figure 8 : Concepts extraits – Analyse de textes – Domaine Assurance

L'analyse et l'interprétation de l'analyse des liens du texte ont déjà débuté précédemment afin de s'assurer de la qualité des données pour les étapes d'analyse de classification supervisée et de segmentation. Puisque l'analyse requiert une connaissance du domaine poussé afin de bien regrouper sous les bonnes catégories les concepts extraits. Nous allons explorer les résultats à un haut niveau.

Regardons, les trois concepts ayant la plus grande fréquence dans le domaine de droit assurance (figure 8). Les trois plus grands concepts sont : tribunal (N : 772 et % 1.313), article (N : 620 et % 1.054) et assuré (N : 403 et % 0.685). Les deux premiers concepts sont reliés au droit tandis que le troisième, le concept assuré, est fort peu étonnant pour le domaine de l'assurance. L'assureur et l'assuré sont lié par contrat, et

il semble normal que soit le défendeur ou le demandeur soit l'un ou l'autre des parties en cause. Continuons avec le corpus complet.

La seconde analyse des liens du texte, sur le corpus complet, permet de comprendre les concepts de droit. Une première opération est effectuée pour partitionner les données puisque le corpus complet est utilisé. Le besoin est de ressortir les concepts pour l'analyse de classification supervisée. Pour l'analyse, les deux méthodes présentées précédemment s'offrent encore à l'explorateur soit : « interactive et directe ». L'avenue de la méthode directe est réutilisée pour les mêmes raisons que l'analyse précédente.



Concept	Global	%	N	Docs	%	N	Type
tribunal		2,638	2393		90,11	164	<Unknown>
juge		2,266	2056		86,813	158	<Unknown>
article		2,105	1910		90,11	164	<Unknown>
preuve		1,88	1706		86,264	157	<Unknown>
accusé		1,362	1236		37,912	69	<Unknown>
personne		1,145	1039		77,473	141	<Unknown>
loi		1,007	914		62,637	114	<Unknown>
jugement		0,986	895		91,209	166	<Unknown>
question		0,974	884		84,066	153	<Unknown>
défendeur		0,921	836		46,703	85	<Unknown>
cour		0,912	827		82,418	150	<Unknown>
demandeur		0,882	800		39,011	71	<Unknown>
dossier		0,841	763		79,67	145	<Unknown>
demande		0,838	760		80,22	146	<Unknown>
demanderesse		0,815	739		37,912	69	<Unknown>
arrêt		0,724	657		70,879	129	<Unknown>
requête		0,708	642		57,692	105	<Unknown>
parties		0,699	634		80,769	147	<Unknown>
appel		0,677	614		68,681	125	<Unknown>
pièce		0,676	613		42,857	78	<Unknown>
décision		0,665	603		68,132	124	<Unknown>
lieu		0,65	590		76,923	140	<Unknown>
cause		0,649	589		67,033	122	<Unknown>
dommages		0,633	574		45,604	83	<Unknown>
motifs		0,63	572		91,758	167	<Unknown>
art		0,626	568		60,989	111	<Unknown>

Figure 9 : Concepts extraits – Analyse de textes – Corpus complet

Observons les trois concepts ayant la plus grande fréquence dans les domaines de droit confondus (figure 9). Ils sont : tribunal (N : 2393 et % 2.638), juge (N : 2056 et

% 2.266) et article (N : 1910 et % 2.105). Ces trois concepts, et même subséquent, sont tous reliés au droit en général. Même si nous continuons la liste, les concepts les plus fréquents semblent tous venir du jargon juridique.

Nous obtenons aussi des résultats semblables entre les deux séries (corpus complet et domaine assurance). Dans les 15 premiers concepts, il y en a 8 communs : « tribunal, article, preuve, loi, demanderesse, défendeur, demandeur et juge ». Ce sont donc des concepts forts dans les textes. Les 7 autres concepts, bien que différents, se ressemblent. Nous pouvons remarquer les parties impliquées : « assuré, assureur et défenderesse » de l'analyse du corpus assuranciel et « accusé et personne » du corpus global. De plus, le litige ressort des deux côtés : « pièce, assurance, contrat et obligation » pour l'assurance et « question, dossier et demande » pour le corpus complet. Ces ressemblances dans la fréquence des concepts provient peut être de la structure des résumés et de l'énoncé des faits.

Ces analyses apportent peu d'information sans la présence d'un dictionnaire, nous ne pouvons pas faire ressortir de sentiment, de domaine ou autres. Continuons par l'analyse par segmentation.

4.2.3.3 Analyse par segmentation

L'analyse par segmentation peut améliorer l'efficacité de la catégorisation de textes en identifiant des groupes d'enregistrements similaires et en répertoriant les enregistrements en fonction du groupe auquel ils appartiennent. Le but de l'analyse

par segmentation est de permettre à la SOQUIJ d'avoir une classification, peut-être nouvelle, des textes et :

- de discerner les thèmes généraux de chaque domaine de droit afin de regrouper automatiquement dans une classe des textes similaires d'un domaine particulier
- de regrouper automatiquement les 459 textes en groupes distincts de documents similaires

Ainsi, cette analyse de segmentation permet d'organiser différemment les textes et ce, autant au sein des domaines de droit ainsi qu'au sein de tous les documents. Des concepts similaires forment un segment (cluster). Les segments sont généré par des algorithmes de classification non supervisée qui se fondent sur la fréquence d'apparition de ces concepts dans l'ensemble de documents et la fréquence d'apparition conjointe des concepts dans le même document (ou cooccurrence). Chaque concept d'une segmentation (cluster) est cooccurrent avec au moins un autre concept de la segmentation (cluster). L'analyse de segmentation (clustering) a pour objectif de regrouper les concepts similaires, alors que les catégories ont pour objectif de regrouper les documents ou les enregistrements en fonction des correspondances existant entre le texte et les descripteurs (concept, règles, patrons) pour chaque catégorie. Un « cluster » adéquat en est un qui présente des concepts fortement similaires et fréquemment cooccurrents, ainsi que dotés de peu de liens vers des concepts d'autres segments (clusters).

L'analyse est effectuée sans l'aide des connaissances existantes sur les classes et leurs caractéristiques. On ne connaît pas le nombre de classe recherchée ni espérée avec ce type d'analyse, mais vrai à dire, en exploration de données il est difficile de prévoir le résultat avant d'avoir effectué l'analyse. Le nombre de classe peut-être exigé par l'explorateur, mais dans le cadre de cette recherche, je n'ai pas trouvé comment et ce autant sur les forums que dans l'aide de l'application.

Trois méthodes de classification sont fournies par SPSS PASW Modeler et elles sont : « le Kohonen, le K-means et le Twostep» (figure 10).

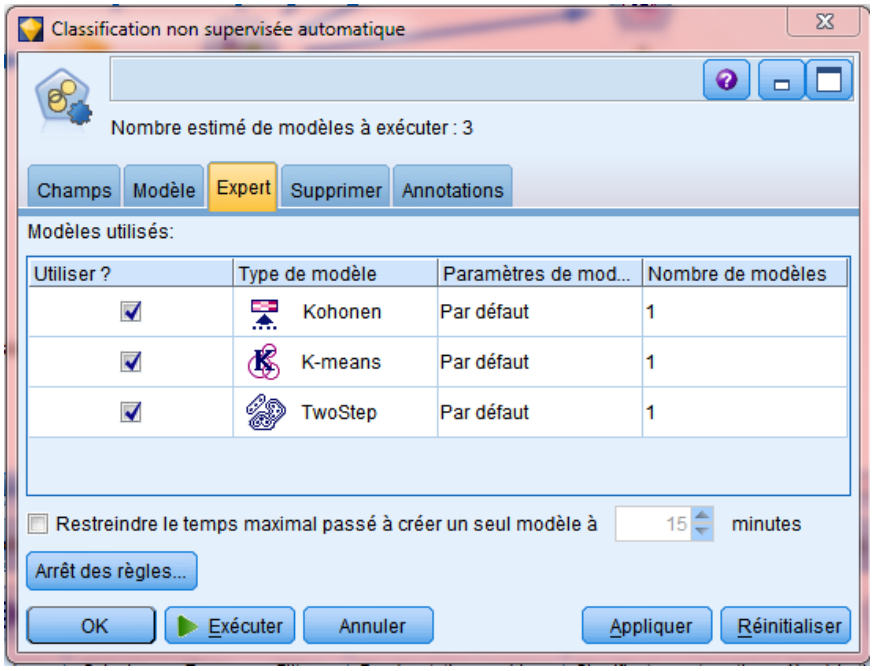


Figure 10 : Analyse de segmentation – Les méthodes fournies par Modeler

La méthode « Kohonen »

Ainsi, le nœud le nœud Kohonen a été ajouté à la fin de l'analyse des liens du texte du domaine de droit « Assurance » (figure 11) et au corpus complet (figure 12).

Les figures 11 et 12 montrent qu'on a obtenu des solutions avec de multitudes de classes. Le coefficient de silhouette permet d''interpréter et valider l'analyse de segmentation. Le coefficient de silhouette n'est pas très élevé, mais au moins se retrouve dans la zone acceptable. Il est donc difficile d'obtenir une solution avec les quatre (4) catégories naturelles. Malgré qu'en segmentation, nous toujours imposer le nombre de classes voulues, je désirais voir le classement naturel. Nous pourrions aussi continuer à explorer plusieurs solutions du Kohonen, mais regardons le K-Means.

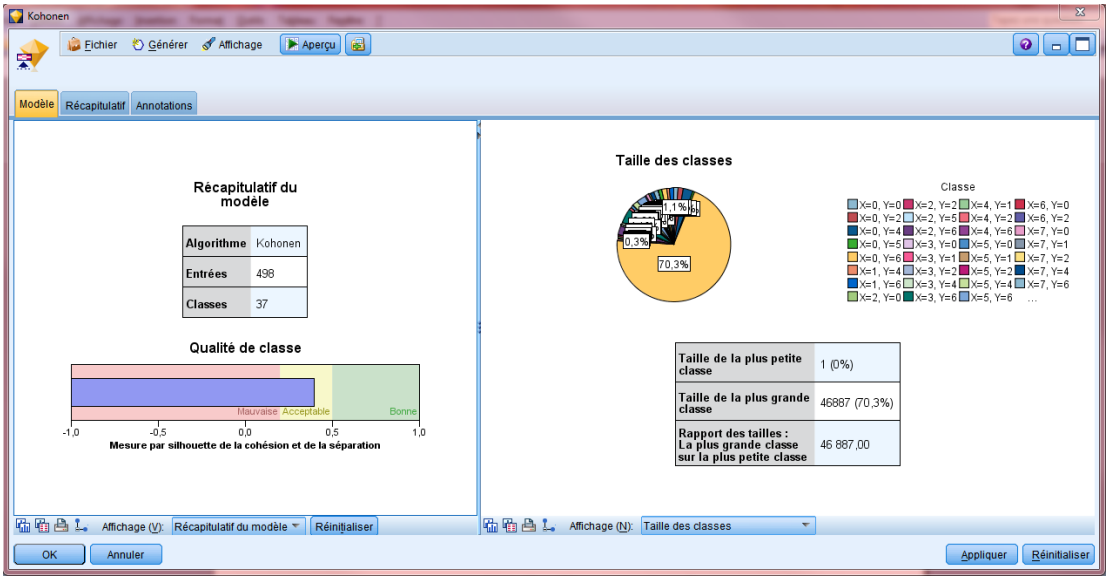


Figure 11 : Analyse de segmentation – Kohonen – Un domaine de droit

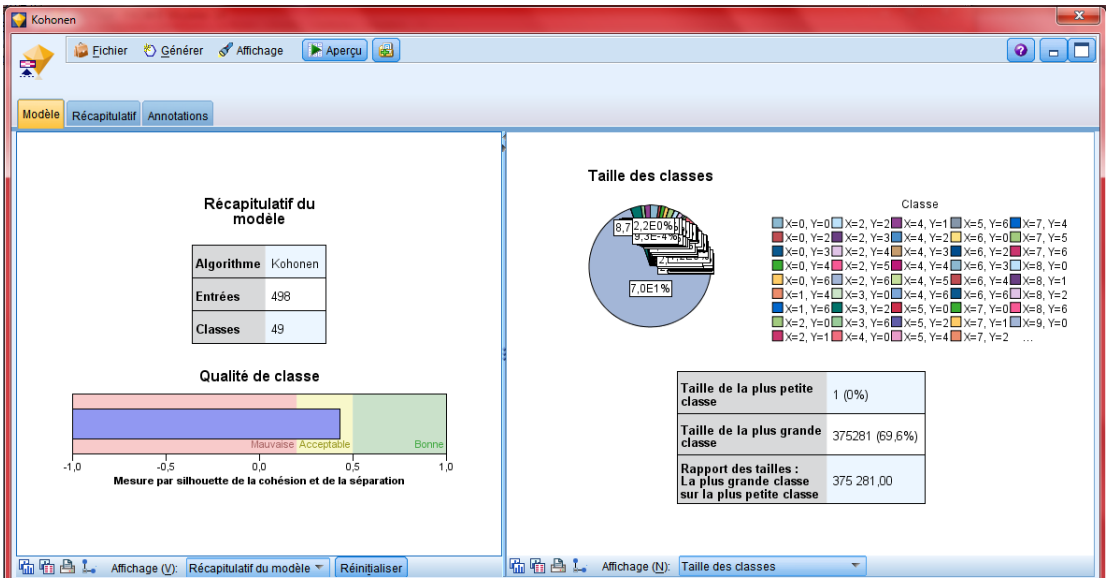


Figure 12 : Analyse de segmentation – Kohonen – Corpus complet

La méthode « K-means »

De la même manière que le Kohonen, le nœud K-Means a été ajouté à la fin de l'analyse des liens du texte du domaine de droit « Assurance » (figure 13) et au corpus complet (figure 14)

Les figures 13 et 14 montrent qu'on a obtenu des solutions à 3 et 4 classes. Malgré que le coefficient de silhouette soit un peu élevé, il se retrouve de plus dans la zone bonne, il serait difficile d'obtenir une solution avec les quatre (4) catégories naturelles puisque plus de 98 % des documents se retrouvent dans la même classe. Allons maintenant voir le Two-Steps.

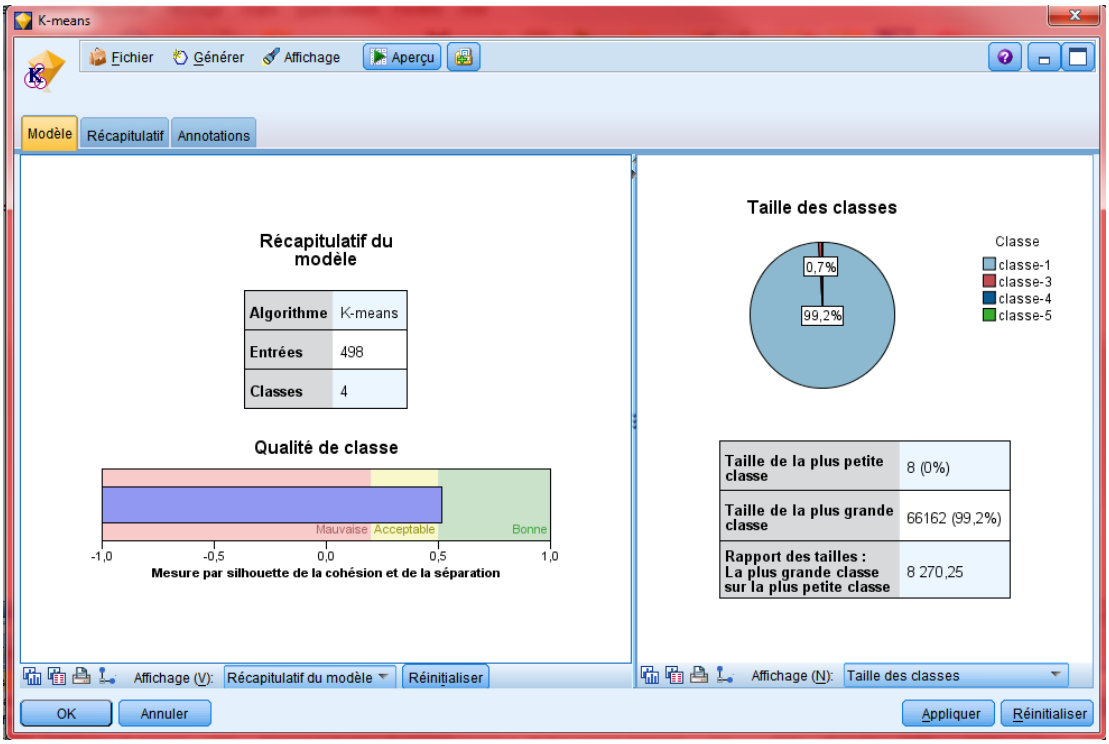


Figure 13 : Analyse de segmentation – K-means – Un domaine de droit

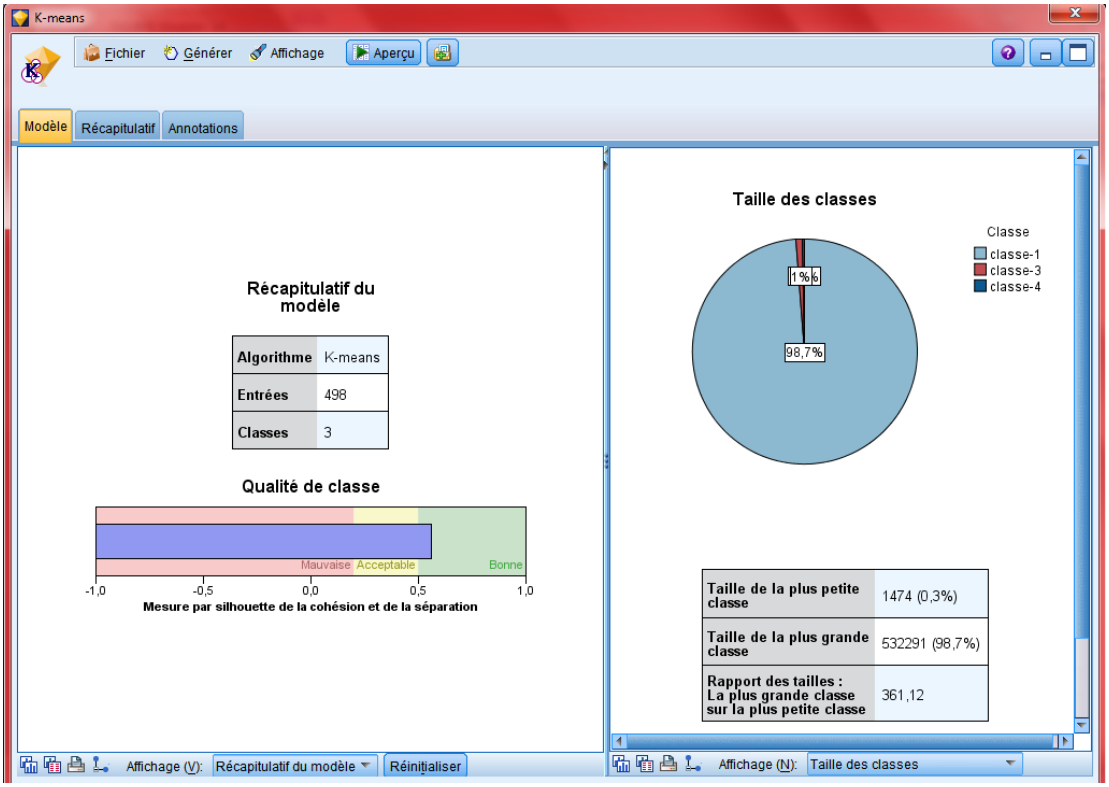


Figure 14 : Analyse de segmentation – K-means – Corpus complet

La méthode « Twostep »

De la même manière que les deux méthodes précédentes, le nœud été ajouté à la fin de l'analyse des liens du texte du domaine de droit « Assurance » (figure 15) et au corpus complet (figure 16)

