Un usage du Text Mining : donner du sens à la connaissance client

Manu CARRICANO[1] & Grégoire DE LASSENCE[2]

[1]Professeur, EADA Barcelona
[2]Responsable Pédagogie et Recherche, SAS Academic

RÉSUMÉ

La technologie du Data Mining a considérablement enrichi les outils traditionnels de traitement de la connaissance client en améliorant en particulier leur potentiel prédictif. Cet outil complexe a récemment connu de nombreux développements. Parmi ceux-ci, émerge la possibilité d'intégrer aux modèles traditionnels des données non structurées, représentant plus de 80 % de la connaissance disponible dans l'organisation. Le Text Mining permet d'exploiter ces données afin d'optimiser la prise de décision dans l'entreprise. L'objet du présent article est tout d'abord de présenter le Text Mining et son utilité en management, puis de démontrer sa valeur ajoutée, en d'autres termes, de quelle manière l'intégration de données textuelles aux modèles de Data Mining classiques améliore le potentiel prédictif de ces outils. Cette démonstration est faite en deux temps : en mettant en évidence tout d'abord son utilité dans la description et l'identification des données textuelles les plus saillantes, puis en mettant en compétition le modèle enrichi de données textuelles avec d'autres modèles prédictifs sur données structurées. Notre étude reprend un cas dans le secteur automobile, et montre de quelle manière, en combinant données structurées et textuelles, un constructeur peut être capable d'anticiper le rappel d'un véhicule, et par suite d'éviter les risques pour la marque liés à une mauvaise gestion de crise.

Mots-clés : CRM, Data Mining, Analyse de Données Textuelles, gestion de la connaissance

ABSTRACT

Data Mining technologies have enhanced management research's predictive capability. In recent years, many improvements have been made, among others by incorporating non-structured data to traditional models. This is an important challenge as non-structured data accounts for more than 80% of an organization's knowledge. Text Mining allows researchers to use this type of data to optimize decision making processes. The goal of this paper is to describe Text Mining implementation and its contribution to management, in other words, the way non-structured data's integration to traditional Data Mining models can optimize the predictive outcome of such analysis. The added-value of Text Mining is demonstrated as follows: first we show that Text Mining allows considerable enrichment of traditional data mining models through identification and analysis of the most relevant textual data; second, through showing that the model with textual data over performs other models with structured data only. We analyze a case in the automotive industry that illustrates how a manufacturer can anticipate vehicles recall by combining structured and non-structured data, and avoid consequently the risk for its brand due to a bad crisis management.

Key-words: CRM, Data Mining, Text Mining, knowledge management

INTRODUCTION

Les entreprises sont noyées dans l'information mais ont soif de connaissance : l'information non structurée, la plupart du temps sous forme de fichiers texte, représente en effet plus de 80 % de la connaissance présente dans l'organisation. Ces données, issues des nombreux points de contact entre l'entreprise et ses clients (e-mail, site web, call center, etc.) sont un enjeu central dans la gestion de la connaissance client mais sont difficiles à collecter, synthétiser et à analyser. De nouvelles technologies permettent désormais d'extraire des éléments utiles pour la décision de vastes volumes de données non structurées, de découvrir des relations entre des variables, et de synthétiser l'information disponible. Extension du Data Mining, le Text Mining – *ou fouille de données textuelles* – permet d'analyser automatiquement ces données textuelles sous-exploitées voire perdues pour l'entreprise, et par suite créer de la connaissance à partir des larges volumes de documents. Il sera particulièrement utile pour traiter par exemple des documents internes (rapports), des articles scientifiques ou des brevets, des commentaires issus de forums ou de call-centers, des questions ouvertes d'enquêtes, ou encore des e-mails. Le Text Mining utilise des méthodes des statistiques textuelles afin d'intégrer le texte non formaté *a priori*, aux puissants et éprouvés modèles de Data Mining.

Les apports du Text Mining en gestion peuvent être classés en deux dimensions principales : améliorer la connaissance du marché et anticiper et détecter des évolutions futures afin de piloter la stratégie de l'entreprise (voir tableau 1).

Le déploiement de ces outils représente un intérêt bien compris de la part de la communauté académique en systèmes d'information et du monde de l'entreprise. Mais les données textuelles étant par nature difficiles à traiter, la question de la pertinence de leur incorporation aux modèles de Data Mining doit être posée. Comme l'avaient fait il y a peu Padmanabhan *et al.* (2006) dans le domaine du *e-CRM*, il s'agit donc de chercher à établir les bénéfices liés à une modélisation issue de données complètes afin de prévenir une « myopie » de la décision. L'objectif de cet article est par conséquent d'illustrer comment le Text Mining améliore le potentiel prédictif des modèles traditionnels en donnant du sens aux données générées à partir d'applications de type CRM.

Après avoir présenté le Text Mining et ses applications en gestion, puis avoir défini les étapes de mise en œuvre d'un projet de Text Mining, nous montrerons par le biais d'un cas d'application de quelle manière l'utilisation de données textuelles permet d'enrichir les modèles prédictifs traditionnels et quelle peut être son utilité pour les gestionnaires. A cette fin, nous utilisons une base de données publique recensant des plaintes d'usagers du réseau autoroutier américain[1] et illustrons de quelle manière, à travers un cas réel, un chef de projet peut : 1) utiliser ce vaste corpus de données non-structurées pour décrire les accidents (cause technique, partie du véhicule endommagée, etc.) ainsi que les marques de véhicules mises en cause, et 2) utiliser ces données afin de prédire des comportements futurs (rappel de véhicule, etc.).

1. DATA MINING, DONNÉES TEXTUELLES ET TEXT MINING

Le Text Mining, extension du Data Mining est un outil de traitement de la connaissance client. Ces technologies relativement récentes (fin des années 1960) « permettent d'analyser et d'interpréter de façon intelligente et automatisée de grandes bases de données » (Piatetsky-Shapiro et

[1] La base de données de l'US National Highway Traffic Safety Administration (NHTSA).

