Mémo technique

LE DATAMINING

46, rue de la Tour
75116 Paris – France

Tél : 00 33 (0)1 73 00 55 00
Fax : 00 33 (0)1 73 00 55 01
http://www.softcomputing.com

SOMMAIRE

1 Synthèse : ce qu'il faut savoir …

Une évolution plutôt qu'une révolution : plutôt que de nouveaux outils révolutionnaires, le Data Mining représente plus simplement la concrétisation d'une évolution d'outils très techniques vers plus de simplicité et de convivialité. Le Data Mining bénéficie par ailleurs de la baisse continue du prix de la puissance informatique. Il est aux outils statistiques traditionnels ce que les PC sont aux terminaux passifs : plus proche de l'utilisateur tout en offrant plus de fonctions et de transparence.

Une destinée incertaine de la technologie : il est certain que le Data Mining et ses extensions perdureront. Il est, en revanche, moins évident de déterminer si la technologie gardera son autonomie ou tendra à se fondre dans nos outils de tous les jours (tableurs, Internet, base de données, requêteurs...), pour nous apporter des services de manière quasi transparente.

Des risques à suivre une démarche essentiellement technologique : les logiciels sont une composante essentielle du Data Mining. Mais la démarche suivie et la formulation du problème conditionneront les résultats. Il est donc primordial de structurer cette démarche et l'organisation du projet avant de plonger dans la sélection d'outils techniques. Qui plus est, il est logique que les responsables fonctionnels, par opposition aux informaticiens, participent au projet, voire pilotent, le Data Mining. Il s'agit, en effet, d'une réponse à des besoins métiers et les résultats n'ont de valeur qu'à la lumière de la connaissance de l'entreprise.

La possibilité de démarrer petit : tout en voyant grand, le Data Mining peut démarrer sur des opérations d'envergure restreinte avec des investissements limités et sur des extractions de données « manuelles ». Il n'est pas nécessaire de constituer au préalable un gigantesque entrepôt de données ni d'acquérir des logiciels de plusieurs millions pour pouvoir profiter du potentiel du Data Mining.

Une maturité des techniques de modélisation : les agents, les knowbots ou les fractales restent, certes, des domaines encore peu développés. Mais, dans leurs grandes lignes (arbres de décision, réseaux neuronaux, algorithmes génétiques, clusterisation, associations), les techniques de modélisation et les outils associés sont éprouvés. Ils apportent les résultats escomptés. En d'autres termes, malgré une présentation ou un vocabulaire parfois ésotérique, ça marche vraiment.

Une véritable opportunité : certains discours placent sans doute la barre trop haute en faisant des promesses inconsidérées. Néanmoins, une utilisation raisonnée du Data Mining apportera dans presque tous les cas des retours sur investissements largement supérieurs à la moyenne. Le pari du Data Mining, dans la mesure où il reste fondé sur des objectifs réalistes, est sans risque et peut apporter des sources de profit. Ses territoires d'applications sont encore très largement inexplorés, et donc à fort potentiel de marge de progrès.

2 Présentation générale et enjeux

2.1 Présentation générale

Le Data Mining est un sujet qui dépasse aujourd'hui le cercle restreint de la communauté scientifique pour susciter un vif intérêt dans le monde des affaires. La littérature spécialisée et la presse ont pris le relais de cet intérêt et proposent pléthore de définitions générales du Data Mining :

- « l'extraction d'informations originales, auparavant inconnues, potentiellement utiles à partir de données » ;
- « la découverte de nouvelles corrélations, tendances et modèles par le tamisage d'un large volume de données » ;
- « un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données » ;
- d'autres, plus poétiques, parlent de « torturer l'information disponible jusqu'à ce qu'elle avoue ».

Plus généralement, les spécialistes du domaine considèrent que la « découverte de connaissances dans les bases de données » (ou Knowledge Discovery in Database, abrégé en KDD) englobe tout le processus d'extraction de connaissances à partir de données. Le mot « connaissance » est compris ici comme étant un ensemble de relations (règles, phénomènes, exceptions, tendances...) entre des données.