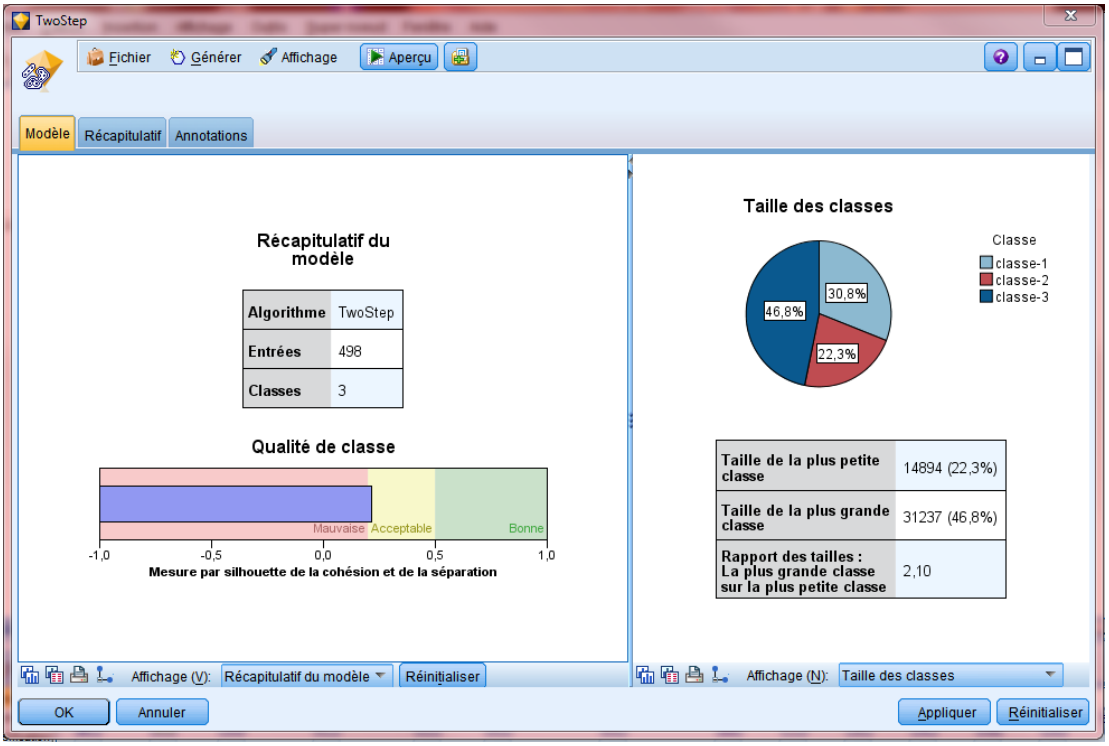


Figure 15 : Analyse de segmentation – TwoStep – Un domaine de droit

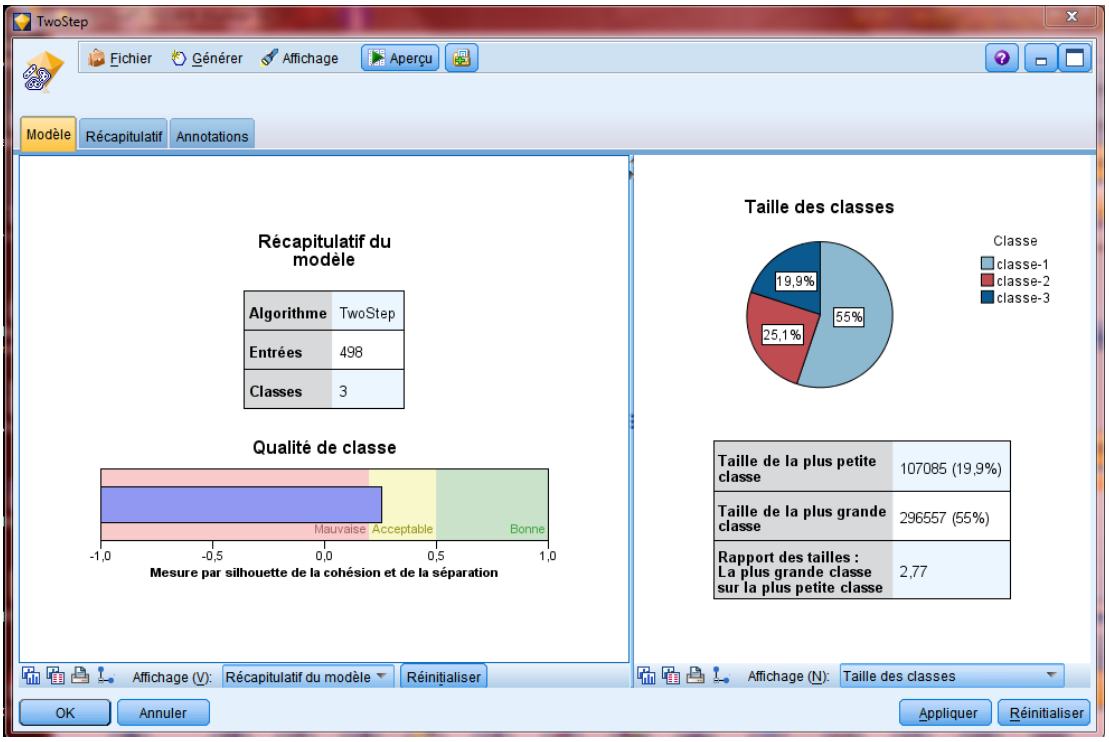


Figure 16 : Analyse de segmentation – TwoStep – Le corpus complet

Les figures 15 et 16 montrent qu'on a obtenu des solutions à 3 classes. Les trois classes ont des volumes intéressants. Le coefficient de silhouette n'est pas très élevé, mais au moins se retrouve dans la zone acceptable. C'est le seul des trois qui obtient une classification intéressante autant dans le corpus de droit que le corpus complet. Il serait intéressant de poursuivre la recherche pour connaître quelles sont les classes créées.

Pour chaque algorithme j'ai pris qu'une solution, et je n'ai pas réussi à la comparer avec les catégories prédéfinies. Même avec l'aide du tutoriel de l'application, l'aide sur internet, je n'ai pas à savoir si la solution à 3 classes avait deux des classes prédéfinies confondues ou bien c'était des nouvelles classes. Ce manque, soit dans

l'aide, soit des mes aptitudes est un frein à la poursuite de cette analyse, allons maintenant voir la classification supervisée.

4.2.3.4 Classification supervisée

Le but de l'analyse de la classification supervisée est de classer automatiquement les 459 textes dans les 4 domaines de droit. D'un point de vue scientifique, l'étude de Pisetta et al. (2006) a déjà prouvé qu'il est possible d'effectuer ce type d'analyse de textes avec deux domaines. Selon Turban et al. (2007), l'analyste doit faire attention à ne pas sélectionner un faux problème pour l'exploration. Mon intérêt est de découvrir des bénéfices potentiels pour la SOQUIJ, d'expérimenter les défis de cette exploration de textes et de valider s'il est possible d'obtenir un modèle assez robuste pour être utilisé avec 4 domaines de droit.

Afin d'éclairer le choix d'un algorithme, le nœud « Classificateur automatique » de SPSS est utilisé. Ce nœud contient les types de modèle : « C5.0, régression logistique, réseau bayésien, discriminant, arbre C&R, Quest, Chaid et réseaux de neurones ». Il est utilisé seulement pour ressortir les statistiques haut-niveau des modèles car les résultats ne sont pas assez descriptifs pour l'analyse (figure 17). Les nœuds de modélisation individuelle offrent eux assez de données pour l'analyse. Les données créées lors de l'analyse des liens du texte sur le corpus complet sont réutilisées. Le nœud de classification automatique a retenu quatre types de modèles : « le C5.0, le Quest, le réseau de neurones et l'arbre C&R». Seul le réseau de neurones n'est pas un

algorithme d'arbres de décisions. Les arbres de décisions sont fréquents en exploration par leurs techniques simples et performantes.

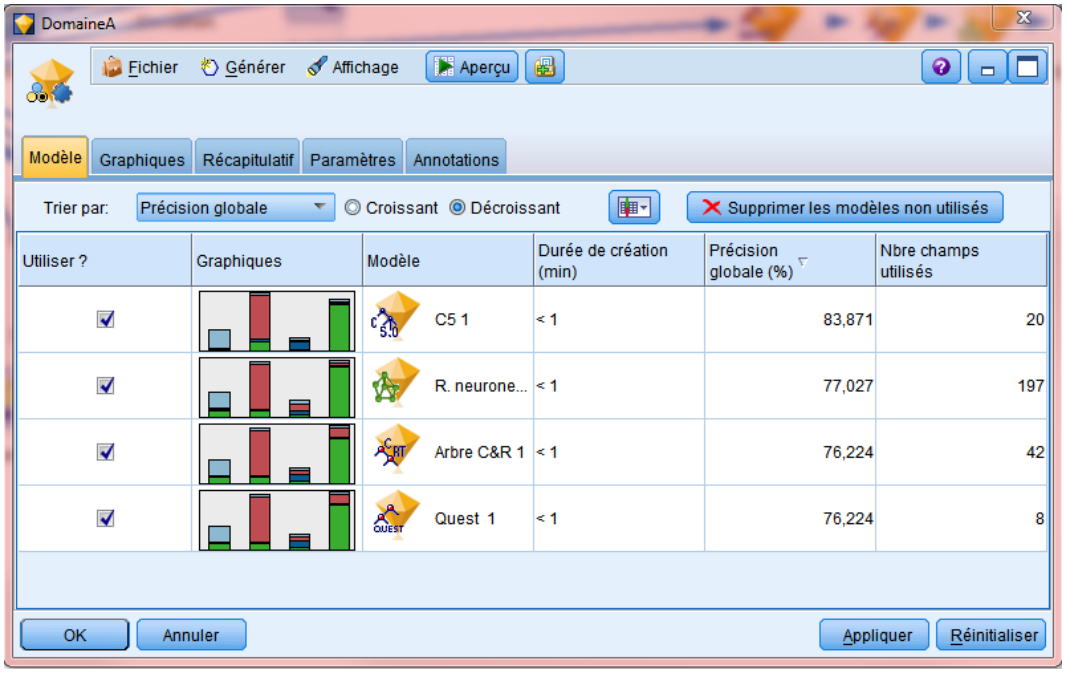
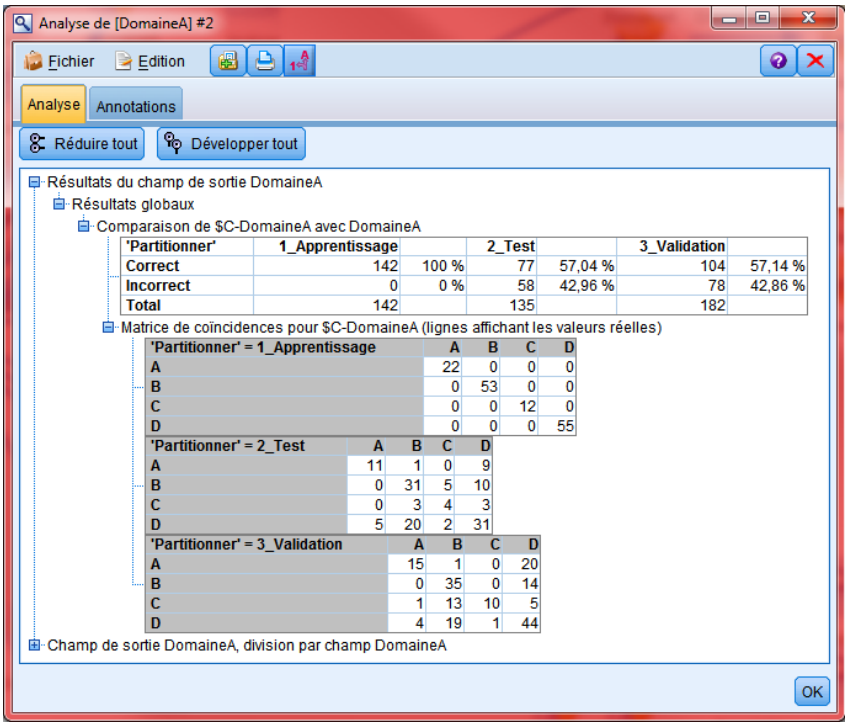


Figure 17 : Type de modèle conservé pour la classification supervisée

Les algorithmes C5.0, C&R, Quest et Réseau de neurones (RN) sont lancés individuellement (figure 18, 19, 20 et 21) puisqu'ils obtiennent une bonne précision globale (figure 17).



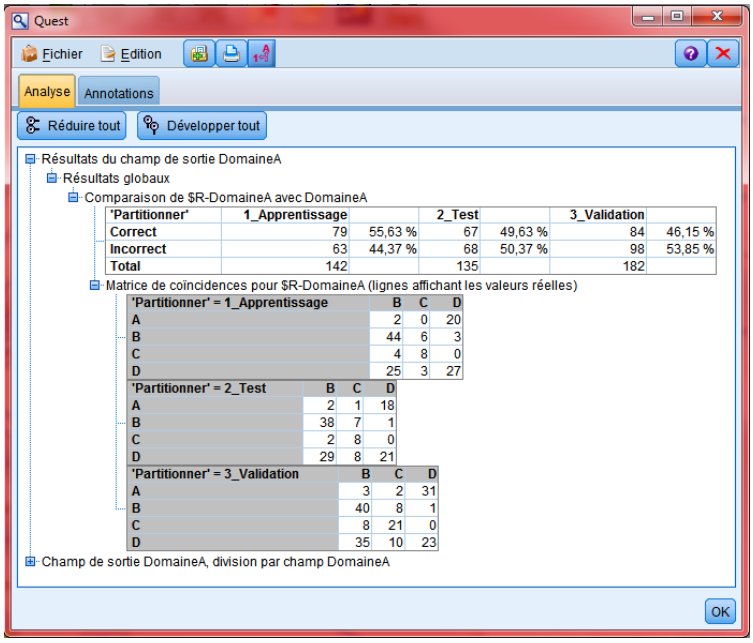
'Partitionner'	1_Apprentissage		2_Test		3_Validation	
Correct	142	100 %	77	57,04 %	104	57,14 %
Incorrect	0	0 %	58	42,96 %	78	42,86 %
Total	142		135		182	

'Partitionner' = 1_Apprentissage	A	B	C	D
A	22	0	0	0
B	0	53	0	0
C	0	0	12	0
D	0	0	0	55

'Partitionner' = 2_Test	A	B	C	D
A	11	1	0	9
B	0	31	5	10
C	0	3	4	3
D	5	20	2	31

'Partitionner' = 3_Validation	A	B	C	D
A	15	1	0	20
B	0	35	0	14
C	1	13	10	5
D	4	19	1	44

Figure 18 : Résultat classification supervisée c5.0



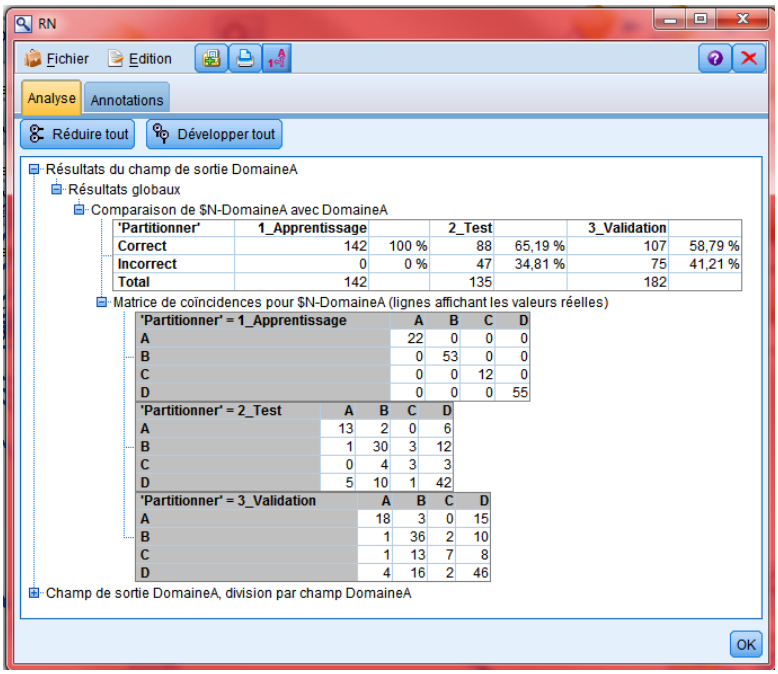
'Partitionner'	1_Apprentissage		2_Test		3_Validation	
Correct	79	55,63 %	67	49,63 %	84	46,15 %
Incorrect	63	44,37 %	68	50,37 %	98	53,85 %
Total	142		135		182	

'Partitionner' = 1_Apprentissage	B	C	D
A	2	0	20
B	44	6	3
C	4	8	0
D	25	3	27

'Partitionner' = 2_Test	B	C	D
A	2	1	18
B	38	7	1
C	2	8	0
D	29	8	21

'Partitionner' = 3_Validation	B	C	D
A	3	2	31
B	40	8	1
C	8	21	0
D	35	10	23

Figure 19 : Résultat classification supervisée Quest



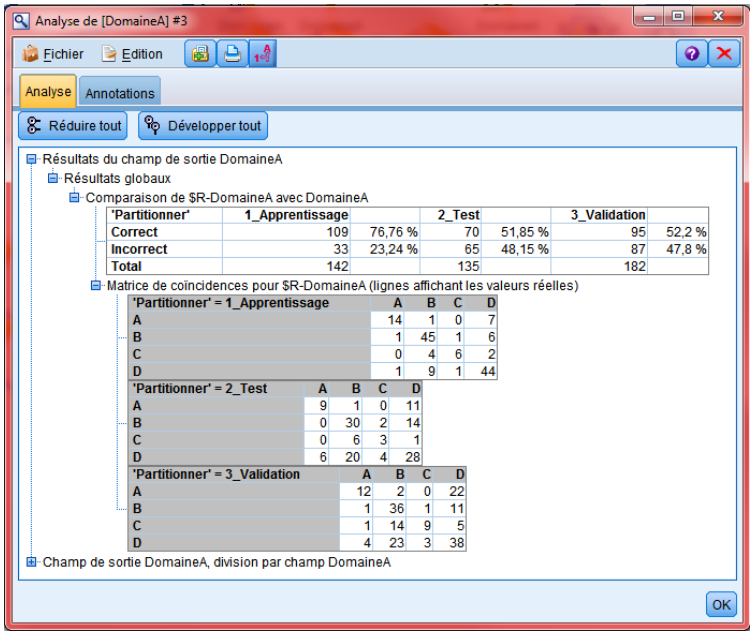
'Partitionner'	1_Apprentissage		2_Test		3_Validation	
Correct	142	100 %	88	65,19 %	107	58,79 %
Incorrect	0	0 %	47	34,81 %	75	41,21 %
Total	142		135		182	

'Partitionner' = 1_Apprentissage	A	B	C	D
A	22	0	0	0
B	0	53	0	0
C	0	0	12	0
D	0	0	0	55

'Partitionner' = 2_Test	A	B	C	D
A	13	2	0	6
B	1	30	3	12
C	0	4	3	3
D	5	10	1	42

'Partitionner' = 3_Validation	A	B	C	D
A	18	3	0	15
B	1	36	2	10
C	1	13	7	8
D	4	16	2	46

Figure 20 : Résultat classification supervisée Réseau de Neurones



'Partitionner'	1_Apprentissage		2_Test		3_Validation	
Correct	109	76,76 %	70	51,85 %	95	52,2 %
Incorrect	33	23,24 %	65	48,15 %	87	47,8 %
Total	142		135		182	

'Partitionner' = 1_Apprentissage	A	B	C	D
A	14	1	0	7
B	1	45	1	6
C	0	4	6	2
D	1	9	1	44

'Partitionner' = 2_Test	A	B	C	D
A	9	1	0	11
B	0	30	2	14
C	0	6	3	1
D	6	20	4	28

'Partitionner' = 3_Validation	A	B	C	D
A	12	2	0	22
B	1	36	1	11
C	1	14	9	5
D	4	23	3	38

Figure 21 : Résultat classification supervisée C&R

Les modèles sont maintenant prêts pour une analyse haut-niveau. Les algorithmes sont semblables puisqu'ils créent, en majorité, un arbre de décisions en divisant de manière itérative les données en sous-groupes de plus en plus petits. Il existe tout de

même des différences entre les algorithmes. Afin de simplifier la lecture, la première étape est de comparer et interpréter la performance des arbres pour la classification et de par la suite souligner quelques faits intéressants et examiner certains points.

Les algorithmes diffèrent par les critères utilisés pour décider des divisions. Dans notre cas, l'arbre C&R prend une mesure de dispersion (le coefficient Gini), le QUEST utilise un test Chi-deux et le C5.0 utilise une mesure théorique des informations appelée le rapport de gain d'informations. Le champ cible, la variable dépendante, est catégoriel. Les catégories sont les quatre domaines de droit :

- **A : Pénal ;**
- **B : Responsabilité ;**
- **C : Assurance ;**
- **D : Procédure civile.**

Les concepts extraits des analyses précédentes (analyse des liens) de type booléen sont utilisés à titre de variables pour la création des arbres.

Tableau 9 : Résumé des performances des arbres en pourcentage de documents bien classés

	R.N.	C5.0	C&R	Quest
Global				
Apprentissage	100,00 %	100,00 %	76,76 %	55,63 %
Test	65,19 %	57,04 %	51,85 %	49,63 %
Validation	58,79 %	57,14 %	52,20 %	46,15 %
Domaine A : Pénal				
Apprentissage	100,00 %	100,00 %	63,64 %	0,00 %
Test	61,90 %	52,38 %	42,86 %	0,00 %
Validation	50,00 %	41,67 %	33,33 %	0,00 %
Domaine B : Responsabilité				
Apprentissage	100,00 %	100,00 %	84,91 %	83,02 %
Test	65,22 %	67,39 %	65,22 %	82,61 %
Validation	73,47 %	71,43 %	73,47 %	81,63 %
Domaine C : Assurance				
Apprentissage	100,00 %	100,00 %	50,00 %	66,67 %
Test	30,00 %	40,00 %	30,00 %	80,00 %
Validation	24,14 %	34,48 %	31,03 %	72,41 %
Domaine D : Procédure civile				
Apprentissage	100,00 %	100,00 %	80,00 %	49,09 %
Test	72,41 %	53,45 %	48,28 %	36,21 %
Validation	67,65 %	64,71 %	55,88 %	33,82 %

La partition du corpus est pour l'échantillon d'apprentissage de 30 %, l'échantillon de test 30 % et donc celui de validation est 40 %.