Tableau 1 : Apport du Text Mining en gestion

Orientation	Problématiques de gestion	Sources d'information textuelles
Marché	Conception du produit	Forums d'opinion, espaces communautaires, emails ou commentaires sur sites commerciaux
	Communication positionnement	Idem
	Suivi des usages, retours produits	Idem
	Prix Prédire les évolutions du marché en fonction des dépêches	Idem
	Place : différence des évaluations selon les communautés, études des contagions	Idem + blogs
	Vente et Service après-vente Optimisation des centres d'appel, réduction du temps d'attente, amélioration du routage Filtrer, router les e-mail Segmentation de remarque client Analyse d'enquête	Emails [+ informations internes (BDD clients)]
Stratégie	Stratégie : • Benchmarking • Anticipation des tendances • Intelligence économique • Evolution des attentes • Prédiction de la satisfaction client	Sites de la concurrence + Blogs + forums et espaces ouverts

Source : adapté de Gauzente (2006)

Frawley 1991). Le développement de l'accès à ces nouvelles technologies a progressivement permis aux praticiens d'utiliser une large palette d'alternatives aux modèles traditionnels (statistiques) d'analyse de la connaissance client. Ces alternatives reposent sur de nombreuses techniques issues de l'apprentissage artificiel (*machine learning*), de la reconnaissance des formes (*pattern recognition*) et des réseaux de neurones. Un des intérêts majeurs de ces nouvelles approches est qu'elles permettent d'outrepasser certaines conditions limitant les approches traditionnelles, rendant ces outils relativement robustes et utiles pour la décision (Kiang et Kumar 2007).

L'autre enjeu – au-delà des progrès techniques réalisés – réside dans l'amélioration de l'interprétation, de l'accès au sens. Les larges volumes à traiter constituent un véritable challenge pour les professionnels en systèmes d'information, et de nombreux chercheurs se sont attelés à la tâche de donner du sens à ces données collectées et analysées par les méthodes de Data Mining afin d'améliorer la performance de l'entreprise (Cooper *et al.* 2000, Speier et Morris 2003).

L'incorporation de données textuelles aux modèles de Data Mining représente bien un enjeu important afin de donner du sens aux données collectées et d'améliorer la performance de l'entreprise. Nous présentons ici, après l'avoir défini, la mise en œuvre d'un projet de Text Mining.

1.1. Le Text Mining

1.1.1. Définition et objet

Depuis longtemps déjà, le langage et les formes de communications sont au centre de nombreux courants de recherche. L'exploitation de ces données est un phénomène plus récent, au cœur d'intérêts bien compris d'entreprises qui cherchent à optimiser leurs techniques de relation clients multicanal. Ces artefacts de communications informatisées, au cœur de nos échanges modernes sont difficiles à exploiter, à analyser, du fait de leur forme non structurée. Les outils de Text Mining, en reposant sur l'analyse de données textuelles, ont permis de réaliser un pas important, de la gestion de la connaissance à une analyse approfondie de cette connaissance présente dans l'entreprise.

Chez les analystes l'ambition de départ de l'analyse des données textuelles (ADT) était de nature essentiellement linguistique : tester le bien fondé ou apporter des éléments aux analyses des grammairiens (Benzécri face à Chomsky p. ex.) par une analyse quantitative des corpus observés. Dans ce domaine, statistique textuelle et statistique lexicale doivent être distinguées, dans le sens où la première ne prend pas forcément le mot comme unité d'analyse. L'analyse des données textuelles repose sur une unité d'analyse (le corpus, les mots, les groupes de mots, les répondants, etc.), des variables – qui ne sont pas définies *a priori* mais induites du texte (mots, lemmes, séquences, occurrences), ainsi que sur des méthodes d'analyse.

A la différence de l'analyse des données textuelles qui part du postulat que l'organisation interne des éléments d'un discours, d'un texte, « mémorise » par sa forme même des processus externes qui ont conduit à sa production (Reinert, 1993), le Text Mining a pour objet l'extraction de connaissance utile pour la décision à partir d'un large volume de données textuelles non-structurées. Son objet est donc moins l'analyse de textes que la modélisation et l'analyse prédictive.

Le Text Mining se définit par « *l'ensemble des techniques et méthodes destinées au traitement automatique de données textuelles en langage naturel disponibles sous forme informatique, en grande quantité, en vue d'en dégager et structurer le contenu, les thèmes dans une perspective d'analyse rapide (non littéraire), de découverte d'informations cachées, ou de prise automatique de décision.* » (Tufféry 2005, p. 323). Le Text Mining est une extension du Data Mining développée autour de la lexicométrie ou statistique lexicale (Benzécri 1981), et nourrie des récents et importants développements permettant de traiter le matériau textuel. A l'instar du Data Mining, le Text Mining et ses applications peuvent être classées en deux approches : explorer les données textuelles et leur contenu – le Text Mining descriptif –, et/ou utiliser cette information pour optimiser la décision et les processus de l'organisation – le Text Mining prédictif –.

Exemples d'application d'outils de Text Mining

Le Text Mining descriptif cherche à découvrir les thèmes et concepts présents dans les données textuelles. Par exemple, les entreprises cherchant à mieux connaître les préférences de leurs clients peuvent utiliser cette approche afin de tirer parti de commentaires et d'informations collectés via leur

site web, les e-mails ou les centres d'appel. L'explorer ces commentaires reviendra à identifier des termes, des phrases, ou autre unité d'analyse, classer les documents en groupes homogènes, ou encore explorer les concepts mis en évidence dans ces groupes de documents. Ces résultats permettent à l'analyste d'avoir une meilleure vue d'ensemble, de mieux comprendre les données collectées.

Le Text Mining prédictif consiste à classer les documents en différentes catégories, et à utiliser l'information implicite dans ces données textuelles afin d'améliorer la prise de décision, comme par exemple identifier des questions récurrentes de consommateurs afin d'automatiser les réponses, prédire la probabilité de rachat, ou autre problématique de modélisation prédictive proche du Data Mining traditionnel. En autorisant l'exploitation de cette information « cachée », le Text Mining prédictif combine la robustesse des modèles de Data Mining et la richesse des données non-structurées : en d'autres termes, une analyse prédictive fondée sur les attitudes, les comportements directement observés, non altérés par le contexte ou par l'instrument de recherche, ou d'une manière plus générale, l'optimisation de la connaissance issue des interactions entre l'entreprise et son environnement.

1.1.2. Text Mining et analyse de données textuelles

Comme cela a été dit, le Text Mining est une extension du Data Mining fondée sur les avancées réalisées en lexicométrie. Le Text Mining utilise donc les outils robustes du Data Mining afin d'aider à découvrir le sens cachés dans de grands corpus textuels, de les catégoriser, de mettre en relation des éléments lexicaux entre eux, de faire émerger des tendances, etc. Le Text Mining est en ce sens une forme d'analyse informatisée de données textuelles qui permet la découverte d'énoncés divers dont la connaissance rapide est indispensable ; son objet est une analyse systématique des contenus, et peut donc être assimilé à une production de données quantifiées destinées à tester un modèle. En d'autres termes, un projet de Text Mining vise à transformer des données textuelles en métadonnées pouvant être analysées par des modèles de data mining traditionnels. Mais quelques mises en gardes s'imposent : « *l'informatique représente un outil précieux (…), et la réalisation de statistiques lexicales peut largement favoriser la compréhension de données textuelles, à condition de connaître précisément quels support et cadres l'analyse lexicale est en mesure d'offrir à un processus ultérieur d'interprétation* » (Gavard-Perret et Moscarola 1998, p. 37).