2.2 Les principales applications du Data Mining

Le tableau ci-dessous propose une ventilation, non exhaustive, des principales applications recensées par secteurs d'activité :

- **Grande distribution et VPC** : Analyse des comportements des consommateurs,recherche des similarités des consommateurs en fonction de critères géographiques ou socio-démographiques, prédiction des taux de réponse en marketing direct, vente croisée et activation sélective dans le domaine des cartes de fidélité, optimisation des réapprovisionnements.

- **Laboratoires pharmaceutiques** : Modélisation comportementale et prédiction de médications ou de visites, optimisation des plans d'action des visiteurs médicaux pour le lancement de nouvelles molécules, identification des meilleures thérapies pour différentes maladies.

- **Banques** : Recherche de formes d'utilisation de cartes caractéristiques d'une fraude, modélisation prédictive des clients partants, détermination de pré-autorisations de crédit revolving, modèles d'arbitrage automatique basés sur l'analyse de formes historiques des cours.

- **Assurance** : Modèles de sélection et de tarification, analyse des sinistres, recherche des critères explicatifs du risque ou de la fraude, prévision d'appels sur les plates-formes d'assurance directe.

- Aéronautique, automobile et industries : Contrôle qualité et anticipation des défauts, prévisions des ventes, dépouillement d'enquêtes de satisfaction,

- **Transport et voyagistes** : Optimisation des tournées, prédiction de carnets de commande, marketing relationnel dans le cadre de programmes de fidélité.

- **Télécommunications, eau et énergie** : Simulation de tarifs, détection de formes de consommation frauduleuses, classification des clients selon la forme de l'utilisation des services, prévisions de ventes.

2.3 Les enjeux

En préambule, les résultats du Data Mining doivent, s'ils veulent prouver leur rentabilité, être intégrés selon les cas, soit dans l'informatique de l'entreprise, soit dans ses procédures. Ainsi, après avoir, par exemple, élaboré un modèle prédictif du départ d'un client à la concurrence, il faudra soit mettre en place des programmes pour calculer le risque de départ de chaque client, soit diffuser une procédure pour que les commerciaux appliquent manuellement ces règles et prennent les mesures adaptées. Cela étant posé, les opérations de Data Mining se soldent généralement par des gains significatifs tant en termes absolus (les francs gagnés) qu'en termes relatifs (les francs gagnés sur les francs investis). A titre indicatif, il n'est pas rare que les premières applications de Data Mining génèrent plus de dix fois l'investissement qu'elles auront nécessité, soit un retour sur investissement de l'ordre du mois !
Afin d'illustrer ce potentiel, nous avons pris trois cas concrets, maquillés pour des raisons évidentes de confidentialité :

- une banque veut améliorer son taux de transformation de rendez-vous commerciaux en vente de produits financiers : 60 millions de retour pour un investissement de 2 millions soit une durée de retour sur investissement de l'opération de datamining en 12 jours.
- un club de disques veut réduire le nombre de retours de son disque vedette : 16 millions de retour sur un investissement de 250.000 Francs soit une durée de retour sur investissement d'une dizaine de jours.
- une entreprise de vente par correspondance (VPC) cherche à améliorer le taux de rendement sur l'envoi de son catalogue spécialisé : 1 million de retour sur un investissement de 80.000 Francs soit une durée de retour sur investissement d'environ 30 jours.

3 Les technologies du datamining

3.1 Le processus du datamining

Quel que soit le domaine d'applications, une opération de datamining suit globalement un processus en huit étapes :

1. poser le problème
2. rechercher des données
3. sélectionner les données pertinentes
4. nettoyer des données
5. transformer les variables
6. rechercher le modèle
7. évaluer le résultat
8. intégrer la connaissance

3.2 Les principaux algorithmes de datamining

On dénombre sept techniques principales dans le domaine du datamining :

1. **Apprentissage fondé sur l'explication** (EBL pour Explanation Based Learning) : Apprentissage formé sur des explications dérivées d'une théorie (généralement incomplète) fournie en entrée. Cette forme d'apprentissage repose sur des déductions pour expliquer les données à partir de la théorie et sur des arbres de décision pour générer de la nouvelle connaissance.