Le tableau 9 souligne dans l'échantillon d'apprentissage que l'arbre C5.0 et le réseau de neurones ont une classification globale parfaite avec un taux de 100,00 %, le C&R a une classification globale avec un taux de 76,76 % et le Quest a une classification globale ordinaire avec un taux de 55,63 %. Globalement, on voit que les modèles classent moins bien les domaines de droit de l'échantillon test avec un taux de bonne classification de 65,19 % pour le réseau de neurones, 57,04 % pour le C5.0, 51,85 %

pour le C&R et de 49.63 % pour le Quest. L'échantillon de validation confirme la dégradation observée entre l'échantillon d'apprentissage et l'échantillon test avec des résultats similaires à l'échantillon de test (R.N. : 58,79 % ; C5.0 : 57,145 % ; C&R : 52,20 % et Quest : 46,1 5%)

La classification des différents domaines de droit pour le C5.0, le réseau de neurones et le C&R respecte généralement la classification globale. Le C5.0, sur l'échantillon d'apprentissage, la classification est bien entendu parfaite, sur l'échantillon de test les taux de bonne classification sont entre 40,00 % et 67,39 % et sur l'échantillon de validation les taux sont entre 34,48 % et 71,43 %. Le réseau de neurones, sur l'échantillon d'apprentissage, la classification est aussi parfaite et sur l'échantillon de test les taux de bonne classification sont entre 30,00 % et 72,41 % et sur l'échantillon de validation les taux sont entre 24,14 % et 73,47 %. Le C&R, sur l'échantillon d'apprentissage, la classification est entre 50,00 % et 84,91 % et sur l'échantillon test les taux sont entre 30,00 % et 65,22 % et sur l'échantillon de validation les taux sont entre 31,03 % et 73,47 %. Le domaine le moins bien classé dans les échantillons de test et de validation est l'assurance (C), qui est aussi le plus petit groupe, avec des taux variant de 24,14 % à 40,00 %.

Le Quest, quant à lui, ressort des résultats qui ont besoin d'être approfondis. Au global, ses taux étaient inférieurs aux autres. De plus, lorsque l'on observe le domaine pénal (A), son taux de bon classement est à 0 % autant sur l'échantillon d'apprentissage, de test et de validation. Il a donc préféré classer un domaine plus

volumineux tel que celui de la responsabilité (83,02 % apprentissage, 82,61 % test, 81,63 % validation) au détriment d'un plus petit.

4.2.3.5 Analyse de l'arbre C&R

Il est intéressant de sortir les statistiques hauts niveaux de la performance des arbres de décisions et du réseau de neurones. Il en est encore plus intéressant de visualiser leur fonctionnement d'un arbre de décision et de l'interpréter.

Le choix entre les trois arbres fut difficile et simple à la fois. Chacun d'entre eux a ses forces et faiblesses. Il faut donc les connaître avant de débuter. J'ai choisi l'arbre C&R pour sa souplesse et sa simplicité d'élagage. Puisque l'arbre C&R recherche de façon exhaustive les séparations possibles et que d'effectuer cet arbre par l'application SPSS peut prendre plus de six (6) heures, ainsi un ajustement a été fait afin de réduire le nombre de valeurs prédites dans l'arbre. Au lieu d'utiliser les concepts ressortis lors de l'analyse de texte précédent (TLA), j'ai restreint la liste aux quinze (15) principaux concepts (15) (figure 9 et 22) ainsi que ceux trouvés préalablement dans l'arbre C&R (figure 23). Cet ajustement permet de réduire le temps d'exécution à quelques minutes seulement (figure 24). Cet ajustement m'a pris quelques semaines avant de le découvrir et m'a permis d'augmenter la fréquence de lancement.

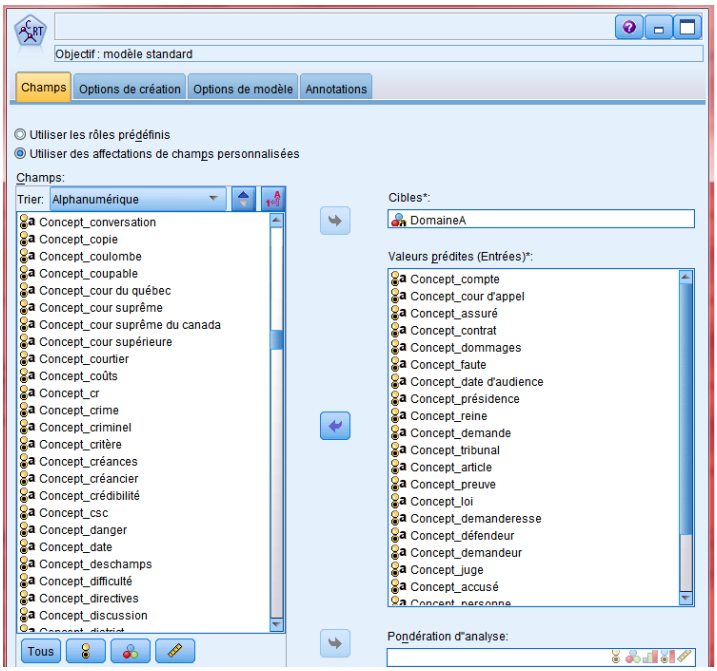


Figure 22 : Concepts utilisés pour l'analyse du C&R

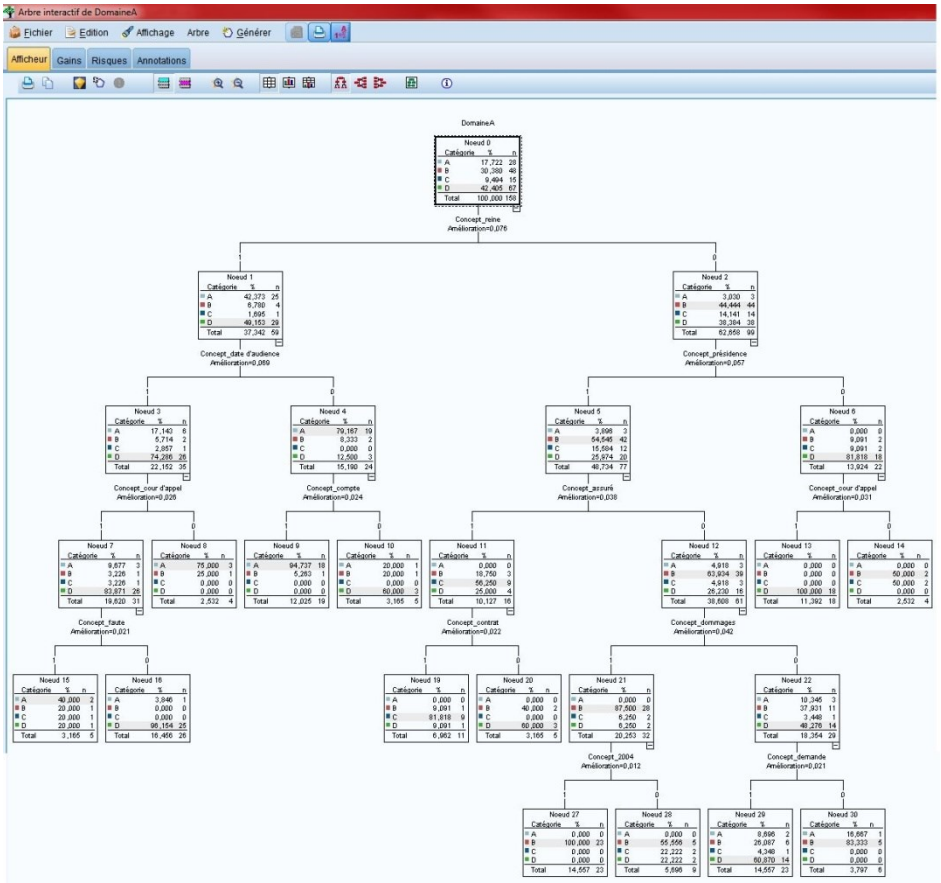


Figure 23 : Arbre original C&R

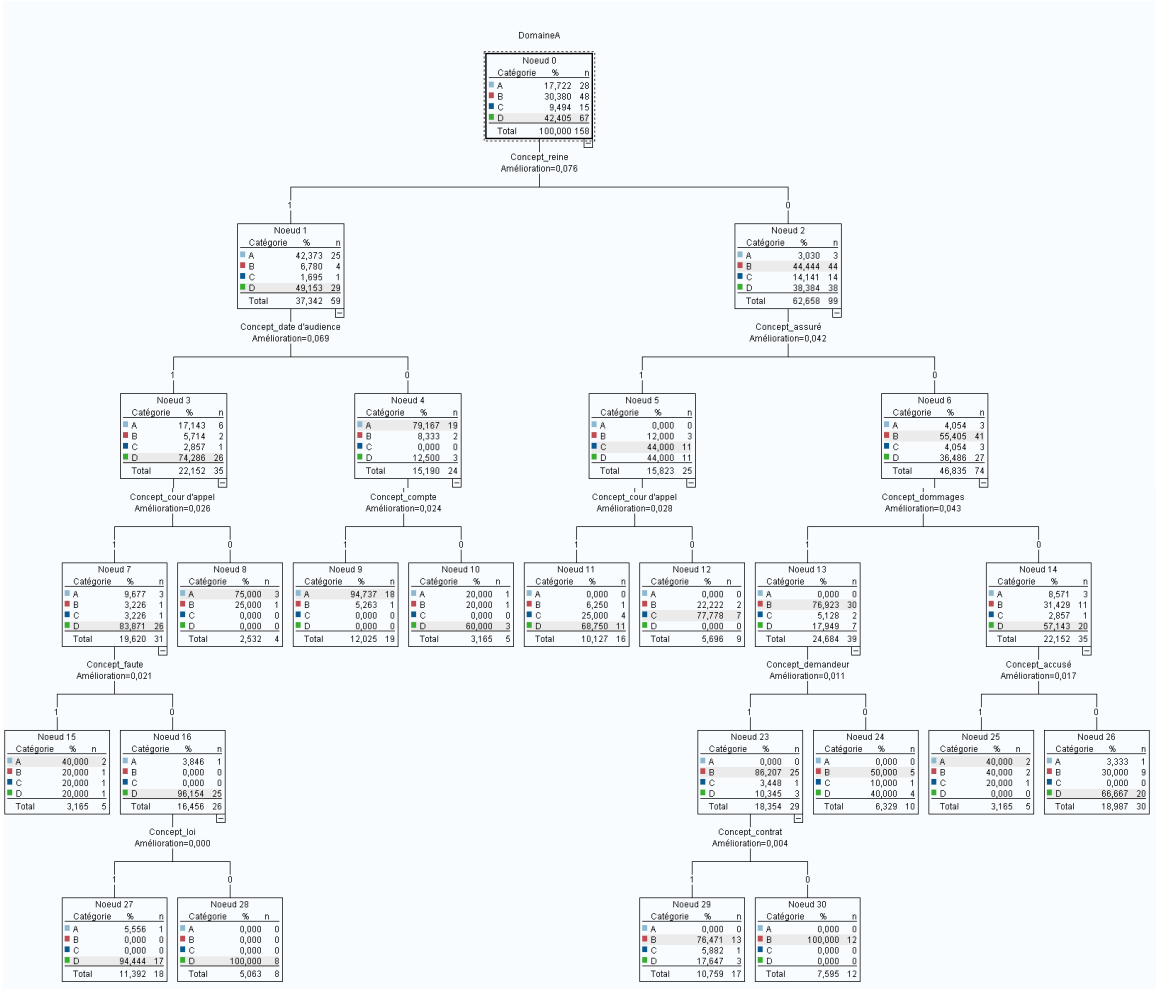


Figure 24 : Nouvel arbre C&R

La figure 24 présente l'arbre obtenu. La racine contient 158 documents. On voit que l'arbre représente bien le ratio des différents domaines. Selon l'algorithme, il semble que ce soit le concept « reine » qui donne la meilleure première séparation. Les documents qui ont le concept « reine » sont envoyés dans le nœud de gauche, les autres à droite. On voit que cette séparation est efficace pour le nœud de gauche : sur les 59 documents envoyés à gauche, seulement 5 sont des domaines B (responsabilité) et C (assurance) (8,5 %), ce qui donne 91,5 % des domaines A

(pénal) et D (procédure civile). Le nœud de droite est plus diversifié: sur les 99 documents envoyés à droite, 44 du domaine responsabilité, 14 du domaine assurance et 38 du domaine procédure civile.

Le restant de l'arbre se lit de la même façon. Cet arbre a 13 feuilles. On peut énoncer une règle pour chacune des feuilles. Par exemple, pour la feuille correspondant au nœud 9, qui contient une majorité de documents du domaine pénal, un document y sera classé s'il a un concept « reine », s'il n'a pas le concept «date d'audience » et il a le concept « compte ». Il sera alors classé dans le domaine pénal, avec une probabilité de 94,7 % d'avoir raison selon cette feuille. En fait, les trois seules feuilles de cet arbre qui classent dans le domaine pénal sont les nœuds 9, 15 et 25.

L'arbre de la figure 25 classe les documents avec l'outil de prévention de surajustement. Le surajustement arrive quand un modèle ne produit plus des prévisions très précises pour les nouvelles observations. On peut alors comparer les performances; ici on constate que la performance des deux arbres est très semblable.

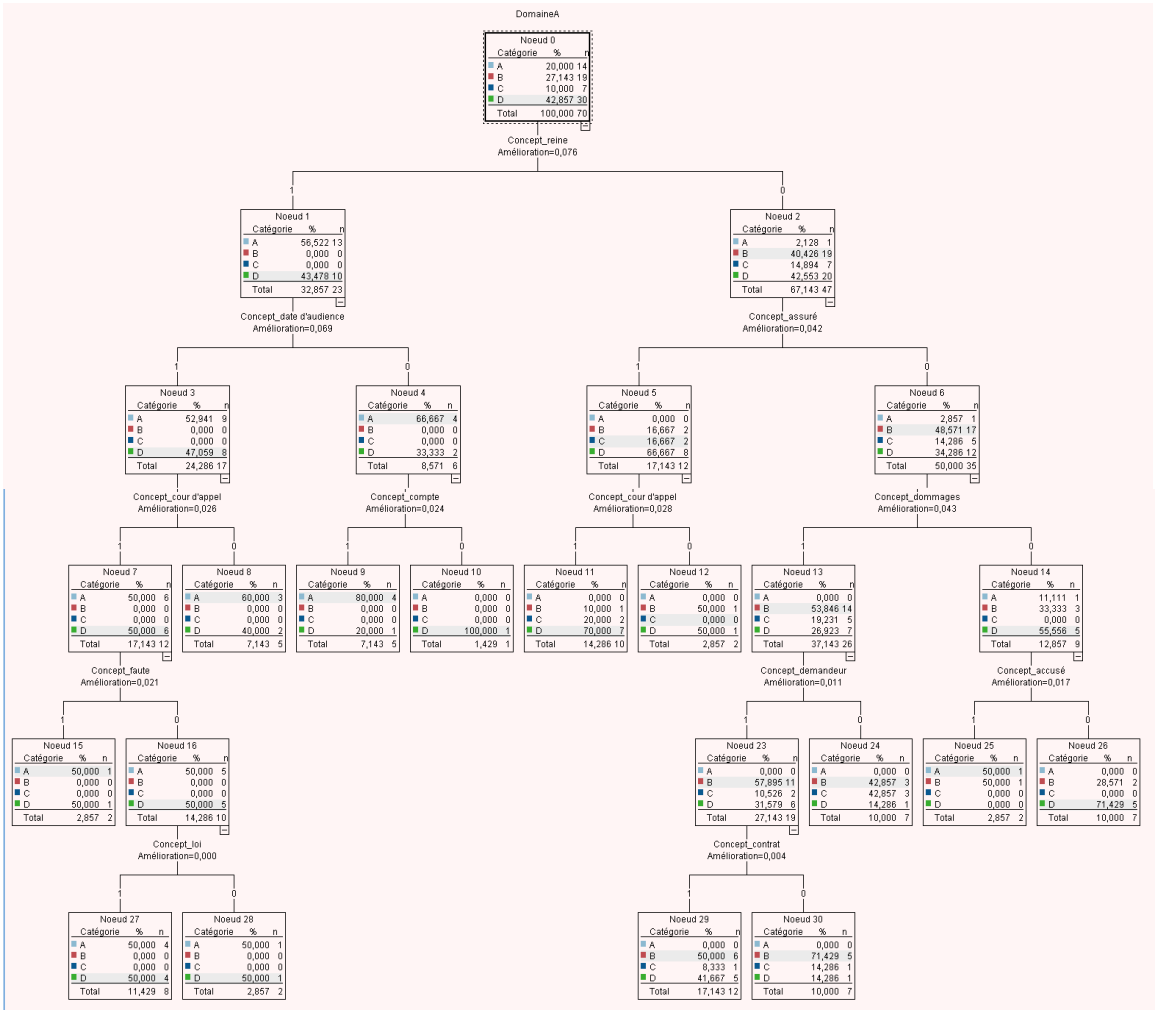


Figure 25 : Arbre C&R sur l'ensemble de prévention de surajustement

Les matrices de classification pour les deux ensembles (figure 26) donnent les performances : sur l'ensemble de développement d'arbre, 78,58 % des documents sont bien classés (89,29 % pour le domaine pénal, 62,5 % pour le domaine responsabilité, 46,67 % pour le domaine assurance et 88,06 % pour le domaine procédure civile), et sur l'ensemble de prévention de surajustement, 58,57 % sont bien classés (64,29 % pour le domaine pénal, 73,68 % pour le domaine responsabilité, 0 % pour le domaine assurance et 60 % pour le domaine procédure civile).

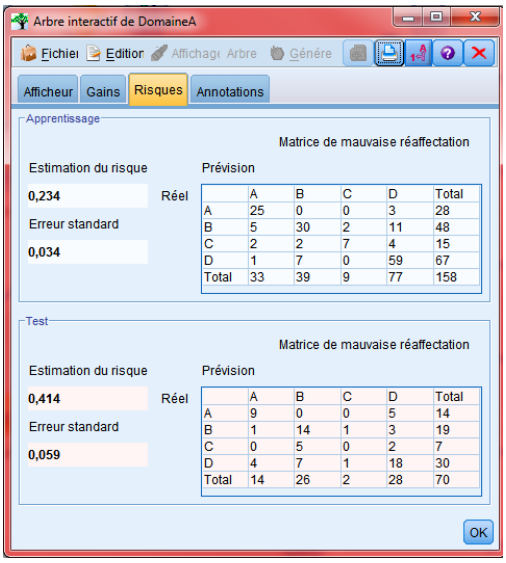


Figure 26 : Matrice de classification

En consultant la sortie de la figure 27, on constate que c'est le nœud 9 qui a les gains les plus importants pour les documents, avec un Index de 534,59. Cette feuille contient 19 documents, ce qui représente 12,03 % de la taille totale de l'ensemble de développement (158 documents). Il y a 94,74 % de ces 19 documents qui sont du domaine pénal, c'est-à-dire 18 documents. Ces 18 documents constituent 64,29 % des documents du domaine pénal. Les documents de cette feuille ont 5,35 fois plus de chances d'être du domaine pénal que les documents constituant l'ensemble de développement.

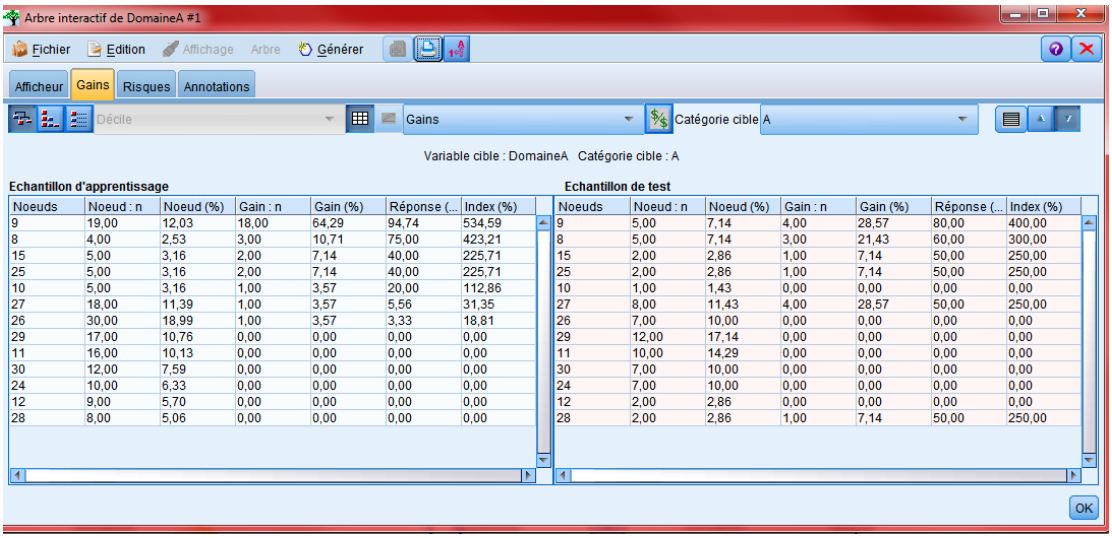


Figure 27 : Matrice de statistique de gain

La figure 28 permet de visualiser les règles. Il y a 13 règles, une pour chaque feuille de l'arbre. Il y a 4 règles qui classent dans le domaine pénal, 3 pour le domaine responsabilité, 1 pour le domaine assurance et 5 dans le domaine procédure civile.

La seule règle pour le domaine assurance énonce qu'un document qui a un concept « assuré » qui n'a pas le concept « cour d'appel » ni le concept « reine » sera classé dans ce domaine et ce, avec une probabilité de 77,8 %. Cette règle a été établie à partir de 9 documents (la probabilité et le nombre de documents proviennent toujours de l'ensemble de développement d'arbre).