Il est important de bien saisir la complexité des données textuelles avant d'entreprendre un projet de Text Mining, et de choisir la solution idéale. En effet, certaines spécificités des données textuelles ont une incidence directe sur les méthodes de classification et de modélisation, et plus généralement, sur les possibilités d'analyse du corpus (Lebart 1998, p. 474).

Les concepts de variable et d'observation sont ainsi plus complexes en Text Mining qu'en statistiques traditionnelles. Les variables par exemple, au lieu d'être définies *a priori*, sont directement inférées du corpus. Les variables en Text Mining peuvent être les unités textuelles suivantes : mots, lemmes (mots ramenés à leur forme canonique), ou segments (séquences de mots apparaissant fréquemment).

Les observations, ou unités statistiques, sont en général des documents (décrits par leur titre ou leur résumé) au sein de bases de données documentaires, des répondants (décrits par leur réponses à des formulaires ou à des questions ouvertes), des segments des textes (phrases, unités de contexte, paragraphes). Les occurrences de mots représentent un second niveau d'unité statistique qui peut également être analysé. En

effet, des tests statistiques peuvent reposer sur les fréquences de mots, de documents, de séquences, de répondants. Ces différents niveaux sont souvent source d'erreurs d'interprétation auquel l'analyste doit prêter attention.

Les problèmes spécifiques liés à l'analyse des données textuelles (syntaxiques, structurels, lexicaux) doivent également être pris en compte, ainsi que le résume le tableau 2.

La taille des bases de données et donc le volume de texte à analyser (des milliers de documents, de mots), ainsi que l'hétérogénéité des données augmentent considérablement la complexité des tableaux lexicaux à traiter. Cependant, la principale difficulté réside dans l'énorme quantité de métadonnées à prendre en compte. En effet, chaque mot peut être comptabilisé plusieurs fois dans un même dictionnaire et de puissants outils linguistiques sont nécessaires afin d'identifier sans ambiguïté le lemme associé à un mot. Ainsi, les règles de grammaire ou de sémantique (en fonction de différentes langues) constituent la base de cette méta-information.

1.2. Les étapes d'un projet de Text Mining

Le Text Mining traite des documents non structurés, écrits en langage naturel. Il est donc primordial de définir un projet d'analyse afin de structurer les données à modéliser (Balbi et Di Meglio, 2004). Nous retiendrons, pour notre part, quatre étapes principales, correspondants aux étapes d'un projet de Data Mining, les spécificités du Text Mining étant traitées dans la troisième étape, la phase de préparation du corpus. Nous précisons entre parenthèses le temps passé à chaque étape, en pourcentage du temps total alloué au projet.

1) Création de la base de données (40 % du projet)

La préparation des données est le préalable à tout travail sur le corpus. Il s'agira 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, de stocker les documents dans la base de stockage. Cette base devra être structurée de telle sorte qu'il y ait une ligne par individu (plaintes,

Tableau 2 : Problèmes spécifiques liés à l'analyse de données textuelles

Catégorie	Type	Exemples
Syntaxique	Coordinations	de, pour, le, sur, qui, quoi, car…
	Ponctuations	!, ?, :, «
	Caractères spéciaux	$, @, #, *, &
Structurelle	Technique	extensions de fichiers, police, …
Lexicale	Valence	Positif / Négatif
	Affect	Bonheur, colère, haine, etc.
	Idiosyncrasie	Expressions idyomatiques, langage vernaculaire
	Géographie	Lieux, adresses, etc.
	Temps	Dates, années, etc.

Source : adapté de Abbasi et Chen (2008)

messages, réclamations, documents etc.) ; et en colonne, le texte complet du document, soit des cellules de milliers de caractères : il n'y a pas de limite. La table peut comporter d'autres colonnes, généralement plus structurées, telles que l'auteur, la date de réception, le destinataire, etc. Une table d'entrée de processus de Text Mining diffère d'une table de Data Mining du fait qu'il y existe au moins une colonne comportant une information textuelle non structurée. Notons qu'il existe des processus automatiques permettant d'extraire du site web l'ensemble textuel s'y trouvant, ce qui peut faciliter le travail de l'analyste.

2) Echantillonnage (10 % du projet) :

Comme dans tout projet de Data Mining prédictif, les données sont séparées en trois échantillons : un échantillon d'apprentissage sur lequel seront calculés les coefficients du modèle, un échantillon de validation grâce auquel le modèle est optimisé, et un échantillon de test et qui permet de juger de sa qualité sur des données différentes de celle sur lesquelles il a appris. En effet, une des caractéristiques du Data Mining prédictif, est de pouvoir être appliqué sur de très grandes bases de données et de générer rapidement un modèle qui, ayant appris sur le passé, peut prédire l'avenir.

Les méthodes utilisées nécessitent d'avoir plusieurs milliers d'enregistrement et c'est notamment en testant les modèles générés sur une base de données différente de celle ayant été utilisée pour leur construction que l'on pourra en évaluer la performance et la robustesse.

3) Préparation du corpus (30 % du projet)

La phase d'échantillonnage permet de préparer une base d'apprentissage et une base de test. L'analyste doit ensuite traiter la (les) colonne(s) comportant un texte afin de la (les) rendre exploitable(s) pas des algorithmes de Data Mining.

Analyse linguistique

Après avoir identifié la langue, et pris en compte un éventuel multilinguisme du document, l'analyste doit procéder à la normalisation (éliminer les formatages ou caractères qui pourraient fausser l'analyse) et à la lemmatisation du corpus. La lemmatisation consiste à regrouper les formes graphiques correspondant à un même mot, épurer le vocabulaire des mots-outils non informatifs (articles…). Cette étape ne doit pas être réalisée trop rapidement auquel cas certains mots-outils caractéristiques d'attitudes ou d'opinions pourraient être omis, ou au contraire, des formes graphiques différentes d'un même mot pourraient bien être assimilées. Ces mots ainsi ramenés à leur forme canonique constituent le dictionnaire général du projet.