2. **Apprentissage statistique** (STL pour Statistical Learning) : Cet apprentissage repose sur des opérations statistiques telles que la classification bayésienne ou la régression pour apprendre à partir de données.

3. **Apprentissage par réseaux neuronaux** (NNL pour Neural Network Learning):Un réseau de neurones est défini par un ensemble d'unités de traitement qui peuvent être des unités soit d'entrée, soit de sortie, soit cachées. L'apprentissage s'effectue par l'injection de cas en entrée et par la mesure des conclusions en sortie.

4. **Apprentissage par algorithme génétique** (GAL pour Genetic Algorithm Learning) : Les algorithmes génétiques sont des procédures de recherche fondées sur la dynamique de la génétique biologique. Ils comportent trois opérateurs, la sélection, la combinaison et la mutation, qui sont appliqués à des générations successives d'ensemble de données. Les meilleures combinaisons survivent et produisent, par exemple, des plannings, des règles...

5. **Apprentissage par similarité** (SBL pour Similarity Based Learning) : Ces techniques utilisent des indicateurs de similarité pour regrouper des données ou des observations et pour définir des règles.

6. **Apprentissage symbolique empirique** (SEL pour Symbolic Empirical Learning) : Cette forme d'apprentissage extrait des règles symboliques compréhensibles par l'utilisateur à partir de données. On retrouve dans cette catégorie les algorithmes ID3/C4.5 et CN2 notamment.

7. **Apprentissage par analogie** (ANL pour Analogy Learning) : L'apprentissage s'appuie sur l'analogie entre un nouveau cas et des cas ressemblants soumis préalablement.

Les principales natures de problèmes qui sont résolues par ces techniques sont au nombre de sept :

1. **Classification** : La capacité de classer des objets ou des événements comme membres de classes prédéfinies.

2. **Prédiction** : Liée à la classification, cette tâche vise à prédire une ou plusieurs caractéristiques inconnues à partir d'un ensemble de caractéristiques connues.

3. **Optimisation** : Il s'agit d'optimiser un ou plusieurs paramètres d'un système, compte tenu d'un ensemble de contraintes.

4. **Planning** : Cette tâche consiste à déterminer un ensemble d'actions ordonnées qui satisfont un ensemble donné de buts.

5. **Ordonnancement** : L'ordonnancement suit le planning et consiste à positionner des actions dans le temps et à leur affecter des ressources.

6. **Acquisition de connaissance** : L'acquisition de connaissances consiste à créer une représentation efficace et fidèle de la connaissance d'experts.

7. **Résolution de conflits** : La résolution de conflits peut, par exemple, aider à départager des experts qui sont en désaccord ou s'appliquer dans le cadre de processus de négociation.

3.3 La segmentation des éditeurs et des prestataires de Data Mining

Les fournisseurs de logiciels peuvent être segmentés selon leurs capacités à couvrir plus ou moins le triangle formé par l'acquisition de connaissance, la planification et la classification. Certains fournisseurs ont une stratégie de niche. Ils optimisent leur outil pour prendre le leadership sur un des sommets du triangle. Cette stratégie est notamment celle suivie par les fournisseurs de réseaux de neurones. D'autres développent une stratégie généraliste. Ils partent d'un des sommets du triangle et font évoluer leur offre pour couvrir l'ensemble de la surface, tout en développant des liens étroits entre leurs différentes offres. Cette stratégie est notamment mise en œuvre par les fournisseurs d'outils statistiques.
Puisque la demande s'oriente vers l'utilisation conjointe de différentes techniques, la stratégie généraliste semble gagnante à terme. Les approches sont différentes selon la taille et la notoriété des fournisseurs. Les plus importants mettent en place d'importantes équipes de chercheurs. Ceux-ci travaillent pour optimiser et intégrer les technologies dans une gamme logicielle. Certains choisissent de s'allier et s'occupent des passerelles de communication entre les produits. Les derniers venus sur le marché pratiquent une politique de rachat de technologies ou d'acquisition de sociétés.
L'offre d'outils de datamining est aujourd'hui atomisée. Aucun fournisseur d'outils ne peut se targuer d'être le standard du marché, ni même d'en détenir une part réellement significative mais une typologie des éditeurs se dégage néanmoins.