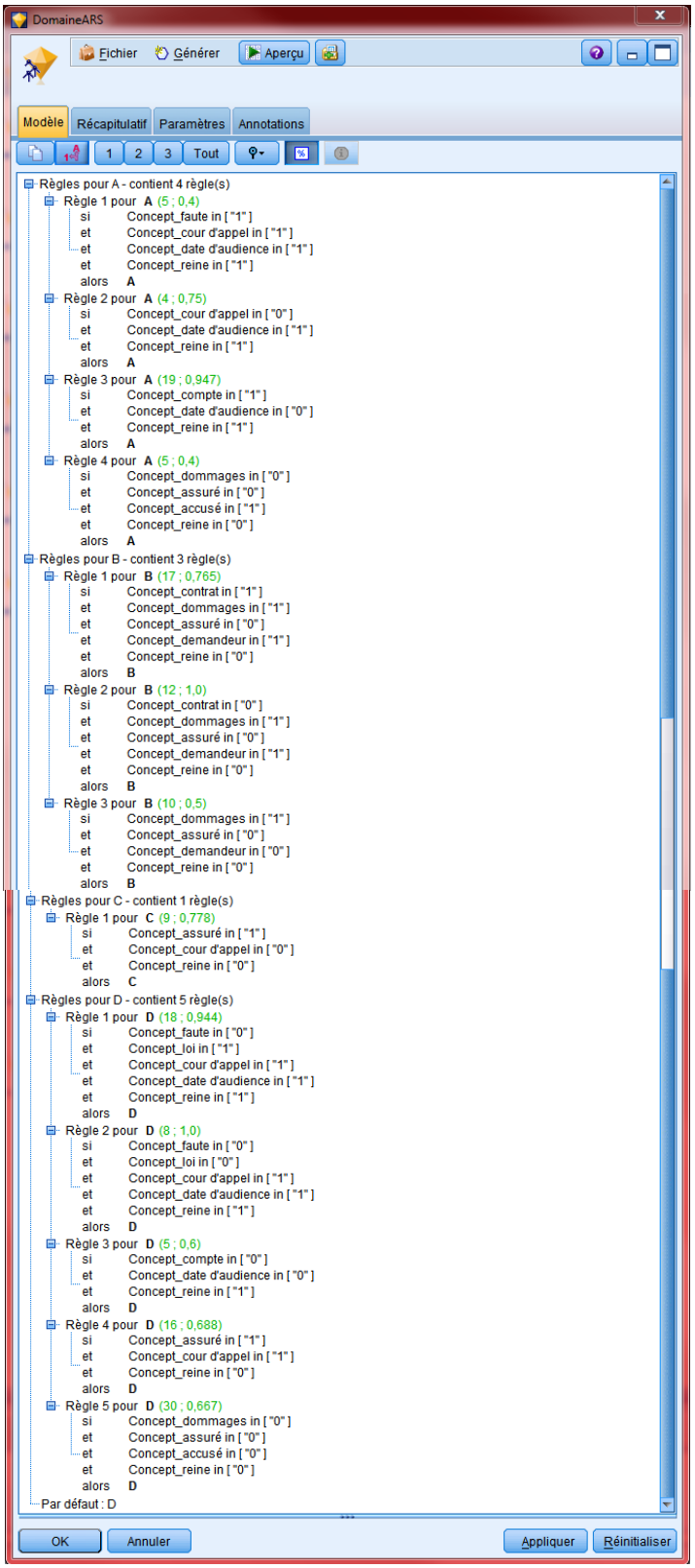


Figure 28 : Règles du modèle généré

Les matrices de classification sont présentées dans la figure 29. On voit que 79,39 % des documents de cet échantillon ont été bien classés, avec :

- 83,33 % des documents du domaine pénal bien classés;
- 86,57 % des documents du domaine responsabilité bien classés;
- 72,73 % des documents du domaine assurance bien classés;
- 74,23 % des documents du domaine procédure civile bien classés.

La performance ne se généralise pas très bien sur l'échantillon test : 64,71 % des documents sont bien classés, avec :

- 77,78 % des documents du domaine pénal bien classés;
- 67,57 % des documents du domaine responsabilité bien classés;
- 56,25 % des documents du domaine assurance bien classés;
- 60,42 % des documents du domaine procédure civile bien classés.

La performance de l'échantillon de validation confirme l'échantillon test : 66,07 % des documents sont bien classés, avec :

- 68,42 % des documents du domaine pénal bien classés;
- 84,09 % des documents du domaine responsabilité bien classés;
- 38,46 % des documents du domaine assurance bien classés;
- 52,78 % des documents du domaine procédure civile bien classés.

Résultats du champ de sortie DomaineA

Comparaison de $R-DomaineA avec DomaineA

'Partitionner'	1_Apprentissage		2_Test		3_Validation	
Correct	181	79,39 %	77	64,71 %	74	66,07 %
Incorrect	47	20,61 %	42	35,29 %	38	33,93 %
Total	228		119		112	

Matrice de coïncidences pour $R-DomaineA (lignes affichant les valeurs réelles)

'Partitionner' = 1_Apprentissage	A	B	C	D
A	35	4	0	3
B	2	58	3	4
C	1	5	16	0
D	7	17	1	72

'Partitionner' = 2_Test	A	B	C	D
A	14	0	0	4
B	1	25	8	3
C	1	5	9	1
D	6	12	1	29

'Partitionner' = 3_Validation	A	B	C	D
A	13	3	0	3
B	0	37	5	2
C	0	7	5	1
D	6	11	0	19

Figure 29 : Matrice de classification

L'arbre C&R doit tenir compte de probabilités a priori lorsqu'il calcule les probabilités d'affectation. Par défaut, les probabilités a priori ont été calculées selon les effectifs présents dans l'échantillon de développement d'arbre. Comme les catégories de domaine pénal et assurance sont plus petites (elles représentent à peu près 28 % de l'échantillon complet), c'est alors le classement des catégories de domaine responsabilité et procédure civile qui sont favorisé. Or il est possible de changer les probabilités a priori et j'ai décidé de tester une probabilité a priori de 0,25 pour chacun des domaines (j'ai laissé les autres options telles quelles) (figure 30).

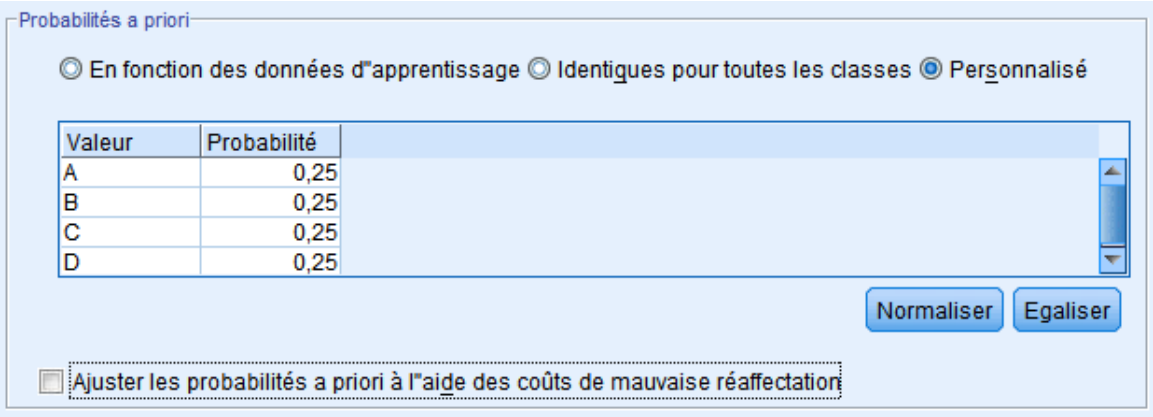


Figure 30 : Probabilité a priori

Le nouvel arbre, qui est présenté dans les figures 31 (développement) et 33 (prévention surajustement).

Cet arbre est plus petit que celui développé précédemment. Il a 11 feuilles. Puisque le nombre de concepts a été limité en amont, il est normal de retrouver les mêmes valeurs de séparation, mais à différents endroits pour tenir compte des priorités a priori.

Puisque j'ai touché aux probabilités a priori, le classement associé à une feuille ne correspond pas nécessairement au groupe qui contient la majorité des documents. Par exemple, si l'on prend le nœud 1, on a 28 documents du domaine procédure civile et 27 du domaine pénal, mais tous ces documents se retrouveraient classés dans le domaine pénal (c'est la partie ombragée du nœud qui indique dans quel groupe les documents sont classés).

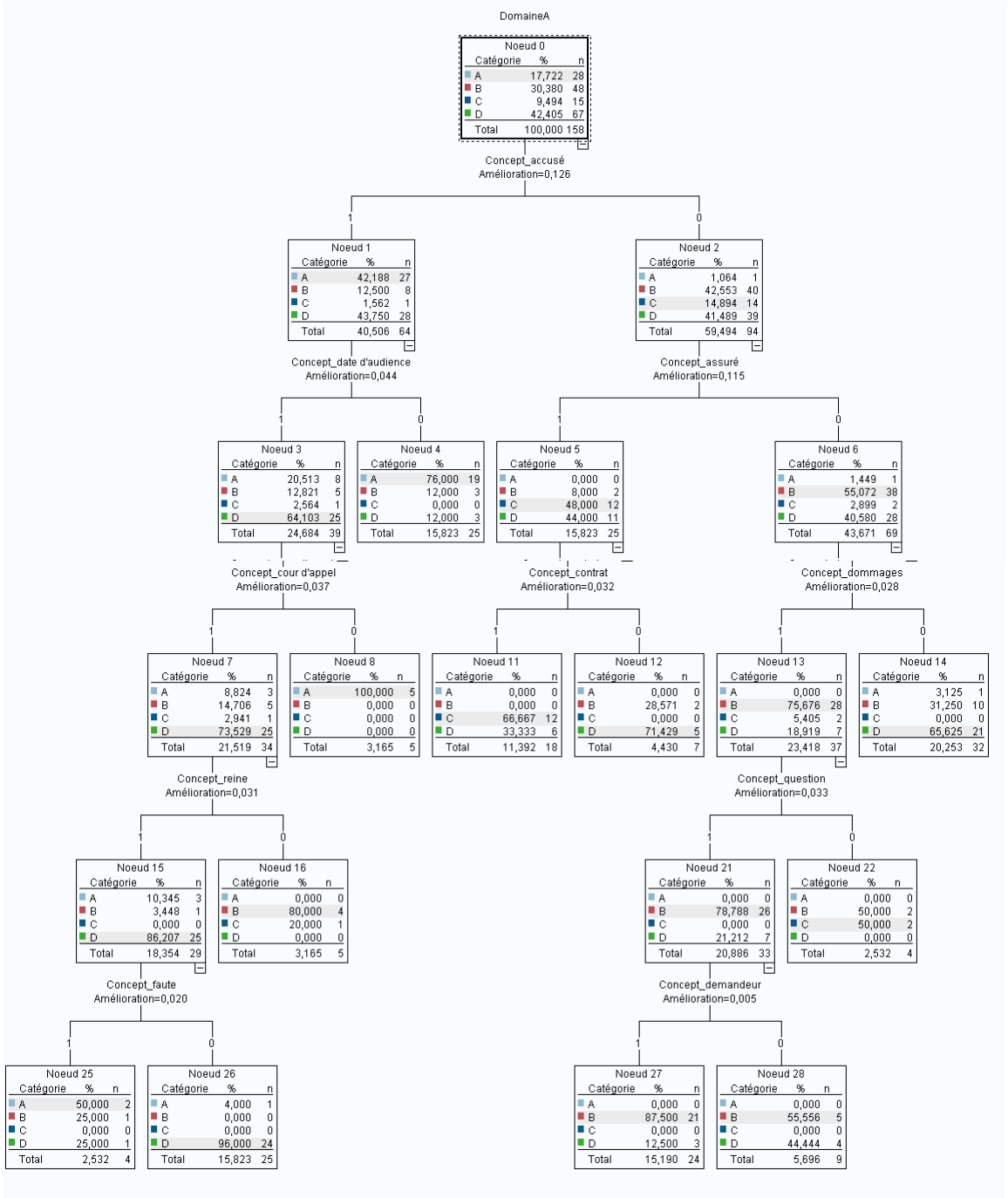


Figure 31 : L'arbre sur l'ensemble de développement

Et encore une fois, la performance de l'arbre se détériore sur l'ensemble de prévention de surajustement, et il restera à vérifier sa performance sur l'échantillon test et de validation.

Est-ce que ce nouvel arbre classe maintenant mieux les documents des plus petites catégories, sans que le classement des plus volumineuses ne se détériore pas trop ? Il semble que oui (figure 32 et 33) en tout cas pour l'échantillon d'apprentissage sur l'ensemble de développement d'arbre, 75,95 % des documents sont bien classés;

- 92,86 % pour le domaine pénal;
- 62,5 % pour le domaine responsabilité;
- 93,33 % pour le domaine assurance;
- 74,63 % pour le domaine procédure civile.

et sur l'ensemble de prévention de surajustement, 48,57 % sont bien classés :

- 64,29 % pour le domaine pénal;
- 52,53 % pour le domaine responsabilité;
- 42,86 % pour le domaine assurance;
- 40 % pour le domaine procédure civile.

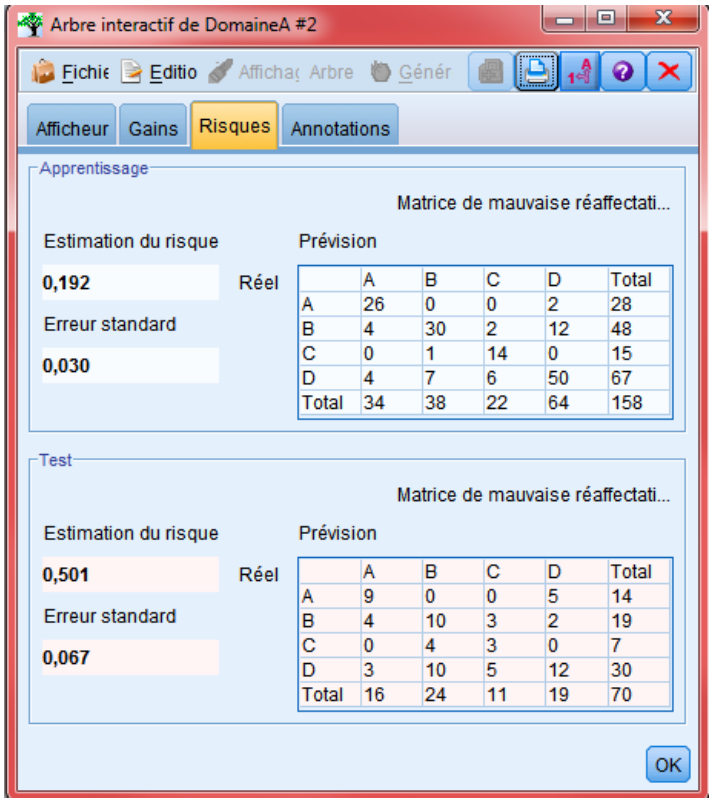


Figure 32 : Matrices de classification

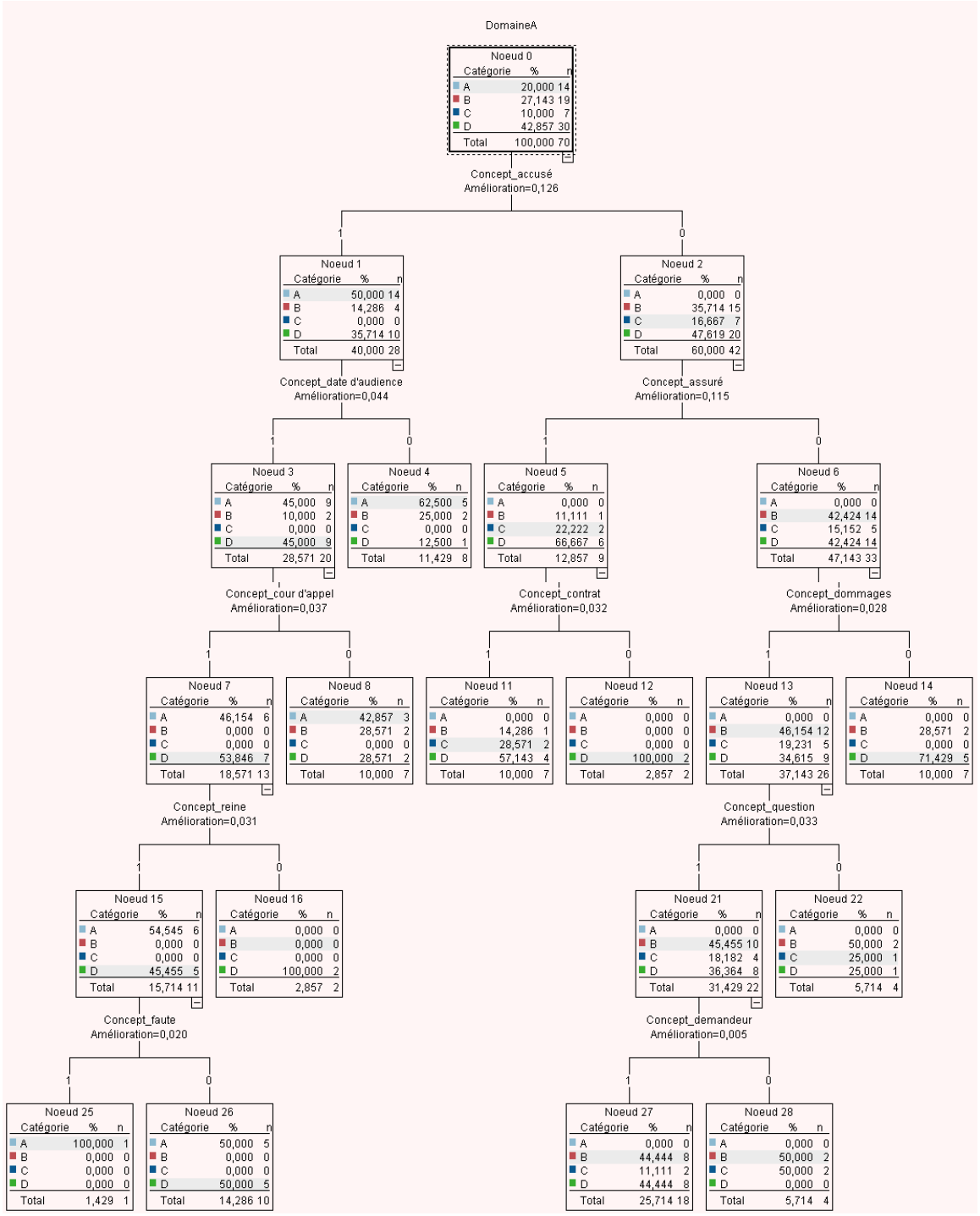


Figure 33 : L'arbre sur l'ensemble de surajustement

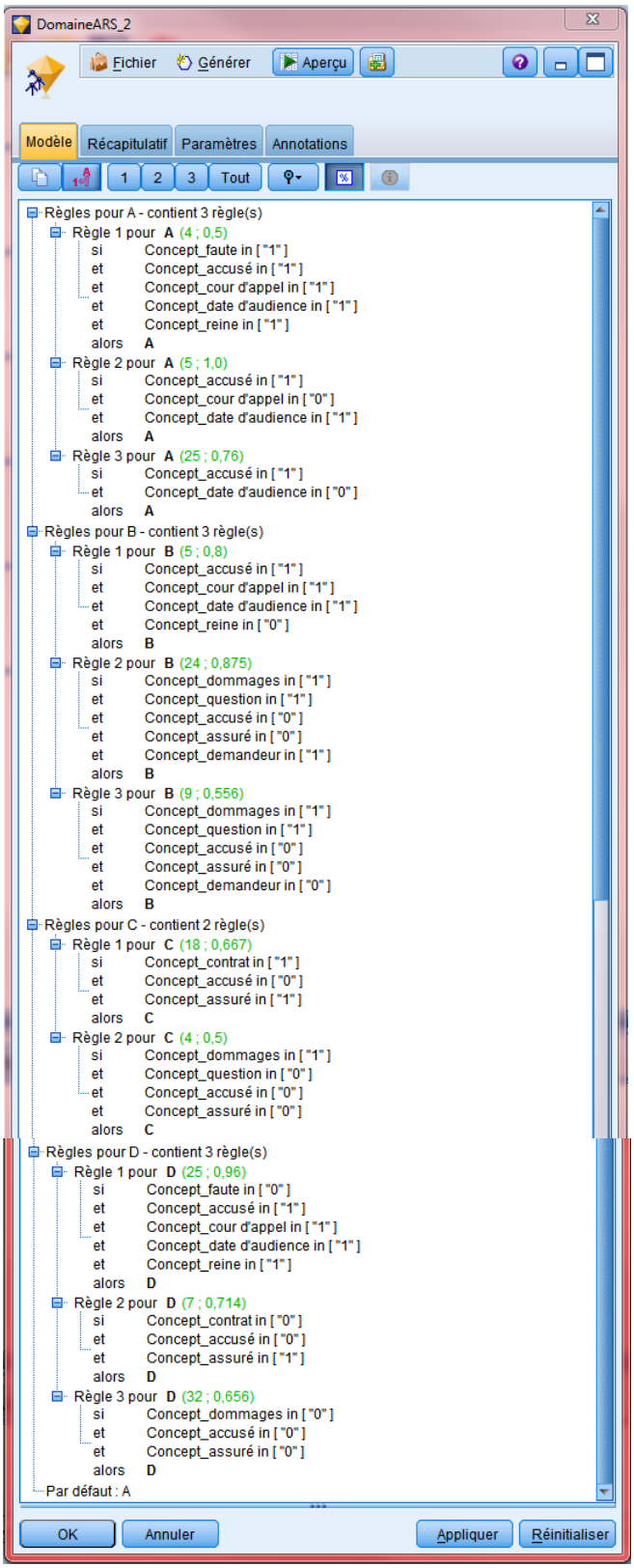
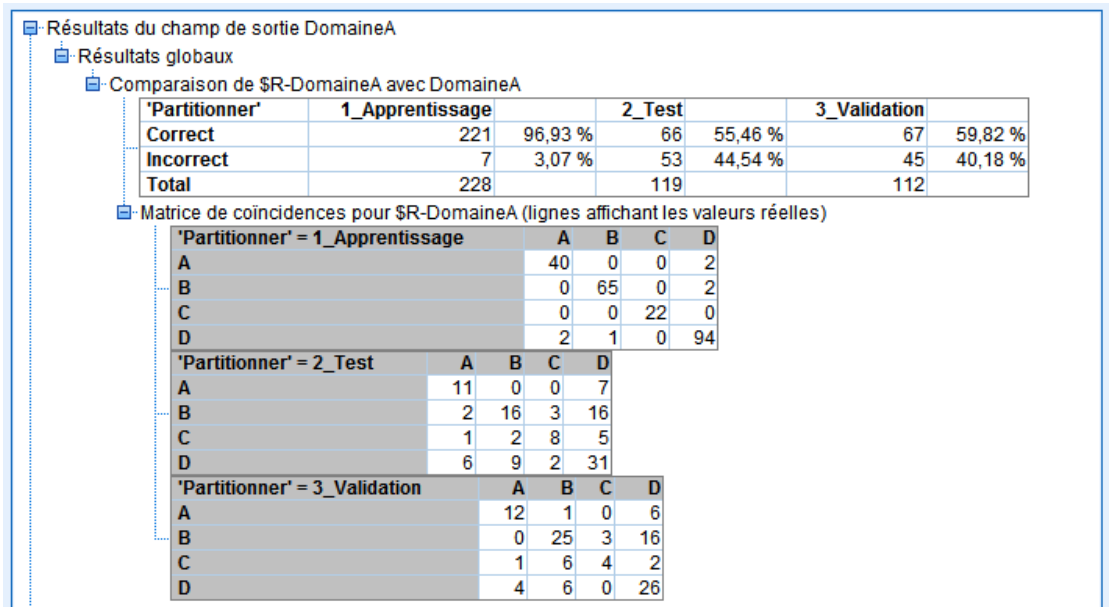


Figure 34 : Règles du modèle généré

La figure 34 montre les règles associées à l'arbre. Il y a 11 règles, une pour chaque feuille de l'arbre. Il y a 3 règles qui classent dans le domaine pénal, 3 pour le domaine responsabilité, 2 pour le domaine assurance et 3 dans le domaine procédure civile.