Analyse syntaxique (parsing)

L'analyse syntaxique, ou *text parsing*, consiste à associer automatiquement au corpus découpé en unités une représentation des groupements structurels et/ou des relations fonctionnelles existant entre ces unités. Ces traitements syntaxiques ne sont pas un but en soi mais sont voués à transformer les données textuelles en vue d'une analyse avec les techniques traditionnelles du Data Mining. Le *parsing* décompose le corpus textuel en une vaste matrice de fréquences de concepts par documents. Les concepts sont définis au sein d'une liste identifiant les mots à retenir, à conserver, et à relier. Ce processus itératif est une étape majeure du projet de Text Mining, et constitue le fondement des analyses statistiques ultérieures. Un corpus pouvant être composé de plusieurs centaines de documents et de milliers de mots, la matrice doit être transformée à son tour afin d'en réduire les dimensions.

Transformation (réduction des dimensions)

La réduction des dimensions consiste à créer une matrice exploitable à des fins de Data Mining. Deux approches sont possibles : la méthode des *N* termes et la *SVD*, ou décomposition en valeurs singulières (Garrouste et Lebourgeois 2002). La méthode *N* termes consiste à retenir les termes pondéreux, permettant de simplifier le document en le réduisant aux *n* termes les plus représentatifs. La *SVD* (décomposition en valeurs singulières) permet quant à elle de résumer la matrice grâce à une analyse factorielle, et de décrire le document en le ramenant à *n* variables principales. Ces variables peuvent être utilisées pour effectuer des typologies de documents et établir des modèles de classification automatique.

4) Modélisation (20 % du temps du projet)

L'objectif du Data Mining est de construire rapidement des modèles sur de grandes bases de données. Le fait de travailler sur de grandes bases permet d'outrepasser certain tests, notamment de normalité. Les modèles utilisés en Data Mining vont de la régression aux arbres de décision en passant par les réseaux neuronaux, le raisonnement à base de cas ou les *SVM* (séparateur à vaste marge ou *Support Vector Machine*). 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 ».

La construction d'un modèle revient à trouver la fonction $f(x) = y$, où x est l'ensemble des variables explicatives d'entrée et y est la (ou les) variable(s) de sortie à expliquer. Parmi les variables explicatives de Text Mining, il faudra donc traiter en plus des variables quantitatives et qualitatives du Data Mining classique, les variables construites par l'analyse de Text Mining.

Analyse, comparaison et intégration du modèle

Suite à l'étape de construction de la base de données de Text Mining, puis à l'échantillonnage et enfin aux différentes modélisations, il faut enfin comparer les modèles candidats pour en sélectionner le plus performant. 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. On s'intéressera donc prioritairement au taux de mauvais classement sur la table de test, et aux courbes de réponse (ou courbes de Lift) montrant la contribution du modèle, ou plus précisément le coefficient multiplicateur du taux de réponse apporté par le modèle par rapport à une approche aléatoire.

Une fois le modèle « champion » sélectionné, il pourra être intégré à un système d'analyse, d'ETL (collecte et préparation automatisée des données), ou bien de gestion des alertes dans un second temps.

2. DÉVELOPPER LE POTENTIEL PRÉDICTIF DES OUTILS DE TRAITEMENT DE LA CONNAISSANCE CLIENT : APPLICATION À UN CAS DE RAPPEL DE VEHICULES

En 2000, plus de 217 millions de véhicules ont été recensés aux Etats-Unis, parcourant plus de 2700 milliards de miles (soit plus de 4 300 milliards de kilomètres). Dans le même temps, les défauts de fabrication ont eu un coût global estimé à 12 milliards de dollars de dommages et intérêts pour les constructeurs automobiles, sans compter le ternissement de l'image de marque et le coût des poursuites judiciaires. Les modèles de Text Mining prédictif peuvent ainsi

permettre de déclencher préventivement des alertes pour rappeler les véhicules défectueux et donc augmenter la sécurité et la satisfaction des conducteurs, ainsi que l'image de marque de l'entreprise.

2.1. Méthodologie

1) Présentation de la base de données

Dans ce cadre d'application nous avons réalisé une analyse portant sur un échantillon de véhicules de tourisme et dont l'objectif de prédire le risque d'accident d'un véhicule. Nous utilisons pour cela la base de données de l'US National Highway Traffic Safety Administration (NHTSA). Cette base comprend des plaintes de particuliers, ayant eu un problème avec une partie d'un véhicule, ayant ou non engendré un accident, et peut être librement téléchargée depuis le site de la NHTSA[7]. Un constructeur peut donc utiliser ces données afin d'identifier les modèles et marques mis en cause, et développer un modèle afin d'anticiper un rappel de véhicule, et atténuer de la sorte les effets sur la marque.

La table de données utilisée dans notre étude contient 50 601 observations, décrites par 52 variables structurées et non structurées (textuelles) dont une variable dépendante binaire (*accident*, codée oui / non) indiquant si le véhicule a été ou non impliqué dans un accident. Plus précisément, la base contient 39 528 véhicules n'ayant pas été impliqués dans un accident, et 17 073 véhicules ayant été accidentés (soit 33,1 % des véhicules).

Après épuration, nous retenons finalement 37 variables explicatives, parmi lesquelles par exemple *kilométrage* (variable continue) ou *plainte* (variable textuelle contenant les plaintes décrivant le comportement du véhicule). Le tableau 3 présente un extrait de cette base de données.

L'analyse de Text Mining a été réalisée avec SAS Text Miner, un composant complémentaire de SAS® Enterprise Miner. Cet outil, par rapport à d'autres solutions potentielles, possède un grand nombre de méthodes de modélisation, et est très efficace pour une modélisation de Data Mining ultérieure, objectif de notre recherche (voir Quatrain *et al.* 2004 pour une approche comparative des outils de Text Mining). Plus précisément, SAS Text Miner permet de transformer une table de Text Mining brute comprenant notamment une colonne où dans chaque cellule est enregistré un texte entier (article, libre, commentaire, etc.) en une table exploitable par les outils de Data Mining classique. En outre, cet outil est particulièrement adapté pour explorer des données non-structurées (Davi *et al.* 2005) : alors que la plupart des logiciels se fondent sur une analyse lexicale (reconnaissance des termes, création de dictionnaire, etc.), SAS Text Miner génère des typologies d'association entre les variables de la table, mettant ainsi en évidence les liens (et leur intensité) entre les données textuelles (voir section 2.2. ci-après). Cette caractéristique représente un avantage certain lorsque l'analyste travaille sur de vastes bases de données (comme c'est le cas ici), mais une limite quand au traitement du contenu textuel, qui doit être menée de manière complémentaire.