Le changement des probabilités a priori amène une règle supplémentaire dans le domaine assurance. La première règle énonce qu'un document qui a un concept « contrat » qui n'a pas le concept « accusé » et qui a le concept « assuré » sera classé dans le domaine assurance et ce, avec une probabilité de 66,7 %. Cette règle a été établie à partir de 18 documents. La seconde règle énonce qu'un document qui a un concept « dommages » qui n'a pas les concepts : « question, assuré, accusé » sera classé dans le domaine assurance et ce, avec une probabilité de 50 %. Cette règle a été établie à partir de 4 documents.


Résultats du champ de sortie DomaineA

Résultats globaux

Comparaison de $R-DomaineA avec DomaineA

'Partitionner'	1_Apprentissage		2_Test		3_Validation	
Correct	221	96,93 %	66	55,46 %	67	59,82 %
Incorrect	7	3,07 %	53	44,54 %	45	40,18 %
Total	228		119		112	

Matrice de coïncidences pour $R-DomaineA (lignes affichant les valeurs réelles)

'Partitionner' = 1_Apprentissage	A	B	C	D
A	40	0	0	2
B	0	65	0	2
C	0	0	22	0
D	2	1	0	94

'Partitionner' = 2_Test	A	B	C	D
A	11	0	0	7
B	2	16	3	16
C	1	2	8	5
D	6	9	2	31

'Partitionner' = 3_Validation	A	B	C	D
A	12	1	0	6
B	0	25	3	16
C	1	6	4	2
D	4	6	0	26

Figure 35 : Matrices de classification (apprentissage, test et de validation)

La figure 35 montre que 96,93 % des documents de l'échantillon d'apprentissage ont été bien classés, avec :

- 95,24 % des documents du domaine pénal bien classés;
- 97,01 % des documents du domaine responsabilité bien classés;
- 100 % des documents du domaine assurance bien classés;
- 96,91 % des documents du domaine procédure civile bien classés.

La performance ne se généralise pas du tout bien sur l'échantillon test : 55,46 % des documents sont bien classés, avec :

- 61,11 % des documents du domaine pénal bien classés;
- 43,24 % des documents du domaine responsabilité bien classés;
- 50 % des documents du domaine assurance bien classés;
- 64,58 % des documents du domaine procédure civile bien classés.

La performance de l'échantillon de validation confirme l'échantillon test : 59,82 % des documents sont bien classés, avec :

- 63,16 % des documents du domaine pénal bien classés;
- 56,82 % des documents du domaine responsabilité bien classés;
- 30,77 % des documents du domaine assurance bien classés;
- 72,22 % des documents du domaine procédure civile bien classés.

Les performances sont très bonnes en apprentissage mais ne se généralisent pas. Il ne semble pas adéquat de se baser sur cet arbre de décisions pour établir quelles sont les différentes catégories de domaine pour les prochains documents.

Ce chapitre des résultats nous a permis de connaitre l'application gagnante et comment elle a réagi sur le corpus de la SOQUIJ. L'application a fait une analyses des liens des textes, a effectué la classification supervisée ainsi que non-supervisée. Nous avons même regardé plus attentivement l'arbre de décision C&R. Poursuivons avec le processus CRISP-DM tout en attaquant la discussion.

5 Discussion

Le but de la section discussion est de dégager des recommandations de ce qui entoure les sujets principaux de la recherche et d'interpréter les résultats des analyses effectuées. Ainsi donc, dans les résultats, nous avons établi comment un outil de TM peut être utilisé pour faciliter la découverte de connaissances sur un corpus (du domaine du droit) ainsi que des défis et bénéfices de l'utilisation d'outil de TM lors de la découverte de connaissances.

La méthodologie CRISP-DM nous permet de revenir analyser, interpréter les résultats et agir sur les connaissances. Je crois que ces sections se prêtent très bien à cette section de discussion.

La recherche est de type « design science » et donc nous voulons savoir comment l'artéfact technologique répond au questionnement de la recherche. Dans le cas

présent, l'artéfact généré dépasse les modèles statistiques démontrés. Il inclut les processus pour le choix de l'application, les méthodologies d'exploration et des pistes de recherches. L'utilisation d'un nouvel outil et d'un nouveau processus inclut ses défis et ses bénéfices. Les principaux éléments répondant aux questions de recherches sont justifiés dans les sections suivantes : « les résultats, le choix de l'outil et agir sur les connaissances».

5.1.1 Analyser, interpréter et vulgariser les résultats explorés. (Évaluer)

Les résultats obtenus ont été analysés, comparés en fonction des différents algorithmes utilisés. L'explorateur de données textuelles est perplexe devant les résultats, il se questionne sur comment améliorer ces résultats. J'ai tout de même passé plusieurs mois (années) pour obtenir si peu de résultats. Les résultats étaient si difficiles à obtenir qu'à un certain point je pensais clore le projet en soulignant seulement les étapes effectuées. Ainsi donc, c'est le gestionnaire en moi qui maintenant se questionne sur le déroulement du projet, doit-on continuer les itérations, les analyses et la collection de résultats. Il s'interroge surtout à l'atteinte des résultats car des résultats il y en a, mais mitigés ils sont.

Tout d'abord, rappelons-nous de l'analyse des liens entre les textes. Les résultats sont semblables entre la série sur le corpus de droit assurance et le corpus complet. Il y avait 8 concepts communs « tribunal, article, preuve, loi, demanderesse, défendeur, demandeur et juge » dans les 15 concepts les plus fréquents. Une analyse plus poussée de chacun des domaines pourrait ressortir des concepts forts par domaine et

pas seulement dans un seul domaine assurance ou bien dans le corpus complet. Un spécialiste pourrait classifier les concepts avec un attribut autre que « unknown » et ainsi faciliter la création d'un dictionnaire.

L'analyse de segmentation (clustering) n'obtient pas de résultats intéressants avec le « K-Means » et le « Kohonen ». Mais le « Two-Steps » a obtenu une classification avec un coefficient de silhouette intéressante pour les deux corpus (spécifique et complet). La poursuite de l'analyse de classification supervisée ne m'a pas permis d'étudier plus attentivement le contenu des 3 classes ressorties. Le manque de temps et d'expertise m'ont contraint de continuer à la classification supervisée. Il serait intéressant d'approfondir l'analyse de segmentation.

L'analyse de la classification supervisée avec l'étude de Pisetta et al. (2006) me laissait miroiter une chance d'obtenir des résultats probants. Les algorithmes (avec les 459 textes dans les 4 domaines de droit) n'ont pas réussi à bien performer dans les échantillons de validation avec des résultats entre 46 et 60 pour cent. Ces résultats ne sont pas assez probants pour les mettre en production. Ici, une analyse plus poussée des concepts par domaine en excluant les concepts génériques auraient pu aider la classification supervisée.

J'ai travaillé sur différentes facettes du TM puisqu'au départ je n'obtenais pas de résultats. À la minute que j'ai obtenu des résultats, même loufoque, je me suis obstiné à travailler sur différent point, l'analyse de lien, de segmentation et de classification

supervisée. Cette dispersion d'analyse à divisé mon temps de recherche et aussi réduit la possibilité d'avoir de meilleurs résultats.

Tel que souligné, un outil de TM peut être utilisé pour retravailler les données, les documents. Il facilite l'analyse des liens du texte et permet de ressortir des concepts des différents textes. Un spécialiste du domaine pourrait les réutiliser, trier, affiner afin d'élaborer de meilleure classification supervisée et non supervisée. L'outil de TM utilisé dans une approche et méthodologie d'exploration jumelée avec un expert peut faciliter la découverte de connaissance sur un corpus. Dans le contexte de la SOQUIJ, les modèles de classifications supervisées qui ont été évalués ne sont pas encore assez performants. Ceux de segmentation (clustering), plus spécialement celui du « Two-steps » apporte un nouveau regard aux données. L'analyse des liens des textes, et bien je ne crois pas qu'un analyste aguerri de la SOQUIJ soit étonné des concepts forts, mais des pistes pourraient s'offrir dans une analyse par domaine. Donc, les résultats obtenus n'offrent actuellement que des pistes et peu d'apports à l'amélioration des processus à la société d'information juridique et ses clients.

5.1.2 Choisir l'application

Le choix de l'application d'exploration de textes est vital car il permet ou non de répondre aux questions posées par le gestionnaire. Le choix de l'outil « IBM PASW SPSS » n'était peut-être pas le meilleur dans le cadre de ce mémoire.

La sous-évaluation de deux critères d'évaluation de l'application d'exploration de données textuelles semble avoir faussé l'obtention de résultats probants. Les critères :

« coût et interroger un dictionnaire » de la section des autres critères, ont été sur et sous-évalués pour l'obtention de résultats probants. L'application la plus dispendieuse comprenait des qualités techniques indéniables et un dictionnaire francophone du domaine juridique. Ce dictionnaire aurait amené une tout autre perspective aux résultats obtenus.

Encore que l'outil « IBM PASW SPSS » n'intègre pas de dictionnaire francophone dans le domaine juridique, il a tout de même un dictionnaire francophone générique et lui permet d'effectuer les analyses. L'interface de l'outil est conviviale et permet assez facilement de créer une analyse statistique. L'outil ne requiert pas de programmation afin d'effectuer les analyses, ce qui est un plus pour les non-programmeurs. Après plusieurs mois, d'ajustement, de précision, j'ai réussi à créer des analyses de segmentation (clustering) et de classification supervisée. Cependant, afin de pousser les analyses, autant l'aide que les forums de discussion, l'application n'offre qu'un support minimal.

Le choix des analyses était varié et intéressant, de plus, l'application comprend un choix varié d'algorithme pour l'exploration de données. L'application s'est facilement branché sur le corpus et offre type de branchement pour les sources de données diverses. Donc, une analyse statistique par un outil commercial peut rendre un corpus de textes plus complet. Malgré les résultats de ce mémoire, l'exploration de données textuelles par un outil commercial pourrait organiser différemment ces textes et les rendre plus simples d'accès tout en étant exhaustifs.

5.1.3 Agir sur les connaissances découvertes. (Déployer)

À la lumière de mon expérience d'exploration de textes avec les données de la SOQUIJ, je peux répéter que la société ne pourrait pas utiliser les modèles résultants de cette recherche directement en production. J'ai déjà soulevé deux points d'améliorations. Le choix d'un outil adapté au domaine juridique et d'affiner le dictionnaire permettant ainsi d'améliorer l'analyse des liens du texte (concepts) et obtenir de meilleurs modèles. Ce serait donc de recommencer la roue du CRISP-DM et pour ce faire, voici quelques autres opportunités et défis sur l'utilisation du TM pour la SOQUIJ.

Je persiste à croire qu'il existe un réel potentiel d'exploration dans les résumés de verdicts que possède la SOQUIJ. Connaissant mieux le domaine et les données, je crois que l'analyse systématique des différents domaines pour la classification est erronée. Ces analyses demandent de partir avec un volume de données trop grand et nécessite une connaissance autant du domaine du droit que de l'exploration de données. Deux options s'offrent à la SOQUIJ. La première option est de créer un dictionnaire des concepts avec la technique des pas de bébés (« baby steps »). Cette option exigerait que les analystes en droit identifient au fur et à mesure des concepts dans les différents domaines de droit qu'ils sont experts. Ainsi, après quelque temps, un explorateur aurait des concepts intéressants de plusieurs domaines de droit et il pourrait améliorer les modèles développés.

La seconde option est de développer des analyses ad hoc personnalisées pour certains bons clients afin d'améliorer leurs recherches. Par exemple, avec une question précise, un analyste chevronné pourrait créer un arbre de décisions et interpréter quelles sont les règles établies dans ce type de cause pour l'avocat. Ainsi, les informations découvertes permettraient d'améliorer les connaissances dans le corpus.

Le problème d'affaire de la SOQUIJ est souligné par Hirschman et Gaizauskas (2001). Ils mentionnent que l'information en ligne (web) est disponible, mais elle est souvent difficile à obtenir. Ils offrent comme piste de recherche d'automatiser les champs de recherche. Sur ce point, Shi et Yang (2007) soulignent que les moteurs de recherche Web sont l'une des solutions les plus populaires disponibles sur le Web. Cependant, il n'a jamais été facile pour les utilisateurs novices d'organiser et de représenter leurs besoins d'information en utilisant des requêtes simples. Les utilisateurs modifient leurs requêtes de recherches jusqu'à ce qu'ils obtiennent les résultats attendus. Par conséquent, il est souvent souhaitable pour les moteurs de recherche d'offrir des suggestions liées aux requêtes des utilisateurs. En outre, en identifiant les requêtes connexes, le moteur de recherche peut effectuer des optimisations sur leurs systèmes, telles que l'expansion de requêtes et l'indexation de fichiers. Les requêtes connexes sont établies dans le « log » de requêtes précédemment soumises par des utilisateurs humains, qui peuvent être identifiées en utilisant un modèle amélioré de règles d'association. Les utilisateurs peuvent utiliser les requêtes liées suggérées d'accorder ou de réorienter le processus de recherche.

Cette option pourrait être effectuée sur les requêtes sur «Tout Azimut ». Cette technique semble avoir l'avantage d'être plus simple et a aussi l'avantage d'amener une technique supplémentaire qui est la recherche d'association (en anglais : market basket analysis) utilisée comme son nom l'indique en anglais principalement dans le domaine du commerce de détail. Cette méthode est différente de l'association simple, car elle nous permet dans le cas des requêtes et des achats de comprendre ce qui est fureté ensemble et ultimement acheté (effet de causalité). Le désavantage de cette technique c'est qu'elle regarde seulement les requêtes et les achats pour la classification et l'association et peut donc laisser tomber des résumés qui sont hors normes et qui ne sont jamais consultés. De plus, je ne connais pas les systèmes derrière « Tout Azimut » et je ne sais pas s'il est possible d'obtenir un échantillon dénormalisé des requêtes et des achats (sans les noms des personnes qui ont fait les requêtes et les achats) pour faire l'exploration.

Développer ces suggestions à la SOQUIJ nécessiterait l'ajout d'un bon architecte en intelligence d'affaires qui comprend aussi le côté relationnel des transactions. Cette personne pourrait développer une architecture propice au développement « d'ODS » pour permettre l'exploration ainsi que le développement de cube. Cet architecte devra aussi travailler avec des analystes qui connaissent bien les différents domaines du droit, mais aussi très bien l'exploration de données. Il est actuellement difficile sur le marché actuel de l'emploi au Québec de trouver des personnes ainsi qualifiées, car je crois que le marché de l'intelligence d'affaires en est encore à ses débuts. Très peu

d'entreprises sortent des analyses statistiques traditionnelles et font de l'exploration de données et donc, que dire de l'exploration de textes !

Je veux conclure sur un point, je n'ai pas exploré le champ de langage naturel qui semble intéressant. J'aurais surtout désiré apporter plus de détails sur les liens entre les concepts étudiés afin d'évaluer leurs proximités comme présentées sur la figure 36… ce sera pour une prochaine recherche, mais pas pour moi.

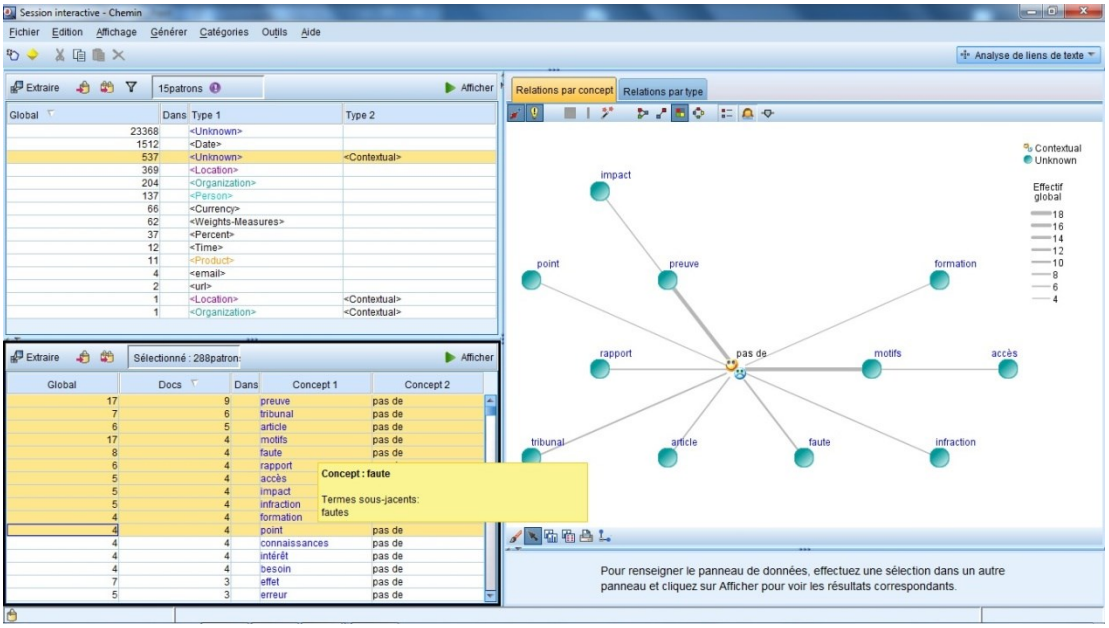


Figure 36 : Liens entre concepts

5.2 Limites et défis associés à la recherche

J'ai toujours été une personne franche et directe, je me permets de conclure sur les limites et les défis associés à ce projet de recherche.

Tout d'abord, ma codirectrice m'a répété à plusieurs reprises que l'envergure du projet était difficile et j'avoue elle avait raison. L'objectif d'effectuer l'exploration de données textuelles dans un domaine nouveau et en français est énorme. Lors d'une conférence SAS, j'ai fait la rencontre de Jean-Paul Isson, vice-président BI & analyse prédictive Monster Worldwide. Il m'avait dit avoir passé 5 ans avec une équipe sur un projet d'exploration de textes afin d'obtenir des résultats probants. De se restreindre dans une section de l'exploration de textes, par exemple la segmentation, est à conseiller.

Les qualités d'explorateurs de données du chercheur sont bonnes, mais l'apprentissage de l'exploration de données textuelles et l'apprentissage du domaine ont été ardues. De l'aide du côté juridique aurait permis une compréhension claire des domaines à catégoriser. De plus, le manque de revenus stables durant le projet a rattrapé son auteur et j'ai dû le terminer en travaillant à temps plein.

Le peu d'aller-retour, la distance, le manque d'interactions entre mes directeurs et moi n'ont pas aidé à la recherche. J'aurais dû mieux planifier mes déplacements et interactions. Les directeurs ne sont en rien responsables de cet item. De plus, je suis une personne qui avance par projet et j'ai sous-évalué le temps nécessaire pour les différentes étapes (jalons), ce point explique la durée en temps de ce mémoire.

La métaphore utilisée dans la section d'introduction, que le monde change, évolue que l'eau coule en dessous du pont est véridique. Malgré cela, et même si ce mémoire a presque pris 5 ans à écrire, les avancées dans le domaine sont peu nombreuses et se retrouvent spécialement dans l'analyse des sentiments et des médias sociaux, ce qui aujourd'hui est très pertinent dans le domaine du marketing.

J'ai beaucoup appris durant le projet et ce, presque autant en gestion de projet TI qu'en exploration de données textuelles. Le défi réel durant ce projet a été beaucoup plus la gestion de mon temps, de mes objectifs, de mon commanditaire principal que le côté technique de l'exploration de données. Ces apprentissages ne vont pas sans heurt, la maîtrise de l'exploration de données textuelles et la création d'un artéfact technologique pour un organisme tel que la SOQUIJ sont difficiles.

Merci.

6 Annexe

6.1 Courriel – Information sur la percée du TM

Seth Grimes

De: Seth Grimes []
Envoyé: 3 novembre 2011 13:22
À: Alexandre Tardif
Cc:
Objet: Re: TR: Use of text mining

Hello Alexandre,

Natural-language processing and text mining (with significant overlap) are used extensively in two legal-domain areas: Embedded in e-discovery solutions and in the creation of information tools knowledge bases at companies such as Thomson Reuters and LexisNexis. I don't have numbers for you, however.

Seth

On Thu, 27 Oct 2011, Alexandre Tardif wrote:

Hi,
My name is Alexandre Tardif and I am presently doing a master degree in Business Intelligence (BI) at Université Sherbrooke in Canada. (Sorry for my written English, it has been a long time since I've written or talked in English). I'm doing some text mining in the legal field and I back at my literature revue when I found that I didn't get fact about the use of text mining in the industry. I've contacted Ezra Steinberg trying to get more information on the use of text in mining in the industry; do you have any data to share? Like % percent are using it, and in which fields??? Or do you know any reports I could access? The only one I found was the Rexer Analytics? Anything in mind to help me would be appreciated?
Alexandre Tardif

Karl Rexer

De: Karl Rexer [
Envoyé: 27 octobre 2011 20:08
À: Alexandre Tardif
Objet: Re: Hard facts text mining...
Pièces jointes: Rexer DM-Survey PAW deck 10-19-11.pdf; ATT00022.htm; Rexer_Analytics_2010_Data_Miner_Survey_Summary_Report.pdf; ATT00025.htm

Alexandre –
Thank you for your interest in our Annual Data Miner Surveys. Attached are:

- Last week's slides from our presentation to the Predictive Analytics World NYC conference (Highlights of the 2011 Survey results).

- The full 37-page summary report from the 2010 Survey.
    - The 2010 Survey highlights and links to previous year highlights are available at http://www.rexeranalytics.com/Data-Miner-Survey-Results-2010.html
    - The 2010 Survey also asked data miners to share best practices in overcoming several key data mining challenges. This page contains a summary of their responses and links to the full verbatim responses (http://www.rexeranalytics.com/Overcoming_Challenges.html).

I have also added you to our Data Miner Survey distribution list:

- Later this Fall, when we complete the full summary report of the 2011 Survey, we will send you the full report.
- In the Spring of 2012 we will send you a request to participate in the 2012 survey (and a request to help us spread the word to other data miners).

Please contact us if you have any questions about this research program or if your company needs analytic assistance. Rexer Analytics is a Boston-based firm specializing in data mining and analytic CRM consulting. More information about Rexer Analytics is available at www.RexerAnalytics.com.

I also want to alert you that some friends of mine have recently written a text mining book. It will be published within the next couple months and available in January. Here's the amazon link: http://amzn.com/012386979X. Title: Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications by Gary Miner et al.

Best regards

-- Karl

Même courriel d'origine que pour Seth Grimes

6.2 Profil de la Société québécoise d'information juridique

Le site web de la Société québécoise d'information juridique (SOQUIJ) permet aisément de faire son profil. Elle fut fondée en 1976 par l'Assemblée nationale et relève du ministre de la Justice du Québec. Elle assure son autofinancement par la vente de ses produits et services.

Le rapport annuel de gestion 2009-2010 décrit son mandat : « … *de promouvoir la recherche, le traitement et le développement de l'information juridique en vue d'en améliorer la qualité et l'accessibilité au profit de la collectivité*. *La Société diffuse et commercialise chaque année une gamme étendue de produits, sous forme de publications imprimées ou électroniques (banques en ligne), auprès de la communauté juridique, du milieu des affaires et du travail et de tout citoyen intéressé à la documentation juridique. SOQUIJ met également à la disposition du grand public, sans frais, les jugements des tribunaux du Québec sur son site Internet*. »

En 1976, SOQUIJ publiait déjà les jugements intéressants qui provenaient des greffiers des tribunaux judiciaires. Les jugements non retenus pour publication devaient être détruits.

En 1987, SOQUIJ et le ministère de la Justice envisagent la création d'une banque de données électroniques. Au début des années 2000, la technologie permet à SOQUIJ d'offrir aux Québécois l'accès à toutes les décisions rendues par les tribunaux judiciaires et certains tribunaux administratifs.

En 2008, un article vient réglementer la diffusion de l'information et sur la protection des renseignements personnels, les organismes publics rendant des décisions motivées dans l'exercice de fonctions juridictionnelles doivent expédier leurs décisions à SOQUIJ, qui a l'obligation de les diffuser.

En 2010, SOQUIJ a reçu plus de 80 000 décisions rendues par différentes instances. Elles sont toutes accessibles sur le site web et ainsi, le Québec est devenu la province canadienne qui diffuse le plus grand nombre de décisions rendues dans son territoire.

Mission de la Société québécoise d'information juridique

Le site web dicte les lignes suivantes pour la mission de la Société : « …

- *de recueillir, d'analyser, de diffuser et de publier l'information juridique en provenance des tribunaux et des institutions,*
- *de présenter cette information sous la forme la plus complète, la plus à jour, la mieux organisée et la plus facile d'accès et*
- *d'offrir une expertise sans égale, des outils de recherche conviviaux, des contenus exhaustifs et un service à la clientèle des plus performants,*

au bénéfice de ses clients des milieux juridiques, des affaires et du travail ainsi que pour le public en général. »

Le traitement de l'information juridique et les interventions de SOQUIJ sont conçus afin de répondre aux attentes de sa clientèle, de faciliter ses tâches. Le site web SOQUIJ est une ressource indispensable pour les professionnels du droit. En

quelques clics, ils obtiennent avec AZIMUT la documentation juridique en ligne, la plus récente du Québec. Ils ont accès à toute l'information concernant les produits et les services en ligne à partir du même endroit: « *le service aux utilisateurs, la section À signaler, les dates de formation, le catalogue en ligne* et finalement l'accès à *AZIMUT »*.

Quelle est la différence entre SOQUIJ et AZIMUT?

SOQUIJ est comme une maison d'édition dans le domaine juridique qui publie la jurisprudence des tribunaux du Québec. Elle publie ses documents sur des supports papier et électroniques. Parmi ses produits, le service AZIMUT permet d'effectuer des recherches dans les banques de données en ligne. La section AZIMUT est l'outil qui permet aux clients d'accéder à différentes banques de données, soit *Juris.doc, Plumitifs* et *le Code civil du Québec annoté Baudouin Renaud***.**

6.3 L'éthique de l'exploration de données

Les quatre problèmes d'éthique avec l'exploration de données de Payne et Trumbach, (2009) sont les suivants.

Protection

Le premier problème d'éthique semble anodin : protéger les données. Les entreprises doivent se protéger de trois problèmes principaux : « l'accès non autorisé par les employés, l'accès non autorisé par des intrusions et l'emploi non éthique des données par les employés ». Les coûts concernant l'acquisition des données soit internes (par exemple les données clients recueillies avec le temps) et externes (par l'achat de données pour parfaire des connaissances sur un sujet en particulier) coûtent cher. Le coût pour entreposer et protéger les vastes données d'exploration est dispendieux. La facilité d'entreposage s'est aussi améliorée, les disquettes de 3,5 pouces pouvaient contenir des centaines de pages de documents imprimés, un lecteur flash peut contenir autant ou plus qu'un disque dur entier qu'il y a une décennie. La capacité de sauvegarder, de copier et de transporter des documents est sans précédent par les lecteurs flash, les téléphones intelligents, les ordinateurs portables, etc. L'entreprise doit prendre conscience que les données sont présentes et doivent être protégées. Une menace a fait changer certaines mentalités sur la protection et la conservation des documents : la loi Sarbannes-Oxley. Maintenant les documents doivent être conservés. Découlant de cette conservation, Barker et al. (2009) soulignent qu'en cas de litige entre une entreprise (qui a 99 % de ses documents numériques) et un tiers, la politique de conservation peut faire la différence entre une défense réussie et une perte coûteuse. Il souligne aussi qu'auparavant les tonnes de documents parlaient que

très peu, à moins d'avoir des analystes minutieux, maintenant les outils de TM peuvent passer au travers plus rapidement.

Conclusions incorrectes

Les applications d'exploration de données sont de plus en plus simples à utiliser. Un analyste se branche à la base de données avec l'application, suit les étapes une à une et peut obtenir un résultat. Il peut donc se glisser des erreurs lors des analyses. De plus, selon le niveau de précision accepté, il peut se glisser des faux positifs ainsi que des vrais négatifs. Un bon client pourrait être catégorisé comme étant mauvais ce qui ferait perdre des ventes… et un mauvais client pourrait être catégorisé comme étant bon ce qui a pour effet d'augmenter les mauvaises créances.

Autre fin que le but premier

La collection des données amène un autre problème d'éthique : l'utilisation autre que celle du but premier. Par exemple, une personne qui s'inscrit dans un concours et fournit ses informations personnelles ; ces données peuvent-elles être utilisées pour connaître d'où provient la clientèle ? Ce point en soulève trois : la vie privée, le consentement et la propriété des données. Est-ce que le client est conscient que ses données seront utilisées par exemple pour une catégorisation géographique de la clientèle afin d'implanter de nouveaux sites. Une utilisation autre peut être inappropriée. Je me souviens d'un film, celui qui se souvient du titre me le fasse parvenir, deux jeunes remplissent une fiche avec un nom fictif chez le marchand de glace pour obtenir un cornet gratuit. Il s'en suit une guerre et le gouvernement veut enrôler ce jeune homme fictif!

Respect de la vie privée

Les individus ont droit au respect de la vie privée et ceci est décrit par le groupe consultatif inter-agences en éthique de la recherche du gouvernement du Canada comme étant: « *Le respect de la vie privée est une valeur fondamentale, vue par beaucoup comme essentielle à la protection et à la promotion de la dignité humaine. En conséquence, l'accès aux renseignements personnels ainsi que le contrôle et la diffusion de telles informations ont une importance considérable pour l'éthique de la recherche.* .» Les points importants sont l'accès, le contrôle et la diffusion. C'est sur ce même sujet que Ann Cavoukian (1998), commissaire à la vie privée en Ontario, relève que les entreprises doivent protéger les données recueillies afin d'assurer la vie privée des gens. Au Québec, il n'y a pas de lignes directrices spécifiques à l'exploration des données, mais la loi s'applique à tous les renseignements personnels qu'ils soient individuels, nominatifs, en liste ou dans une base de données. La définition d'un renseignement personnel est: *« information qui permet d'identifier une personne* ». Donc, si un croisement, un tri, une sélection dans une base données permet d'identifier une personne, ce sont des renseignements personnels et la loi s'applique. (Courriel reçu de la Commission d'accès à l'information).

Finalement, Wright et al. (2008) identifient un manque de sensibilisation du public envers ce que les entreprises possèdent comme information sur eux et des moyens qu'ils peuvent utiliser pour mieux les connaître. Le public a l'illusion que ses données sont en sécurité et que personne, ni aucune entreprise, peut les connaître parfaitement. Un exemple, une entreprise qui achète votre dossier de crédit au complet (accès), qui connaît votre dossier médical électronique au complet et qui revend (diffusion) ces

informations. Cette information est acheminée à un futur employeur pour y attester de votre bonne santé financière et médicale et ce, sans que vous le sachiez (respect de la vie privée). Cela représente un futur inquiétant et ce, sans compter ce que l'on peut tirer comme analyse statistique en croisant toutes ces données.

6.3.1 Courriel – Commission d'accès à l'information

Courriel reçu de cai.communications@cai.gouv.qc.ca

Site Internet: www.cai.gouv.qc.ca

Bonjour monsieur Tardif,

Il n'y a pas de lignes directrices spécifiques à l'exploitation des données, mais la loi s'applique à tous les renseignements personnels qu'ils soient individuels, nominatifs, en liste ou dans une base de données. Définitions d'un renseignement personnel: *information qui permet d'identifier une personne*. Si un croisement, un tri, une sélection dans une base de données permet d'identifier une personne, ce sont des renseignements personnels et la loi s'applique.

Esther Turcotte
Service des communications
Commission d'accès à l'information
575, rue St-Amable, bur. 1.10
Québec (QUEBEC)
G1R 2G4

>>> "Alex Julie" < > 16:30 2010-10-19 >>>

Bonjour, Je rédige mon mémoire de maîtrise en Stratégie de l'intelligence d'affaires à l'Université de Sherbrooke. Je traite de l'exploration des données textuelles (text mining) dans le cadre de la prise de décision dans un contexte analytique. Je suis tombé sur un texte écrit par une de vos collègues en Ontario Ann Cavoukian. Il a pour titre : "*Data Mining: Staking A Claim on Your Privacy*" écrit en 1998. J'aimerais savoir si nous avons au Québec des lignes directrices concernant l'exploration de données (data mining) commercial et la vie privée.

Merci à l'avance,

Alexandre Tardif / Étudiant à la Maîtrise en Administration /Stratégie de l'intelligence d'affaires Université de Sherbrooke

6.4 Méthodologie de l'exploration de données

Voici les différentes méthodologies de l'exploration de données

6.4.1 SEMMA

La méthodologie **SEMMA** (*sample, explore, modify, model, assess)* peut se traduire en français par : échantillonner, explorer, modifier, modéliser, évaluer).

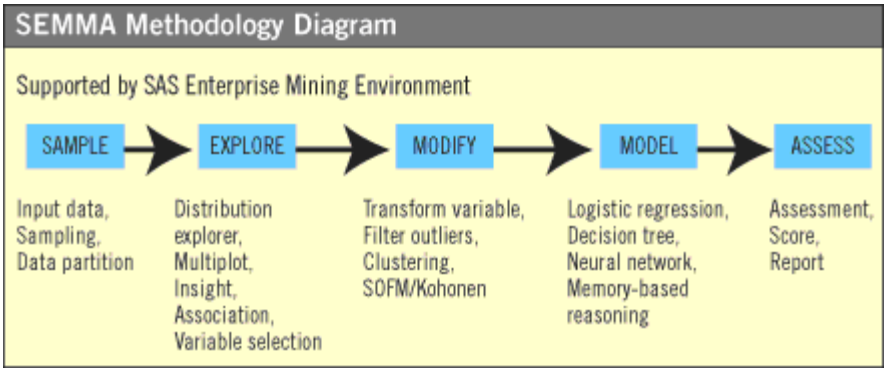


Figure 37 : Radha, R. (2008) Modèle Semma

Sampling = Échantillonner :

Accéder aux données, tirer un échantillon significatif, créer des données d'apprentissage, de validation et de test

Exploration = Explorer :

Devenir familier avec les données et découvrir

Manipulation = Manipuler :

Préparer, nettoyer les informations, coder, grouper des attributs

Modelling = Modéliser :

Construire des modèles (statistiques, réseaux de neurones, arbres de décisions, règles associatives, etc.)

Assessment = Valider :

Comprendre, valider, expliquer, répondre aux questions

Le SEMMA est basé uniquement sur le travail d'exploration de données à effectuer. Il débute par un échantillonnage statistiquement représentatif des données, il continue par l'application des techniques exploratoires statistiques, de visualisation, de choix et transformation des variables prédictives les plus significatives, de modélisation des variables pour prévoir des résultats, et de confirmation de l'exactitude d'un modèle. Il ne s'intéresse par au problème d'affaires, mais seulement aux données et aux résultats.

6.4.2 CRISP-DM

Le CRISP-DM est une méthode en six étapes. Il permet d'effectuer le travail d'exploration de données et de l'inclure dans une problématique d'affaires.

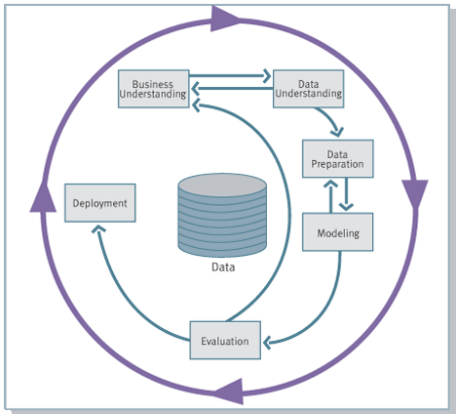


Figure 38 : Shearer, C. (2000). Modèle CRISP-DM

Les étapes principales sont les suivantes :

Business Understanding = Comprendre les affaires :

Formaliser un problème d'affaire que l'organisation cherche à résoudre en termes de données et mettre en place un plan initial pour réaliser cet objectif

Data Understanding = Comprendre les données :

Accéder aux données, les comprendre et utiliser les données appropriées

Data Preparation = Préparer les données :

Préparer les données pour les traitements et l'utilisation future

Modeling = Modéliser :

Modéliser les données avec les différents algorithmes d'analyse

Evaluation = Évaluer et valider :

Évaluer et valider les connaissances extraites

Deployment = Déployer :

Mettre en place les modèles analysés dans l'entreprise

Le CRISP-DM est une méthodologie connue qui décrit les étapes de définition, de développement et d'implémentation d'un projet d'exploration de données en incluant la problématique d'affaires. Le CRISP-DM ne fournit pas d'instruction pour définir l'étendue (scope) d'un projet et le planifier. Il inclut le principe de continuité et de répétition des tâches.

6.4.3 Cycle vertueux (Virtuous Cycle)

Le cycle vertueux est une méthode en quatre étapes. Il est très simple et très intuitif. Comme la méthodologie précédente, il permet d'effectuer le travail d'exploration de données et de l'inclure dans une problématique d'affaires.

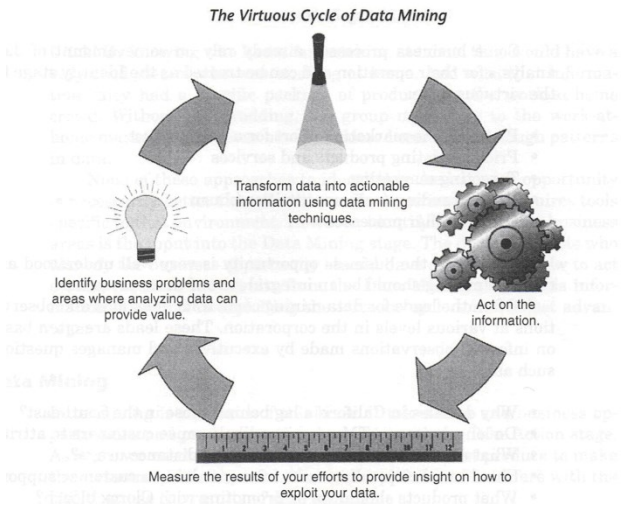


Figure 39 : Berry et Linoff (2004). Modèle Cycle Vertueux

Identify the business problem = Identifier le problème d'affaires

Formaliser un problème d'affaire que l'organisation cherche à résoudre en termes de données, et mettre en place un plan initial pour réaliser cet objectif.

Use data mining techniques to transform the data into actionable information. = Utiliser les techniques d'exploration de données pour transformer les données en information.

Accéder aux données, les comprendre et utiliser les données appropriées

Préparer les données pour les traitements et l'utilisation future

Modéliser les données avec les différents algorithmes d'analyse

Évaluer et valider les connaissances extraites

Act on information = Agir avec l'information

Mettre en place les modèles analysés dans l'entreprise

Measure the results = Mesurer les résultats

Évaluer et valider les résultats

Le cycle vertueux une méthodologie connue qui comme la précédente décrit les mêmes étapes, mais en rajoute une qui est de mesurer les résultats des modèles implantés. Cet ajout permet le calcul du ROI des projets d'exploration de données.

6.4.4 Six Sigma (DMAIC)

Le DMAIC (Define, Mesure, Analyse, Improve, Control) est une méthode structurée, orientée donnée, dont le but est l'élimination des défauts, des redondances et des problèmes de contrôle qualité de toutes sortes. Cette méthodologie étant générique, nous n'irons pas plus loin.

6.5 Courriels pour l'obtention du corpus

15/01/2010

Date: Fri, 15 Jan 2010 10:58:43 -0500 [2010-01-15 10:58:43 EDT]
De: Harpin, Geneviève < >
À: Alexandre Tardif < >
Objet: RE: Données juridiques...
Salut Alex,
Nous chez SOQUIJ, nous fournissons bien sûr tous les jugements sous forme électronique (encore que c'est possible d'avoir certains jugements qui sont rassemblés sous forme de recueil), mais si tu vas à la Cour tu peux avoir une copie papier. Quant à savoir s'il y a des analyses qui ont été faites là-dessus, alors là je n'en sais rien.....
Geneviève Harpin
Analyste en droit

Direction de l'information juridique
SOQUIJ
514 842-8741 poste 418

De : Alexandre Tardif []
Envoyé : 15 janvier 2010 10:49
À : Harpin, Geneviève
Objet : Données juridiques...
Bonjour Genevieve, (Excuse-moi les accents passent pas bien sur le courriel de l'université)
Comme tu dois le savoir déjà, j'entreprends ma recherche en intelligence d'affaires et mon champ de recherche est le forage de données. Je me demandais sous quel format les jugements des différentes cours étaient sauvegardés, électronique ou papier, et si c'est électronique, est-ce qu'il y a des analyses faites sur cette masse d'information?
Alexandre

20/01/2010

Date: Wed, 20 Jan 2010 11:19:48 -0500 [2010-01-20 11:19:48 EDT]
De: David, Hélène
À: Alexandre.Tardif>
Objet: accès aux données de SOQUIJ

Bonjour,
On m'a transmis votre courriel. Afin d'être en mesure de répondre à votre demande j'aimerais savoir quels sont vos besoins qu'elles soient les informations que vous voudriez accéder? Résumés, textes intégraux, les deux? Voulez-vous un accès sans frais à nos banques ou désirez-vous recevoir les informations dans un format donné?
La banque Textes intégraux de SOQUIJ contient à ce jour 611 568 décisions. Chaque décision a en moyenne 15 pages. Vous en retrouvez plusieurs sur le site jugements.qc.ca. Ce site offre un accès sans frais aux décisions.
La banque résumés quant à elle contient 138 889 résumés de décisions à ce jour.
J'attends donc de vos nouvelles.
Bonne fin de journée!
Hélène David, avocate
Conseillère d'affaires juridiques et responsable de l'accès et de la protection de l'information SOQUIJ

25/01/2010

Date: Mon, 25 Jan 2010 11:13:43 -0500 [2010-01-25 11:13:43 EDT]
De: Champagne, Daniel
À: Alexandre.Tardif
Bonjour M. Tardif,
Me David m'a fait part de votre demande d'accès à certaines de vos données dans le cadre de votre projet de maîtrise. Dans le but d'éviter un long échange de courriels pour préciser vos attentes et nos questionnements, je vous invite à me téléphoner au numéro indiqué ci-dessus, au moment qui vous conviendra.