2) Modèle et échantillonnage

Le modèle de Text Mining (figure 1) que nous mettons en œuvre en fonction des étapes précédemment abordées, permet d'échantillonner la table de données en trois (Data Partition). Nous retenons les critères de partition conventionnels en Data Mining :

[7] ftp://ftp.nhtsa.dot.gov/Consumer Complaints/

Tableau 3 : Extrait de la BDD NHTSA

	MODELTXT	CRASH	FAIL_DATE	FIRE	CSUMMARY	INJURIES	DEATHS	CO
1	TOWN AND COUNTRY	N	97-05-22	Y	TRANSMISSION COOLING UNIT AND LINES LEAK CAUSING VEHICLE FIRE AK	0	0	POWER
2	BLAZER	N	97-07-02	N	BRAKES THE BRAKES WERE INOPERATIVE TOOK VEHICLE TOOK DEALER AND WAS TOLD THAT THE RETURN SPRINGS FOR THE S	0	0	BRAKES
3	S10	N		N	WINDSHIELD WIPERS ARE NOT WORKING AT ALL AK	0	0	VISUAL
4	BLAZER	N	97-07-02	N	BRAKES THE BRAKES WERE INOPERATIVE TOOK VEHICLE TOOK DEALER AND WAS TOLD THAT THE RETURN SPRINGS FOR THE S	0	0	BRAKES
5	VOYAGER	N	96-12-19	N	TRANSMISSION SHUDDERS WHEN STARTED UP FROM A STOPPED POSITION AFTER THE VEHICLE HAS BEEN WARMED UP SOMETI	0	0	POWER
6	VOYAGER	N	96-12-19	N	TRANSMISSION SHUDDERS WHEN STARTED UP FROM A STOPPED POSITION AFTER THE CAR HAS WARMED UP AT TIMES IT ALSO	0	0	POWER
7	PICKUP	Y	95-09-24	N	FRONT SUSPENSION BROKE LOST CONTROL OF VEHICLE THEN IT FLIPPED OVER AK CONSUMER WAS INJURED IN ACCIDENT	1	0	SUSPEN
8	CARAVAN	Y	91-09-25	N	BRAKE FAILURE CAUSED ACCIDENT LW	1	0	BRAKES
9	K35	Y	94-04-15	N	BRAKES FAILED WITHOUT WARNING CAUSED ACCIDENT REPORTED BY TROOPER BONNELL OWNER WILL PROVIDE ADDITIONAL I	1	0	BRAKES
10	S10	Y	91-02-18	N	LEFT FRONT UNIROYAL TIRE SEPARATED FROM RIM VEHICLE SWERVED HIT A UTILITY POLE RESULTED IN NO APPARENT TIRE DA	0	0	TIRES
11	CARAVAN	Y	95-03-11		MR PHILIPS WITNESSED INCIDENCE OF EJECTION OF A CHILD WHEN VEHICLE WAS HIT BROADSIDE THEN REAR LIFTGATE FLEW O	1	1	STRUC
12	AEROSTAR	Y	94-06-30	N	ABS APPLIED BRAKES EXPERIENCED EXTENED PEDAL FADED IMPACT 12:00 POSITION SPEED UNKNOWN AIR BAG DID NOT DEPL	1	0	BRAKES
13	S10	Y	94-04-15	N	PCV VALVE BLEW OFF HOSE CAUSING HIGH VEHICLE IDLE APPROX 3000 RPMS TT	1	0	FUEL F
14	VANAWAGON	Y	91-07-08	N	HILE DRIVING VAN PULLED LEFT, ENDED UP IN CENTER DIVIDER and FLIPPED OVER LW	1	0	SUSPEN
15	VANAWAGON	Y	91-07-08	N	WHILE MAKING RIGHT TURN, REAR END OF VAN LOST TRACTION, WHEELS REACHED EDGE OF ROAD, and VAN ROLLED OVER LW	1	0	POWER
16	LUMINA APV	Y	94-03-06	N	REAR SEAT BELTS DID NOT LOCK UP UPON IMPACT EXTENDED OUT TO EXTREME LEAVING OCCUPANTS UNRESTRAINED TT	3	0	INTERI
17	CARAVAN	Y	94-11-04	N	UPON IMPACT AT 35MPH 2ND ROW BUCKET SEAT BELT ANCHOR DETACHED FROM FLOOR PASSENGER WAS THROWN OUT OF SEA	1	0	INTERI
18	DAKOTA	Y	89-08-08	N	TIE ROD BROKE RESULTING IN ACCIDENT TRUCK ROLLED OVER TWICE BI	0	0	STEERI
19	MPV	Y	91-01-31	N	WHILE PARKED IN DRIVEWAY VAN DRIFTED BACKWARDS HITTING FENCE SLIPPED TRANSMISSION SYSTEM	1	0	POWER
20	VOYAGER	Y	92-01-18	N	VEHICLE CROSSED CENTERLINE CAUSING ACCIDENT LDG	3	1	STEERI
21	VOYAGER	Y	93-09-21	N	FRONTAL COLLISION 11:00 POSITION VEHICLE WAS AT COMPLETE STOP AIR BAG DID NOT DEPLOY LAP / SHOULDER BELT FAILE	2	0	INTERI

1. Une table d'apprentissage pour construire les modèles (40 %).
2. Une table de validation pour optimiser les modèles (30 %).
3. Une table de test pour tester les modèles sur des données différentes de celle sur lesquelles ils ont été construits (30 %).

Suite à cet échantillonnage, l'outil Text Miner est utilisé pour préparer les données textuelles. Une des étapes clés consiste à mettre en compétition différents modèles afin d'évaluer leur capacité prédictive. Sept modèles sont comparés dans ce projet :

- Un réseau de neurone
- Un algorithme de raisonnement à base de cas (*Memory-Based Reasoning*)
- Deux régressions robustes (*Dmine Regression*)
  1. Une régression sans les données textuelles
  2. Une régression avec les données textuelles
- Trois arbres de décision identiques :
  1. Un arbre sans les données textuelles
  2. Un arbre avec uniquement les données textuelles
  3. Un arbre de décision avec les données textuelles et les données de Data Mining classique.

3) Préparation du corpus

La phase la plus longue dans ce type de projet est celle du paramétrage du processus de traitement Text Mining en lui-même. Une première étape consiste à vérifier la table de données. La langue et le format des documents doivent être contrôlés à ce stade, même si les langues et formats pris en charge sont importants. La table de données contient soit le texte à traiter, soit un lien vers ce texte. Le texte est réduit (lemmatisation) afin de ne conserver que l'information la plus riche et

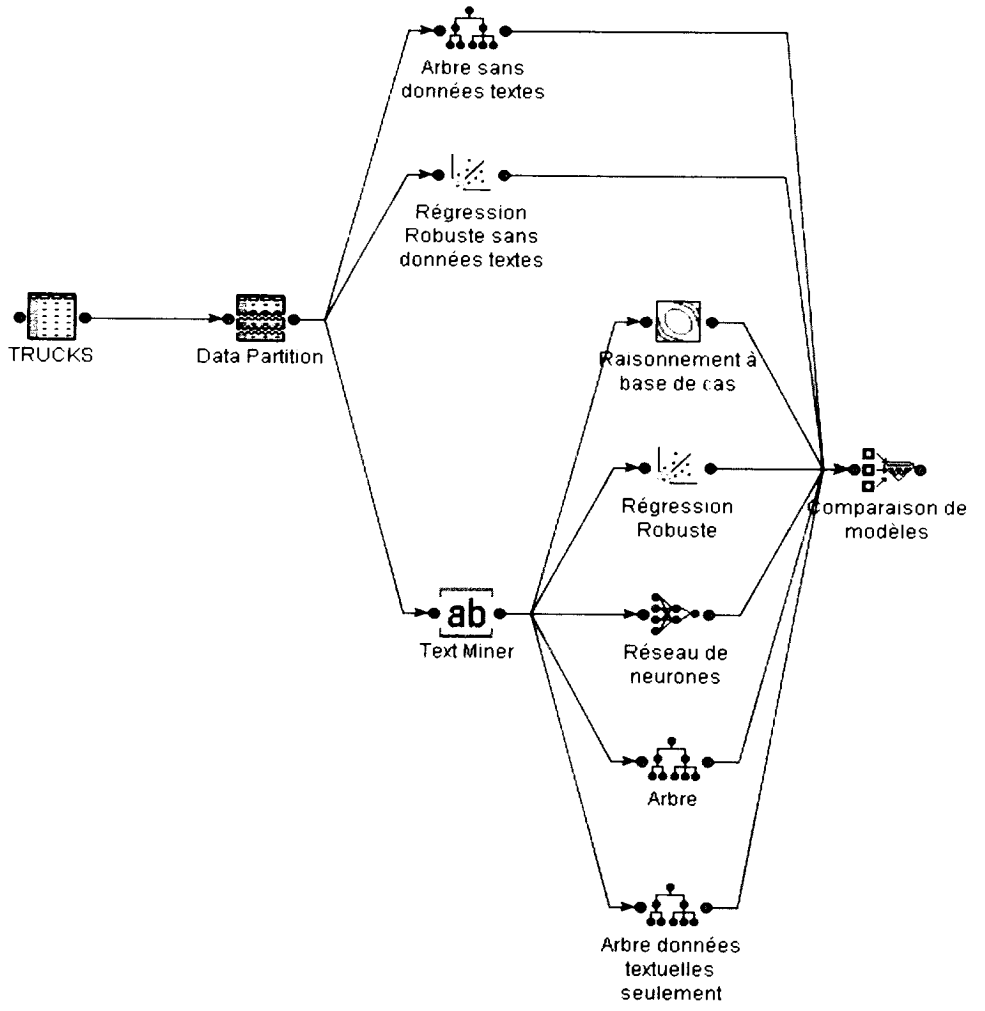

Figure 1. Représentation du modèle de Text Mining.

écarter les mots-outils par exemple. L'analyse syntaxique (*parsing*) transforme ensuite la table en une matrice de présence/absence de concepts par texte et/ou document.

Après avoir exécuté Text Miner une première fois, les entités textuelles « *'s, *ak, ak, tt vehicle, vehicles et have* » ont été rajoutées à la stop liste, liste des mots que l'on exclut de l'analyse[3].

[3] Text Miner propose une « stop liste », une « start liste », liste des mots à absolument utiliser, et une « synonyme liste », ces dictionnaires devant être complétés et adaptés au contexte spécifique de l'analyse.

En comparant sur ce modèle deux régressions robustes identiques, l'une pour laquelle le processus de Text Mining a utilisé la stop-liste de Text Miner par défaut, et l'autre où la stop liste a été complétée par les éléments ci-dessus, nous obtenons un taux d'erreur pour la première régression de 8,1 %, de 8,0 % pour la seconde, dont la stop liste a été modifiée. Le gain en terme de taux d'erreur des modèles est modeste, mais peut être assez significatif en coût monétaire.

La section suivante présente les résultats de l'étude en suivant les étapes du projet

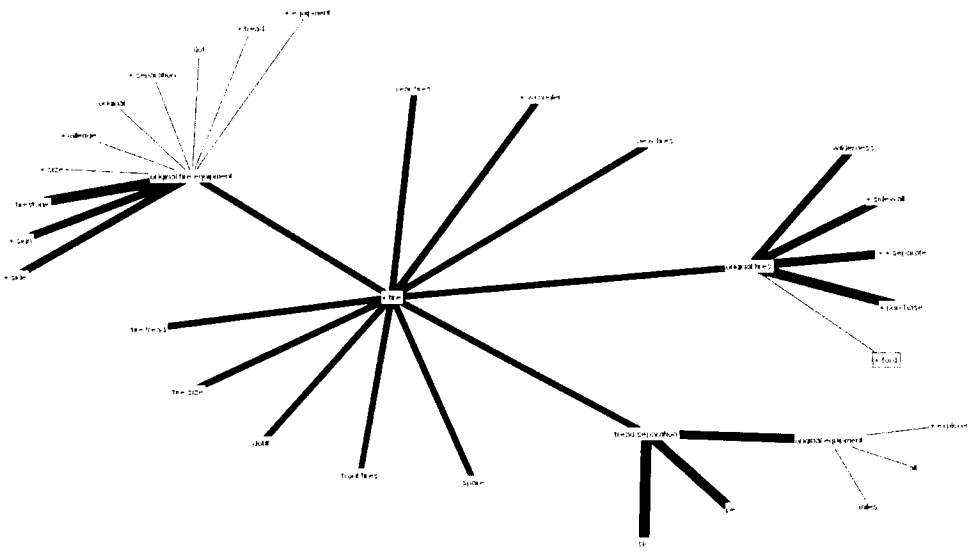

Figure 2. Exemple de représentation des données textuelles.

de Text Mining telles quelles ont été présentées dans la première partie. Les deux dernières parties du projet sont traitées : l'analyse descriptive et la modélisation.