Daniel Champagne, avocat
Directeur de l'information juridique
SOQUIJ

05/02/2010

Date: Fri, 5 Feb 2010 11:49:31 -0500 [2010-02-05 11:49:31 EDT]
De: Champagne, Daniel
À: Alexandre Tardif
Objet: RE: Appel

C'est parfait.
Daniel Champagne, avocat
Directeur de l'information juridique
SOQUIJ

-----Message d'origine-----
De : Alexandre Tardif
Envoyé : 5 février 2010 10:45
À : Champagne, Daniel
Objet : RE: Appel
Rebonjour, juste après dîner, 13;30?
Alexandre

Selon "Champagne, Daniel":
Bonjour, Certainement. À ce moment-ci, je suis disponible entre 9h00 et 15h00. SVP me préciser le plus tôt possible l'heure de notre rencontre. Bonne journée!
Daniel Champagne, avocat
Directeur de l'information juridique
SOQUIJ

-----Message d'origine-----
De : Alexandre Tardif
Envoyé : 5 février 2010 10:11
À : Champagne, Daniel
Objet : RE: Appel
Bonjour,
Est-ce que je peux aller vous rencontrer mercredi 10 février?
Alexandre

10/02/2010

Date: Wed, 10 Feb 2010 16:04:49 -0500 [2010-02-10 16:04:49 EDT]
De: Alexandre Tardif
À: olivier.caya
Cc: Jessica.Levesque
Objet: SOQUIJ

Bonjour à vous deux,
Je reviens d'une rencontre à la SOQUIJ. J'y ai rencontré pendant environ 45 minutes Daniel Champagne (Directeur de l'information juridique). Il est ouvert pour une recherche, mais me mentionne qu'il n'a pas d'argent et pas de ressources pour me soutenir.
Ils ont les verdicts de cours en formats numériques (fichier texte) et ils utilisent déjà un outil de résumé de leur texte. Deux sujets semblent ressortir d'eux-mêmes de la masse. Le premier, l'étude statistique des clients en décroissance, voir les liens ce qu'ils abandonnent. Le second aider la recherche de verdict de leur base de données, la recherche se fait actuellement par mot clé, et donc un potentiel de passer à côté d'un texte important pour une jurisprudence pour un client et une vente en moins pour la SOQUIJ (un genre Amazon : les autres clients ont aussi consulté ou bien un clustering de texte de même sujet).
Du côté SOQUIJ, M. Champagne doit présenter le projet au niveau hiérarchique supérieur au sien et s'attend à une demande de recherche clairement présentée sous peu. On s'en reparle, je suis à l'Université jeudi et vendredi.
Alexandre

26/02/2010

Date: Fri, 26 Feb 2010 10:34:35 -0500 [2010-02-26 10:34:35 EDT]
De: Champagne, Daniel
À: Alexandre Tardif
Objet: RE: Accès et formation...

Bonjour Alexandre,
Voici le Code d'accès : XXX
Et le Mot de passe : XXX
Le code sera automatiquement fermé le 31 mai 2010, à la fin de la journée. Par ailleurs, tu dois me confirmer que ce code ne sera utilisé qu'aux fins de ta recherche et que tu prendras connaissance de la licence d'utilisation.
Daniel Champagne, avocat
Directeur de l'information juridique
SOQUIJ

25/05/2010

Date: Mon, 10 May 2010 14:32:29 -0400 [2010-05-10 14:32:29 EDT]
De: Champagne, Daniel
À: Alexandre Tardif
Objet: RE: Semaine prochaine...

Parfait! On se voit donc le 21 mai à 12h30.
Daniel Champagne, avocat
Directeur de l'information juridique
SOQUIJ

15/06/2010

Date: Tue, 15 Jun 2010 10:03:11 -0400 [2010-06-15 10:03:11 EDT]
De: Champagne, Daniel
À: Alexandre Tardif
Cc: Harvey, Richard
Objet: RE: Avancement

Bonjour Alexandre,
Comme je te l'avais laissé déjà entendre, et après étude plus approfondie des données disponibles, la deuxième option de projet que tu proposes ne sera pas possible.
Par contre, si tu veux aller de l'avant avec la première option, fais-moi signe et nous préparerons des fichiers contenant des textes intégraux de jugements et les résumés et indexation correspondants.
J'attends de tes nouvelles.
Daniel Champagne, avocat
Directeur de l'information juridique
SOQUIJ

28/06/2010

Date: Mon, 28 Jun 2010 10:52:10 -0400 [2010-06-28 10:52:10 EDT]
De: Carré, Johanne
À: Alexandre Tardif
Objet: RE: TR: TR: Envoi de documents

Bonjour,
Vous recevrez les fichiers la semaine prochaine, car la personne chargée de les produire est en vacances cette semaine.
Johanne Carré
Coordonnatrice Réception des jugements,
documentation et édition
Direction de l'information juridique

Société québécoise d'information juridique

-----Message d'origine-----
De : Alexandre Tardif
Envoyé : 25 juin 2010 12:42
À : Carré, Johanne
Objet : Re: TR: TR: Envoi de documents
Bonjour,
Tel que mentionné dans mon courriel précédent, le format et le corpus me conviennent et vous pouvez préparer les fichiers.
Alexandre

09/07/2010

Date: Fri, 9 Jul 2010 16:19:56 -0400 [2010-07-09 16:19:56 EDT]
De: Champagne, Daniel
À: Alexandre Tardif
Objet: RE: Avancement
Partie(s): 2 ENGAGEMENT DE CONFIDENTIALITÉ.doc 42 Ko

Bonjour Alexandre,
Nous sommes prêts à te transmettre les données. À quelle adresse devons-nous l'envoyer? Ci-joint , un engagement à signer. Merci de me le faire parvenir!
Bonne fin de semaine!
Daniel Champagne, avocat
Directeur de l'information juridique

6.6 Matériel - définition des catégories

Voici quelques explications concernant les critères utilisés pour l'évaluation de l'application.

6.6.1 Performance

La performance est la capacité à gérer une variété de sources de données d'une manière efficace. La configuration matérielle (hardware) a un impact majeur sur l'application du point de vue informatique. En outre, certains algorithmes d'exploration de données sont intrinsèquement plus efficaces que les autres. Cette catégorie met l'accent sur les aspects qualitatifs de la capacité d'un outil de gérer facilement des données sous une variété de circonstances plutôt que sur les variables de performance qui sont entraînés par des configurations matérielles et/ou inhérentes aux caractéristiques algorithmiques.

Tableau 10: Critères de performance **pour l'outil d'exploration de textes**

Critères	**Description**
Plate-forme de variétés	Est-ce que le logiciel peut s'exécuté sur une grande variété de plates-formes informatiques?
Architecture logicielle	Est-ce une utilisation de logiciels architecture client-serveur ou une seule architecture stand-alone? Est-ce que l''utilisateur a le choix d'architectures?
Accès hétérogène des données	Dans quelle mesure l'interface du logiciel a accès à une variété de sources de données (RDBMS, ODBC, etc)? Est-il besoin d'un logiciel auxiliaire de le faire? L'interface est transparente?
Taille des données	Dans quelle mesure l'application permet l'utilisation de grands ensembles de données? La performance est-elle linéaire?
Efficacité	Le logiciel produit-il des résultats dans un délai raisonnable par rapport à la taille des données, les limites de l'algorithme, et d'autres variables?
Interopérabilité	Est-ce que l'interface de l'application fonctionne avec d'autres applications?
Robustesse	L'application s'exécute systématiquement sans se planter?

6.6.2 Fonctionnalité

La fonctionnalité est l''inclusion d'une variété de capacités, des techniques et méthodologies pour l'exploration de données.

Tableau 11: Critères de *fonctionnalité* pour l'outil d'exploration de textes

Critères	Description
Algorithmique variée	Est-ce que l'application fournit une diversité suffisante de techniques d'exploration de données et d'algorithmes?
Méthodologie prescrite	Est-ce que l'application propose une méthode étape par étape pour effectuer une exploration afin d'éviter des résultats erronés?
Validation du modèle	Est-ce que l'application encourage la validation dans le cadre de la méthodologie?
Flexibilité du type de données	Est-ce que les algorithmes pris en charge face peuvent travailler avec une grande variété de types de données?
Algorithme modifiable	Est-ce que l'utilisateur a la possibilité de modifier et d'affiner les algorithmes de modélisation?
Données d'échantillonnage	L'outil permet-il un échantillonnage aléatoire des données pour la modélisation prédictive?
Rapport	Les résultats d'une analyse peuvent-ils se présenter sous différentes formes? L'application fournit-elle un résumé des résultats ainsi que les résultats détaillés?
Exportation du modèle	Après qu'un modèle soit validé, l'application fournit-elle une variété de moyens pour exporter le modèle.

6.6.3 Convivialité

La convivialité est l'adaptation de l'application aux différents niveaux et types d'utilisateurs et ce sans perte de fonctionnalité ou d'utilité. Un problème avec les applications faciles à utiliser c'est leur mauvaise utilisation. Non seulement il devrait être un outil facile à apprendre, il doit aider à orienter l'utilisateur vers la bonne utilisation de l'exploration de données. Les utilisateurs en règle générale ajustent leur modèle afin de générer des modèles plus valables. Un bon outil fournira des diagnostics utiles pour déboguer les problèmes et améliorer le rendement.

Tableau 12: Critères de *convivialité* pour l'outil d'exploration de textes

Critères	Description
Interface utilisateur	L'interface utilisateur est-il facile à naviguer et simple? La présentation des résultats dans une interface est faite de manière significative?
Courbe d'apprentissage	L'application est-elle facile à apprendre? L'application est-elle facile à utiliser correctement?
Types d'utilisateurs	L'application est conçue pour les débutants, intermédiaires, les avancés ou une pour combinaison des types? Comment bien adapté est l'outil pour son type d'utilisateurs ciblés? Quelle est la facilité de l'outil pour les analystes à utiliser? Est-il aussi facile à utiliser pour les utilisateurs finaux?
Visualisation des données	Comment bien l'application présente les données? Comment présente-t-elle les résultats de la modélisation? Y a-t-il une variété de méthodes graphiques utilisées pour communiquer l'information?
Rapport d'erreurs	Est-ce que le rapport d'erreurs est significatif? Est-ce que les messages d'erreur aident l'utilisateur à déboguer ses problèmes?
Historique	L'application maintient-elle un historique des actions prises dans le processus? L'utilisateur peut-il modifier ce processus et ré-exécuter le script?
Domaine varié	Dans quelle mesure l'application est spécifique à un domaine?

6.6.4 Travail de soutien

L'utilisateur doit exécuter une variété de tâches préalables, tel le nettoyage des données, l'extraction, la transformation, la visualisation et autres pour ultimement faire l'exploration de données. Ces tâches comprennent la sélection des données, le nettoyage, l'enrichissement, la substitution de valeur, des données de filtrage, etc. Il est rare qu'un ensemble de données soit vraiment propre et prêt à l'exploration de données, l'utilisateur doit être en mesure de facilement raffiner les données pour la phase de construction du modèle.

Tableau 13: Critères de *travail de soutien* pour l'outil d'exploration de textes

Critères	Description
Nettoyage de données	Est-ce que l'application permet à l'utilisateur de modifier les valeurs des parasites dans l'ensemble des données ou effectuer d'autres opérations de nettoyage des données?
Valeur de	L'application permet-elle la substitution d'une valeur globale de

substitution	données avec un autre (c.-à-d., remplacer «F M avec 1 ou 0 pour l'uniformité)?
Filtrage des données	L'application permet-elle la sélection des sous-ensembles de données fondées sur des critères de sélection définis utilisateur?
Attributs	L'application permet-elle la création d'attributs dérivés basés sur les caractéristiques intrinsèques? Y a-t-il une grande variété de méthodes disponibles pour calculer les attributs (par exemple, les fonctions statistiques, fonctions mathématiques, fonctions booléennes, etc.)?
« Randomisation »	L'application permet-elle de « randomiser » les données préalablement à la construction de modèles? Quelles sont l'efficacité et l'efficience de randomisation?
Suppression d'enregistrement	L'application permet-elle la suppression d'enregistrements qui peuvent être incomplets ou peut fausser la modélisation (outlier)? Comment effectue-t-elle cette tâche?
Manipulation des vides	Est-ce que l'application manipule bien les vides? L'application permet-elle les vides d'être substitués par une variété de valeurs dérivées (par exemple, moyenne, médiane, etc.)? L'application permet-elle aux vides d'être substitués par une valeur définie par l'usager? Si oui, cela peut-il être fait globalement ainsi que valeur par valeur?
Manipulation des métadonnées	L'application présente-t-elle à l'utilisateur les métadonnées avec les descriptions des données, etc.? Dans l'affirmative, l'application permet-elle à l'utilisateur de manipuler ces métadonnées?
Résultat d'évaluation	L'application permet-elle aux résultats d'une analyse d'être renvoyé dans une autre analyse pour la construction de modèles supplémentaires?

6.6.5 Autres critères

L'utilisateur doit dans le cadre d'exploration de données textuelles effectuer des tâches supplémentaires, certaines d'entre elles peuvent être automatisées. Dans le cadre du projet, deux tâches sont essentielles, soit la classification supervisée des textes ainsi que le « clustering ». Nous retrouvons aussi d'autres critères pour départager les applications.

Tableau 14: Critères *autres* pour l'outil d'exploration de textes

Critères	Description
Coût	Quel est le coût par licence?

Segmenter	L'application divise-t-elle le texte en phrases?
« Tokéniser »	L'application scinde-t-elle un texte en unité lexicale élémentaire?
Analyser morphologiquement	L'application renvoie-t-elle la forme normalisée (le lemme) et les catégories grammaticales potentielles pour tous les mots identifiés durant la phase de « tokenisation »?
Désambiguïser les mots	L'application détermine-t-elle la catégorie grammaticale exacte d'un mot en fonction de son contexte?
Interroger un dictionnaire	L'application identifie-t-elle le contexte d'un mot pour l'orienter vers l'entrée du dictionnaire correspondant?
Reconnaître des expressions idiomatiques	L'application reconnaît-elle les expressions présentes dans un texte?
Windows 7 64 bits	L'application fonctionne-t-elle sur un client Windows 7 64 bits?
« Clustering »	L'application classe-t-elle et regroupe-t-elle automatiquement les documents en fonction de leur ressemblance et de leur proximité thématique?
Catégorisation	L'application classe-t-elle automatiquement les documents dans des catégories prédéfinies?

6.7 Commentaires du conseil d'éthique de la recherche

« Il n'y a pas lieu de faire évaluer par un CÉR la recherche fondée exclusivement sur de l'information accessible au public si l'une ou l'autre des conditions suivantes est remplie:

a) l'information est légalement accessible au public et adéquatement protégée en vertu de la loi;

b) l'information est accessible au public et il n'y a pas d'attente raisonnable quant à la protection de la vie privée.

Application aux fins de la politique, sont considérés comme de l'information accessible au public les documents, fichiers ou publications existants qui peuvent ou non contenir des renseignements identifiables et que la loi traite comme étant

accessibles au public ou légalement accessibles au public, sous réserve de protections adéquates.

Certains genres d'information sont légalement accessibles au public sous une certaine forme et à certaines fins, comme le prévoient souvent des lois ou règlements. C'est le cas par exemple des registres de décès, ***des jugements des tribunaux****, des archives publiques et des statistiques accessibles au public (comme les fichiers à grande diffusion de Statistique Canada). Toutes les archives accessibles au public (nationales, provinciales ou municipales) font l'objet de politiques régissant l'accès à leur contenu. Une pièce d'archives ou une base de données qui est soumise à des restrictions, par exemple les restrictions prévues par les lois sur l'accès à l'information et la protection des renseignements personnels ou imposés par le donateur des documents, peut néanmoins être considérée aux fins de la Politique comme étant accessible au public. »*

Les données de ce mémoire les données proviennent de la SOQUIJ, elles sont des résumés de jugements de tribunaux publics accessibles à tous. La mission de la Société d'information juridique du Québec lui mandate à diffuser et publier l'information au bénéfice de ses clients qui proviennent des milieux juridiques, des milieux des affaires et ainsi que pour le public en général. Et donc, avec ces données publiques nous n'avons pas de souci au niveau de la provenance et ainsi que de l'utilisation des données.

6.8 Applications d'exploration de textes considérés

Tableau 15: Descriptif des différents outils d'exploration de textes

Compagnie	Nom de la suite	Site Web	Description
			Commercial
Rocket Software	Rocket AeroText	http://www.rocketsoftware.com/	Rocket Software est une suite de module d'exploration de données textuelles. Dans la suite de produits il y a Rocket AeroText qui analyse les textes par : entité, phrase clé, phrase grammaticale, entité association, catégorisation et résolution temporelle. Il ne supporte pas le français, il supporte l'anglais, l'arabe, le chinois, l'espagnol et le bahasa (Indonésie)
Attensity	Attensity360 Attensity intelligence	http://www.attensity.com/	Attensity 360 utilise la technologie de traitement du langage naturel pour analyser les médias sociaux et forums, les courriels, la gestion de la relation client, les e-services et autres analyses d'information textuelle L'implantation a coût élevé (plus de 100 000$) et ne supporte pas le français.
Autonomy	Analytics & Taxonomy	http://www.autonomy.com/	Autonomy propose une suite d'exploration de données textuelles, de clustering et de catégorisation supervisée pour une variété d'industries. L'implantation a coût élevé (plus de 100 000$).
Basis Technology	Rosette7	http://www.basistech.com/	Basis Technology fournit des solutions pour l'extraction de renseignements utiles à partir de textes non structurés dans différentes langues. Il aide les entreprises et les gouvernements, mais pas les étudiants.
Expert System S.p.A.	Cogito	http://www.expertsystem.net/	Expert System offre des solutions novatrices et flexibles qui peuvent être personnalisées pour répondre aux besoins du client. Il analyse en temps réel les nouvelles, les pages Web, la correspondance par courriel, RSS, blogues, etc. De plus, il traite les données non structurées, par exemple dans les fichiers texte, les documents Word et les documents PDF, etc.

FICO	FICO Predictive Analytics	http://www.fico.com/	Fico est un fournisseur de solutions analytiques spécialisé dans les services financiers.
Inxight		www.sap.com	Inxight regroupe six modules d'analyse de texte : • LinguistX : analyse de textes, • StarTree : hiérarchise des textes, • Summarizer : résume les textes, • ThingFinder : outil de langage naturel, • TableLens : visualise les tendances dans des grands volumes de données TimeWall : un outil de visualisation des évènements dans le temps. (Inxight a été acheté par Business Objects qui eux ont été achetés par SAP AG en 2008)
LexisNexis	LexisNexis Analytics	http://www.lexisnexis analytics.com/	La solution LexisNexis aide la prise de décision par l'analyse de textes par la couverture de presse et analyse de la presse. Ainsi, un décideur peut avoir une vue d'ensemble de la couverture médiatique de ses activités.
Nstein Technologies	Nstein TME 5	http://www.nstein.com/	TME est une solution d'exploration textuelle qui crée des métadonnées riches pour permettre aux éditeurs d'augmenter les pages vues, d'automatiser l'étiquetage, d'améliorer l'expérience de recherche, la productivité éditoriale augmenter. En combinaison avec les moteurs de recherche, il est utilisé pour créer des applications de recherche sémantique. (maintenent propriété de Open Text depuis le 22 -02- 2010)
Open Text	Open Text Content Analytics	http://www.opentext.com/	Open Text Content Analytics (anciennement Nstein Text Mining Engine (TME))
SAS	SAS Text Miner	http://www.sas.com/	SAS Text Miner (anciennement Teragram) analyse les textes, traite avec le langage naturel, et fait la taxonomie.
Silobreaker	Enterprise Software Suite	http://www.silobreaker.com/EnterpriseSoftware.aspx	Silobreaker fournit l'analyse de textes, de clustering, de recherches et des technologies de visualisation. (mash-up)

IBM	IBM SPSS Modeler Premium	http://www.spss.com/	IBM SPSS Modeler Premium extrait et classe les concepts à partir de données non structurées. Il intègre cette nouvelle source de données structurées dans vos efforts d'exploration de données afin de créer des modèles plus précis.
IBM	LanguageWare		LanguageWare est la nouvelle génération de plate-forme linguistique d'IBM. Elle a été conçue dès le départ pour répondre aux exigences posées par les applications mondiales d'aujourd'hui. http://www-01.ibm.com/software/globalization/topics/languageware/index.jsp
StatSoft	Statistica Text Miner	http://www.statsoft.com/	*STATISTICA Text Miner* est une extension de STATISTICA Data Miner et est idéale pour donner un sens aux données textuelles non structurées comme un grappage précieux pour la prise de décision.
Temis	Luxid	http://www.temis.com/	Les principaux avantages de Luxid sont **: Extraire des connaissances ciblées :** Pour détecter et extraire des informations stratégiques à forte valeur ajoutée à partir de données non structurées, les annotateurs multilingues identifient de manière précise les entités et les relations sémantiques. **Filtrer les données :** Pour cibler les documents les plus pertinents, les utilisateurs peuvent aisément affiner leurs requêtes en utilisant de multiples filtres contextuels. **Identifier les tendances :** Une large gamme de tableaux interactifs, tableaux croisés dynamiques, tableaux de bord et rapports permet de révéler le sens caché des données. Les utilisateurs ont un accès immédiat à de multiples vues et peuvent ainsi analyser l'information en profondeur en l'organisant et la croisant selon leur scénario d'analyse. **Naviguer dans la connaissance :** Pour révéler des dépendances entre les informations, les utilisateurs peuvent naviguer intuitivement à travers un réseau de connaissance constitué d'entités reliées par leur proximité statistique ou par des relations sémantiques. **Accélérer les découvertes collaboratives :** Pour favoriser le partage d'information au sein de l'entreprise, les utilisateurs disposent de tableaux de bord dynamiques et personnalisables.

Thomson Reuters	*Thomson Data Analyzer*	http://thomsonreuters.com/	*Thomson Data Analyzer* est un logiciel de bureau qui offre une puissante interface de gestion et d'extraction de connaissances. Il est destiné pour les brevets et les données scientifiques dans les bases de données internes ou externes. Grâce à sa technologie d'analyse avancée, il permet aux utilisateurs d'analyser les données sur les brevets et la littérature scientifique de toute base de données structurée directement de leur ordinateur.
			Logiciel libre
GATE	GATE	http://gate.ac.uk/	GATE (General Architecture for Text Engineering) est un outil d'ingénierie linguistique et de traitement de langage naturel.
UIMA		http://uima.apache.org/	UIMA - (Unstructured Information Management Architecture) est un cadre de composantes pour l'analyse des contenus non structurés tels que du texte, audio et vidéo. À l'origine développé par IBM.
Rapid-i	RapidMiner	http://rapid-i.com/	L'extension textuelle ajoute tous les opérateurs nécessaires à l'analyse statistique textuelle. Vous pouvez charger des textes de différentes sources de données ou de vos jeux de données, les transformer ensuite par un vaste ensemble de différentes techniques de filtrage, et enfin d'analyser vos données textuelles.
Carrot2.		http://project.carrot2.org/	Carrot^2 est un engin de clustering de résultats de recherche web. Il peut organiser automatiquement de petites collections de documents, par exemple des résultats de recherche web dans des catégories thématiques. Carrot^2 offre des composantes prêtes à utiliser pour retrouver les résultats de recherche provenant de diverses sources, y compris YahooAPI, GoogleAPI, MSN Live API, Lucene, SOLR, Google Desktop et autres.

6.9 Introduction pour certains algorithmes

Twostep

Le TwoStep est une méthode d'analyse de segmentation en deux (Two) étapes (step). La première consiste en une exploration des données visant à les compresser en sous-classes. La seconde est une classification hiérarchique permettant de fusionner progressivement les sous-classes en classes de plus en plus importantes. La technique TwoStep évalue le nombre de classes optimal pour les données d'apprentissage. Elle prend en charge de manière efficace des types de champ mixtes et des ensembles de données volumineux.

K-means

Le K-means définit un nombre de classes fixe et affecte à plusieurs reprises des enregistrements à des classes et ajuste les centres de classe et ce, jusqu'à ce que le modèle ne puisse plus être amélioré. Le modèle généré dépend essentiellement de l'ordre des données d'apprentissage et donc si les données sont réorganisées, le modèle final sera différent.

Kohonen

Le Kohonen représente un type de réseau de neurones classificateur. Il est également appelé « knet » ou cartes auto-organisatrices. Les unités de base sont les neurones. Elles sont organisées en deux couches : la couche d'entrée et la couche de sortie

(également appelée connexion de sortie). Tous les neurones d'entrée sont connectés à tous les neurones de sortie. Ces connexions ont une puissance ou une pondération associée. Au cours de l'apprentissage, chaque unité entre en compétition avec les autres pour «gagner» des enregistrements. La connexion de sortie est une grille de neurones à deux dimensions, sans connexion entre les unités.

C5.0

Le C5.0 peut traiter tout type de variable explicative et créer plus de deux branches lorsque le critère de séparation se base sur une variable nominale ou ordinale. Le fonctionnement d'un modèle C5.0 repose sur un découpage de l'échantillon basé sur le champ qui fournit le gain d'informations le plus important. Chaque sous-échantillon issu du premier découpage est de nouveau découpé. Ce processus est itératif jusqu'à ce que les sous-échantillons ne puissent plus être découpés. À la fin, les sous-échantillons sont réétudiés et ceux qui n'ont pas d'influence significative sur la valeur du modèle sont supprimés ou élagués. Le C5.0 ne génère pas seulement un arbre, mais aussi un ensemble de règles qui se veulent plus simples et générales que celles issues directement de l'arbre (il y a cependant parfois une perte en précision). Chaque « feuille » décrit un sous-ensemble particulier des données d'apprentissage ; chacune des observations contenues dans les données d'apprentissage correspond à un seul nœud terminal de l'arbre. Autrement dit, chacun des enregistrements présentés à l'arbre de décisions ne peut donner lieu qu'à une seule prévision.

Quest

Le Quest (Quick, Unbiased, Efficient Statistical Tree) est une méthode de classification binaire permettant de créer des arbres de décisions. L'une des principales raisons pour lesquelles cette méthode a été développée était de réduire le temps de traitement nécessaire aux analyses C&RT importantes, qui utilisaient alors de nombreuses variables ou observations. QUEST avait également pour objectif de limiter la tendance, observée parmi les méthodes d'arbre de classification, à favoriser les entrées autorisant un nombre supérieur de divisions, à savoir des champs d'entrée continus (intervalle numérique) ou ceux dotés de nombreuses catégories.

Réseau de neurones

Le réseau neurones est un modèle simplifié de la manière dont le cerveau humain traite les informations. Le fonctionnement de ce modèle repose sur la simulation d'un grand nombre d'unités de traitement interconnectées, qui sont en quelque sorte des versions abstraites de nos neurones. Ces unités de traitement sont organisées en couches. Il existe généralement trois types de couche dans un réseau de neurones : une couche d'entrée dans laquelle les unités représentent les champs d'entrée, une ou plusieurs couches cachées, ainsi qu'une couche de sortie dans laquelle des unités représentent les champs cibles. Les unités sont reliées entre elles par des connexions de puissance (ou de pondération) différentes. Les données d'entrée sont présentées dans la première couche et les valeurs sont transmises entre les neurones d'une couche à l'autre. Le résultat final est obtenu à partir de la couche de sortie.

Arbre C&R

L’arbre C&RT (classification et régression) est une méthode de classification et de prévision basée sur un système d'arborescence. Similaire au C5.0, cette méthode utilise la technique de partition récursive afin de diviser les données d'apprentissage en segments présentant des champs de sortie similaires. Il examine en premier lieu les champs d'entrée, afin de définir la meilleure segmentation. Celle-ci est mesurée en fonction de la réduction de l'index d'impureté résultant de la segmentation. Le découpage définit deux sous-groupes qui sont à leur tour découpés en deux nouveaux sous-groupes. Le découpage se poursuit jusqu'à ce que l'un des critères d'arrêt soit atteint. Toutes les divisions sont binaires (deux sous-groupes uniquement).

7 Références

7.1 Livres

Berry, M.J. et Linoff, G.S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd Edition.* Indianapolis, IN : Wiley Publishing Inc. ISBN: 978-0-471-47064-9

Cooper, D.R. et Schindler, P.S. (2000). *Business Research Methods*. McGraw-Hill College. ISBN: 978-0072314519

Davenport, T.H. et Prusak, L. (1998). *Working Knowledge: How Organizations Manage What They Know.* Harvard Business School Press. ISBN: 0-87584-655-6

Turban, E., Sharda, R. et Aronson, J. (2007). *Business Intelligence: A Managerial Approach*. Upper Saddler River, NJ : Prentice Hall ISBN: 013234761X ISBN-13: 9780132347617

Société Québecoise d'information juridique (2010). *Rapport Annuel de Gestion 2009-2010* (2010), Gouvernement du Québec. ISBN : 978-2-7642-712-3

7.2 Articles

Abd-Elrahman, A., Andreu, M., et Abbott, T. (2010). Using text data mining techniques for understanding free-style question answers in course evaluation forms, *Research in Higher Education Journal*, 9, 1-11.

Agrawal, R., Grandison, T., Johnson, C., et Kiernan J. (2007). Enabling the 21st century health care information technology revolution. *Association for Computing Machinery. Communications of the ACM*, 50(2), 34-42

Ahmed, S., Coenen, F. et Leng, P. (2006). Tree-based partitioning of date for association rule mining. *Knowledge and Information Systems*, 10(3), 315-331

Allentuck, A. (2006). Making statistical connections, managing the info glut. *Canadian Grocer*, March 120(2), 29-30

Anthes, G. (2009). Deep Data Dives Discover Natural Laws. *Communications of the ACM,* 52(11), 13

Apte, C., Liu, B., Pednault, E.P.D. et Smyth, P. (2002). Business applications of data mining. *Association for Computing Machinery, Communications of the ACM,* 45(8), 49

Bacon, J. et Webb, C. (2009). Mine Your Data. *Engineering and Mining Journal,* 210 (6), 30

Barker, R.M., Cobb, A.T. et Karcher, J. (2009). The legal implications of electronic document retention: Changing the rules. *Business Horizon*, 52, 77-186

Basit, H., et Jarzabek, S. (2009). A Data Mining Approach for Detecting Higher-Level Clones in Software. *IEEE Transactions on Software Engineering*, 35(4), p. 497-514

Bertrand-Gastaldy, S. et Lanteigne, D. (1995). La modélisation de l'analyse documentaire: à la convergence de la sémiotique, de la psychologie cognitive et de l'intelligence articifielle. *Canadian Association for Information Science; Proceedings of the 23rd Annual Conference / Association canadienne de sciences de l'information; Travaux du 23e congrès annuel. Connectedness: Information, Systems, People, Organizations, H.A. OLSON et D.B. WARD (ed.)*. Edmonton: University of Alberta, School of Library and Information Studies, 1995: 1-11

Better, M., Glover, F., et Samorani, M. (2010). Classification by vertical and cutting multi-hyperplane decision tree induction. *Decision Support Systems*, 48(3), 430

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R. et Ramanujam, K. (1997). Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery*, 1(1), 121-125

Brisson, L., Collard, M. et Pasquier, N. (2006). Ontologie et base de connaissances pour le pré-traitement et post-traitement en fouille de données. *Atelier Fouille de Données Complexe de la conférence EGC'2006 sur l'Extraction et la Gestion des Connaissances*, Lyon : France (2006)

Buck, N. (2001). Eureka! Knowledge Discovery. *Software Magazine - Westborough*, 20(6), 24-29

Carricano., M., et deLassence., G. (2009). Un usage du Text Mining : donner du sens à la connaissance client, *Systèmes d'Information et Management*, 14(2), 85-107

Cavoukian, A (1998). Data Mining: Staking A Claim on Your Privacy. *Report by Information and Privacy Commissioner, Ontario, Canada.*

Chan, P.K., Fan, W., Prodromidis, A.l. et Stolfo, S.J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems & Their Applications*, 14(6), 67-74

Chen, M.S., Han, J. et Yu, P.S. (1999). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, 8(6), 866-883

Chou, C.H., Sinha, A.P. et Zhao, H. (2008). A text mining approach to Internet abuse detection. *Information Systems and E-Business Management*, 6(4), 419-439

Choudhary, A., Oluikpe, P., Harding, J., et Carrillo, P. (2009). The needs and benefits of Text Mining applications on Post-Project Reviews. *Computers in Industry*, 60, 728-740

Chung, W., Chen, H. et Nunamaker, J.F.Jr. (2005). A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration. *Journal of Management Information Systems*, 21(4), 57

Cody, W.F., Kreulen, J.T., Krishna, V. et Spangler, W.S. (2002). The integration of business intelligence and knowledge management. *IBM Systems Journal*, 41 (4), 697-713

Cohen KB, Hunter L (2008). Getting started in text mining. *PLoS Comput Biol* 4(1)

Collier, K., Carey, B., Sautter, D. et Marjaniemi, C. (1999). A Methodology for Evaluating and Selecting Data Mining Software. *Proceedings of the 32nd Hawaii International Conference on System Sciences*, IEEE 0-7695-0001-3/99

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. et Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547.

Crowsey, M.J., Ramstad, A.R., Gutierrez, D.H., Paladino, G.W. et White, KP (2007). An Evaluation of Unstructured Text Mining Software. *IEEE Systems and Information Engineering Design Symposium*, 2007. SIEDS 2007, 1-6

Cui, G., Wong, M.L. et Lui, H.K. (2006). Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science*, 52(4), 597-613

Dahl, S. (2010). Current Themes in Social Marketing Research: Text-Mining the Past Five Years. *Social Marketing Quarterly*, 16(2), 128-136.

Dozier, C., et Jackson, P. (2005). Mining Text for Expert Witnesses, *IEEE Computer Society*, May-June, 94-100.

Fan, W., Wallace, L., Rich, S. et Zhang, Z. (2006). Tapping the power of text mining. *Association for Computing Machinery. Communications of the ACM*, 49(9), 76-82.

Fayyad, G., Piatetsky-Shapiro et P. Smyth (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54.

Foley, E., et Guillemette, M.G. (2010). What is Business Intelligence. *International Journal of Business Intelligence Research*, 1(4), 1-28

Gkoulalas-Divanis, A. et Verykios, V.S. (2009). Hiding Sensitive knowledge without side effets. *Knowledge and Information System*, 20, 263-299

Glymour, C., Madigan, D., Pregibon, D. et Smyth, P. (1997). Statistical themes and lessons for data mining. *Data mining and knowledge discovery*, 1(1), 11-28

Goh, T., et Huang, Y. (2009). Monitoring youth depression risk in Web 2.0. *VINE: The journal of information and knowledge management systems*, 39(3), 192-202.

Hair, J.F. Jr. (2007). Knowledge creation in marketing: the role of predictive analytics. *European Business Review*, 19(4), 303-314, ISSN 0955-534X

Herschel, R.T. et Jones, N.E. (2005). Knowledge management and business intelligence: the importance of integration. *Journal of knowledge management*, 9 (4), 45-55

Hevner, A.R (2007). A Three Cycle View of Design Science Research, *Scandinavian Journal of Information Systems*

Hevner, A.R., March, S.T. et Ram, S. (2004). Design Science In InformationSystems Research, *MIS Quartely*, 28 (1), 75-105

Hirschman, L, et Gaizauskas, R. (2001). Natural language question answering: the view from here. *Natural Language Engineering*, 7, 275-300

Hoffman, T. (1998). Banks Turn to IT to Reclaim Most Profitable Customers. *Computer World*, December 7

Hoffman, T. (1999). Insurers Mine for AgeAppropriate Offering. *Computer World*, April 19

Imhoff, C. (2004), Predicting the Future – A Crystal Ball. *DM Review*, August, 3 , ISSN 1521-2912

Jukic, N. et Nestorov, S. (2006). Comprehensive data warehouse exploration with qualified association-rule mining. *Decision Support Systems*, 42(2), 859

King, W.R. (2009). Text Analytics: Boon to Knowledge Management? *Information Systems Management*, 26(1), 87

Kloptchenko, A., Eklund, T., Kar;sson, J., Back, B. et Vanhar, H. (2004). Combining Data and Text Mining Techniques for Analysing Financial reports. *Intelligent Systems in Accounting, Finance and Management*, 12(1)

Laurent, W. (2008). Mining the Business Intelligence from Unstructured information. *DM Review*, 2008 (April).

Lee, S., Song, J., et Kim, Y. (2010). AN EMPIRICAL COMPARISON OF FOUR TEXT MINING METHODS. *Journal of Computer Information Systems*, 51(1), 1-10

Leong, E., Ewing, M., et Pitt, L. (2004). Analysing competitors' online persuasive themes with text mining. *Marketing Intelligence & Planning*, 22(2/3), 187-200

Liu, B., et Tuzhilin, A. (2008). MANAGING LARGE COLLECTIONS OF DATA MINING MODELS. *Association for Computing Machinery. Communications of the ACM*, 51(2), 85

Liu, D. R. et Shih, Y.Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management*, 42(3), 387

Lu, Y., Luo, X., Polgar, M. et Cap, Y. (2010). Social Network Analysis of a Criminal Hacker Community. *The Journal of Computer InformationSystems*, 51(2), 31

March, S.T. and G.F. Smith (1995). Design and natural science research on information technology. *Decision Support Systems*, 15, 251-266

Marcoux, Y. et Rizkallah, É. (2009). Intertextual semantics: A semantics for information design. *Journal of the American Society for Information Science and Technology*, 60(9), 1895-1906

Menon, S. et Sarkar, S. (2007). Minimizing Information Loss and Preserving Privacy. *Management Science*, 53(1), 101-116.

Menzies, T., Dekhtyar, A., Distefano, J. et Greenwald, J. (2007). Problems with Precision: A Response to Comments on 'Data Mining Static Code Attributes to Learn Defect Predictors'. *IEEE Transactions on Software Engineering*, 33(9), 637

Nemati, H.R. et C.D. Barko. (2001). Issues in Organizational Data Mining: A Survey of Current Practices. *Journal of Data Warehousing*, 6(1)

Ong, T. H., Chen, H., Sung, W.K. et Zhu, B. (2005). Newsmap: a knowledge map for online news. *Decision Support Systems*, 39(4), 583

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., Mertens, P., Oberweis, A., Sinz, E. J., et al. (2011). Memorandum on design-oriented information systems research. *European Journal of Information Systems*, 20 (1), 7-10

Owen, C. (1998). Data Modelling, Data Warehousing and Data Mining. Information Management, 1998 (November).

Payne, D. et Trumbach, C.C. (2009). Data mining: proprietary rights, people and proposals. *Business Ethics: A European Review,* 18(3), 241-252

Pechenizkiy, M, Puuronen, S. et Tsymbal, A. (2008). Towards more relevance-oriented data mining research. *Intelligent Data Analysis* 237-249

Pisetta, V.,Hacid, H., Bellal, F., Ritschard, G. et Zighed, D.A. (2006). Traitement automatique de textes juridiques, *Actes de SdC*

Ponmuthuramalingam, P., et Devi, T. (2010). Effective Term Based Text Clustering Algorithms. *International Journal on Computer Science & Engineering*, 2(5), 1665-1673

Pouponnot (2009). Tirer profit des solutions analytiques. *Relation Client Magazine*, 80(Avril)

Quatrain, Y., Nugier, S., Peradotto, A. et Garrouste, D. (2004). Évaluation d'outils de Text Mining : démarche et résultats. *7ièmes Journées Internationales d'Analyse Statistique des Données Textuelles.*

Radha, R. (2008). A Scoring and Choice Model for Multistage Cross-Selling in the Insurance Industry, Part 1. *Information Management Online*, August 13

Ruthledge, J. (2009). Top Ten Algorithms in Data Mining. *Journal of Quality Technology*, 41(4), 441

Schmid, H, (1994). Probabilistic Part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing,* 44-49.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), 13-22

Shi, X. et Yang, C.C. (2007). Mining related queries from Web search engine query logs using an improved association rule mining model. *Journal of the American Society for Information Science and Technology*, 58(12), 1871-1883

Sinha, A.P. and Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers : An application in indirect lending. *Decision Support Systems,* 46 (1), 287-299

Spiegler, I. (2003). Technology and knowledge: Bridging a "generating" gap". *Information & Management*, 40(6), 533

Spinakis, A. et Chatzimakri, A. (2005). Comparative Study of Text Mining Tools. *Studies in Fuzziness and Soft Computing*, 185, 223-232

Stodder, D. (2010). Text Analytics Drives Customer Insight. *InformationWeek*, 25 janvier, 35-38

Swanson, D. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science,* 38(4), 228–233

Tseng., Y., Lin., C., et Lin., Y. (2007). Text mining techniques for patent analysis. *Information Processing and Management*, 43, 1216-1247

Wang, H. et Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. *Industrial Management and Data Systems,* 108 (5) 622-634

Wei, C., Yang, C. et Lin, C. (2008). A Latent Semantic Indexing-based approach to multilingual document clustering. *Decision Support Systems,* 45 (3), 606-620

Wright, D., Friedewald, M., Schreurs, W., Verlinden, M., Gutwirth, S., Punie, Y., Maghiros, I., Vildjiounaite, E. & Alahuhta, P. (2008). THE ILLUSION OF SECURITY. *Association for Computing Machinery. Communications of the ACM,* 51(3), 57

Wu, H., Gordon, M, DeMaagd, K. et Fan, W. (2006). Mining web navigations for intelligence. *Decision Support Systems,* 41(3), 574-591.

Yang, Y.Y., Akers, L., Klose, T. et Barcelon Yang, C.S. (2008). Text mining and visualization tools – Impressions of emerging capabilities. *World Patent Information,* 30, 280–293

7.3 Autres documents

"Vie privée et confidentialité des données" Gouvernement du Canada Groupe consultatif interagences en éthique de la recherche

Rexer. K. (2010) 4th Annual Data Miner Survey. Tech. report, Rexer Analytics, suggéré par Jessica

KDnuggets (2002) What main methodology are you using for data mining?

http://www.kdnuggets.com/polls/2010/analytics-data-mining-industries-

applications.html , Septembre 2010 .