2.2. Résultats

1) Analyse descriptive

Des analyses de contextes sémantiques sont menées de manière itérative afin d'affiner le dictionnaire et de minimiser le taux d'erreur. Ces analyses sont fondées sur la force de l'association entre deux concepts (distributions binomiales), fonction logarithmique d'une probabilité conditionnelle calculée de la manière suivante :

$$Strength = log_{20}(1/Prob_k)$$

$$avec\ Prob_k = C(n,r)\, p^r (1-p)^{n-r}$$

Où :

a : nombre de documents contenants le terme a

r : nombre de documents dans la base

n : nombre de documents contenant le terme b

k : nombre de documents contenant le terme a et le terme b

r : le nombre de documents à tester

$p = k/n$ est la probabilité conditionnelle que le terme a soit présent lorsque le terme b est présent, les deux étant indépendants

Nous remarquons par exemple[4] sur cette extraction que la marque Firestone se trouve dans le même périmètre lexical que les occurrences « *Ford* » et « *équipement d'origine* », elle-même proche du mot « *pneu* ». L'analyse montre effectivement des problèmes significatifs avec les pneus *Firestone* sur les véhicules de la marque *Ford* type *Explorer* (équipés d'origine par les pneus incriminés). Il est intéressant de noter que ce problème avait fait à l'époque l'objet d'une importante couverture médiatique (Kumar 2001).

[4] Pour des contraintes évidentes de volume, l'ensemble des extractions et analyses ne peuvent être présentés. Nous présentons ici des extraits afin d'éclairer notre propos, ainsi que les principaux résultats.

Tableau 4 : Problèmes spécifiques liés à l'analyse de données te: tuelles

Modèle	Taux d'erreur
DmineReg2	0,229048
Tree	0,187172
MBR	0,180812
Tree3	0,12592
Tree2	0,113847
Neural	0,101243
DmineReg	0,083397

2) Modélisation

Si l'on s'intéresse maintenant à l'estimation du potentiel prédictif des modèles, grâce aux résultats de la comparaison de modèle (Tableau 2), on peut noter que l'arbre n'utilisant que les données structurées (*tree*) a un taux d'erreur sur la table de test de 18,7 %, celui n'utilisant que des données textuelles (*tree3*), a un taux d'erreur sur la table de test de 12,5 % et celui utilisant toutes les données, quantitatives, qualitatives et textuelles (*tree2*) a un taux d'erreur sur la table de test de seulement 11,3 %. Ceci permet donc, sur un même modèle, de représenter l'apport du Text Mining dans la modélisation.

De plus, si l'on compare l'ensemble des modèles mis en compétition, la régression robuste utilisant les données textuelles (DmineReg) est le modèle vainqueur avec un taux de mauvais classement sur la table de test de seulement 8,0 % contre 22 % pour la régression robuste sans données textuelles (DmineReg2), devant le réseau de neurone à 10,1 %.

3. CONCLUSION

Le Massachusetts Institute of Technology (MIT) classait le Data Mining il y a peu comme l'une de dix technologies émergentes qui «*changeront le monde au XXI siècle*». Il paraissait à ce titre important d'exploiter ces outils et leurs extensions les plus récentes, en particulier le Text Mining, et de montrer de quelle manière ils permettent d'enrichir les modèles prédictifs classiques largement utilisés en management afin d'utiliser la masse d'information disponible dans l'entreprise. Au-delà de la dimension prédictive communément associée au Data Mining, il semble que le Text Mining permette aujourd'hui à la fois de valoriser la connaissance client, et plus encore, de lui donner du sens en exploitant les données textuelles.

En outre, de nombreux chercheurs travaillent depuis peu sur les problèmes nouveaux suscités par l'explosion de l'accès aux données, via le Web par exemple. Les récents développements autour du web sémantique (à partir de la « vision » de Berners-Lee, Hendler et Lassila 2001) et qui consiste à faire un saut quantique de l'hypertexte à une représentation automatisée des connaissances, représentent un champ d'application fertile pour des applications de type Text Mining, afin de fournir par exemple des outils de qualification de connaissances issues du web pour différents types d'utilisateurs.

L'analyse empirique que nous avons menée est une illustration, parmi d'autres possibles, du potentiel de cette méthodologie innovante pour les professionnels en systèmes d'information. Les résultats de notre étude montrent en effet que le modèle combinant données quantitatives, qualitatives et textuelles permet de minimiser le taux d'erreur. Les données non-structurées permettent donc d'enrichir de manière significative les modèles de Data Mining. Au-delà de ces résultats, nous avons vu qu'un modèle de Text Mining permet également d'enrichir considérablement l'interprétation des modèles de Data Mining conventionnels.

Les implications managériales sont importantes. L'intégration des résultats de

cette analyse de Text Mining dans une plateforme décisionnelle permettront par exemple de créer facilement un processus itératif qui lira automatiquement et quotidiennement les nouvelles saisies sur le site de la NHTSA, appliquera le modèle précédemment créé – la régression logistique utilisant le traitement de Text Mining –, et en cas de dépassement d'un seuil critique, enverra une alerte spécifiant que les véhicules, en fonction de caractéristiques retenues dans le modèle, maximisent le taux de risque d'avoir un accident.

Malgré ces résultats encourageants, quelques limites doivent être abordées. L'objet même du Text Mining, qui relève de l'analyse automatisée de données non-structurées et rapproche la méthode d'outils de statistique multidimensionnelle, en limite la portée linguistique. Ainsi certains documents ne peuvent être analysés par de tels outils : écriture manuscrite de type chèques p.ex., textes littéraires manipulant les sous-entendus, l'ironie, etc. Même si nous avons montré que le Data Mining permet de modéliser de façon enrichie des comportements et que l'intégration d'un traitement de Text Mining permet d'améliorer significativement la performance de ces modèles, il semble tout de même qu'une étude plus approfondie du corpus de mots – en mobilisant des logiciels d'analyse textuelle (Nvivo, Alceste, Tropes) par exemple –, permette d'améliorer significativement cette performance en enrichissant notamment les dictionnaires. L'analyste pourra donc à profit combiner statistique textuelle et Text Mining afin de donner du sens aux données clients.

4. BIBLIOGRAPHIE

Abbasi A. et Chen H. (2008), Cybergate : a design framework and system for text analysis of computer-mediated communication, *MIS Quarterly*, 32, 4, 811-37.

Adam J-M. (2006), Autour du concept de *texte*. Pour un dialogue des disciplines de l'analyse des données textuelles, *Journées internationales d'Analyse des Données Textuelles (JADT)*, Besançon.

Balbi S. et Di Meglio E. (2004), A Text Mining strategy based on local contexts of words, *Journées internationales d'Analyse des Données Textuelles (JADT)*, Louvain-La-Neuve.

Benzécri J.P. (1981), *Pratique de l'analyse de données, linguistique et lexicologie*, Dunod, Paris.

Cooper B.L., Watson H.J., Wixom B.H. et Goodhue D.L. (2000), Data Warehousing supports corporate strategy at first American corporations, *MIS Quarterly*, vo. 24, n°4, 547-67.

Berners-Lee T., Hendler J. et Lassila O. (2001), The Semantic Web, *Scientific American*, 284, 5, 35-43.

Davi A., Haughton D., Nasr N., Shah G., Skaletsky M., et Spack R. (2005), A review of two text-mining packages: SAS Text Mining and WordStat, *The American Statistician*, vol 59, fev., 89-103.

Garrouste D. et Lebourgeois S. (2002), La méthodologie du Text Mining illustrée par un cas concret, *Séminaire Exploitez vos données textuelles*, Paris.

Gavard-Perret M-L. et Moscarolla J. (1998), Enoncé ou énonciation ? Deux objets différents de l'analyse lexicale en marketing, *Recherche et Application en Marketing*, vol. 13, n°2, 31-47.

Gauzente C. (2006), E-marketing et textmining – Une application à l'analyse des opinions de consommateurs sur Internet, *8èmes Journées Internationales d'Analyse des Données Textuelles (JADT)*, Besancon.

Guernsey L. (2003), Digging for Nuggets of Wisdom, *The New York Times*, 16 octobre.

Helme-Guizon A. et Gavard-Perret M-L. (2004), L'analyse automatisée de données textuelles en marketing : comparaison de trois logiciels, *Décisions Marketing*, vol. 36, 75-90.

Kiang M.Y. et Kumar A. (2007), An evaluation of Self-Organizing Map Networks as a

robust alternative to Factor Analysis in Data Mining applications, *Information Systems Research*, vol. 12, n° 2, 177-94.

Kumar, A. (2001), Attention shifts from Firestone to Ford Explorer, *St Petersburg Times*, 17 juin.

Lebart L. (1998), Classification problems in text analysis and information retrieval, *in : Advances in Data Science and Classification*, A. Rizzi, M. Vichy, H-H.Bock (eds), p. 473-82, Springer, Berlin.

Lebart L. (2004), Validation techniques in Text Mining, *in S. Sirmakessis (Ed.) Text Mining and its applications – Results of the NEMIS Launch Conference*, 169-79.

Lebart L. et Salem A. (1994), *Statistique Textuelle*, Dunod, Paris.

Padmanabhan B., Zheng Z., et Kimbrough S.O. (2006), An empirical analysis of the value of complete information for eCRM models, *MIS Quarterly*, vol. 30, n° 2, 247-67.

Piatetsky-Shapiro G. et Frawley W.J. (1991), *Knowledge Discovery in Databases*, AAAI Press/ The MIT Press, Cambridge, MA.

Quatrain Y., Nugier S., Peradotto A. et Garrouste D. (2004), Evaluation d'outils de Text Mining : démarche et résultats, *7èmes Journées internationales d'Analyse des Données Textuelles (JADT)*.

Reinert M. (1993), Les « mondes lexicaux » et leur logique à travers l'analyse statistique d'un corpus de récits de cauchemars, *Langage et société*, vol. 66, 5-39.

SAS Institute Inc. (2006), *Data Mining Using Enterprise Miner Software: A Case Study Approach*. Cary, NC: SAS Institute Inc.

Speier C. et Morris M.G. (2003), The influence of query interface design on decision-making performance, *MIS Quarterly*, vol. 27, n° 3, 397-423.

Tufféry S. (2005), *Data Mining et Statistique Décisionnelle*, Technip, Paris.

James DESMECHT est chargé de recherche au LENTIC (HEC-Ecole de Gestion de l'Université de Liège, Belgique). Il s'intéresse à l'évolution du secteur TIC et aux innovations technologiques (Web 2.0, Open Source, etc.), aux stratégies e-commerce des entreprises et aux techniques de WebMarketing.

Adresse : LENTIC – HEC-ULg, Bd du Rectorat, 19, Bât. B.51, B-4000 Liège (Belgique).

Carine DOMINGUEZ Maitre de conférences à l'Université de Saint-Etienne, responsable du Master Management de projet à l'ISEAG et chercheur Coactis (EA4161). Docteur en sciences de gestion, spécialité systèmes d'information, agrégée d'économie et de gestion, ancienne étudiante de l'ENS Cachan et de l'INT Management. Domaine de recherche : business model, place de marché électronique, création de valeur, organisation et pilotage des SIIO, achats, Supply Chain.

Adresse : 6, rue basse des rives, 42023 Saint-Etienne Cedex 2

Mail : carine.dominguez(@)univ-st-etienne.fr

Manu CARRICANO est Professeur à l'EADA Barcelone. Ses travaux s'orientent vers l'optimisation des décisions en marketing (et en particulier la fixation des prix) et la convergence des méthodes quantitatives et qualitatives sur Internet.

Adresse : EADA Barcelona c/ Arago 204, 08011 Barcelona - SP

Mail : mcarricano(@)eada.edu

Grégoire DE LASSENCE. Responsable Pédagogie et Recherche chez SAS Academic. DESS SIAD Systèmes d'Information et d'Aide à la Décision.

Adresse : Domaine de Grégy – BP 5, 77166 Grégy-sur-Yerres - FR

Mail : Gregoire.DeLassence(@)fra.sas.com

Anthony HUSSENOT. Maitre de conférences à l'université de Paris Dauphine. Mes travaux portent sur les relations entre le collectif et les dispositifs de travail. Ces recherches visent à identifier les dynamiques sociales et techniques en oeuvre dans les processus organisationnels.

Adresse : Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16

Mail : anthony.hussenot(@)dauphine.fr.

Olivier LISEIN est assistant à HEC-Ecole de Gestion de l'Université de Liège (Belgique) et chargé de recherche au LENTIC (HEC-ULg). Ses travaux portent essentiellement sur l'introduction des TIC dans les organisations, les stratégies des entreprises et leurs politiques e-business/e-commerce, le management de l'innovation et la gestion du changement. Il conduit des recherches et assure des interventions en organisation dans ces domaines ; il anime également plusieurs modules d'enseignement et de formation sur ces thématiques.

Adresse : LENTIC – HEC-ULg, Bd du Rectorat, 19, Bât. B.51, B-4000 Liège (Belgique).

Mail : O.Lisein(@)ulg.ac.be

François PICHAULT, docteur en sociologie, est professeur ordinaire à HEC-Ecole de Gestion de l'Université de Liège (Belgique). Il préside, à l'Université de Liège, le LENTIC, un centre de recherche et d'intervention spécialisé dans l'étude des aspects humains et organisationnels des processus de changement et d'innovation technologique. Il est actuellement Directeur de la recherche de HEC-Ecole de gestion de l'Université de Liège. Il est également professeur affilié à l'Ecole Supérieure de Commerce de Paris (ESCP Europe).

Adresse : LENTIC – HEC-ULg, Bd du Rectorat, 19, Bât. B.51, B-4000 Liège (Belgique).

Mail : F.Pichault(@)ulg.ac.be