3.3.1 Les fournisseurs de logiciels statistiques

SAS Institute et SPSS, pour ne citer qu'eux, vendent depuis bien longtemps des outils de datamining… à l'image de Monsieur Jourdain qui faisait de la prose : sans le savoir. Ils restent aujourd'hui des intervenants majeurs du marché. La qualité et les capacités d'évolution de leurs offres sont démontrées. La connaissance des statisticiens, lesquels se positionnent, à tort ou à raison, comme les futurs « dataminers » dans leurs entreprises, leur assure un accès plus facile au marché. Ces éditeurs disposent de ressources conséquentes et de revenus récurrents (la plupart des logiciels statistiques font l'objet d'une location et non d'une vente) qui leur permettent de suivre une stratégie généraliste. Citons pour exemple la société SAS : elle propose son offre traditionnelle SAS System, ensemble de modules statistiques, et a su construire une offre intégrée de datamining connue sous le nom de SAS Enterprise Miner.

3.3.2 Les vendeurs de matériels

Il est naturel que les vendeurs de machines développent ou acquièrent des technologies de datamining afin de justifier la débauche de puissance qu'ils proposent à leurs clients. A ce titre, IBM, NCR, Siemens ou Silicon Graphic, pour ne nommer qu'eux, développent des offres d'outils de datamining. Ils s'appuient pour cela, soit sur leurs propres laboratoires de développement soit sur des accords de distribution avec des sociétés qui proposent des technologies de datamining. A l'heure actuelle, seuls IBM, avec Intelligent Miner, et Silicon Graphic, avec Mineset ont choisi une stratégie globale et fondée sur des développements internes.

3.3.3 Les startups

La plupart des algorithmes de base du datamining sont du domaine public. On peut en trouver tous les détails de fabrication dans les thèses de doctorat ou dans les comptes rendus des congrès spécialisés. Il est donc naturel que des individus développent à moindre coût des produits qui peuvent tout à fait rivaliser en performance et en qualité avec des solutions « poids lourd », tout au moins sur une ou deux des techniques de modélisation du datamining. La pérennité de ces entreprises monoproduits se jouera, comme c'est toujours le cas sur les marchés des technologies, avant tout sur leur capacité commerciale et marketing. Les meilleures auront dès lors deux solutions : se vendre à une entreprise avec laquelle elles présentent une complémentarité ou racheter des startups concurrentes pour enrichir leur gamme ou prendre des parts de marché.

3.3.4 Les intégrateurs

Ce panorama ne serait pas complet sans parler des sociétés de services, telles que Soft Computing en France. Elles proposent des prestations d'application ou de transfert de technologies autour du datamining. Elles n'ont pas d'offres d'outils à proprement parler, mais disposent de nombreux logiciels. Elles apportent des conseils pour choisir les outils, former les équipes internes et apporter une aide méthodologique dans la résolution de problèmes concrets. Elles présentent une alternative viable aux entreprises qui ne justifient pas de la taille critique, qui ne disposent pas des compétences internes requises, ou qui, tout simplement, souhaitent que l'intégration de ces technologies soit progressive et guidée par des professionnels expérimentés.

4 Le marché

4.1 Evolution du marché : taux de croissance supérieur à 40 % par an

Les entreprises s'intéressent de plus en plus au datamining, probablement à cause des promesses de rentabilité immédiate que vantent les fournisseurs de technologies et dont les médias se font l'écho. Une étude récente du cabinet IDC, spécialisé dans les études quantitatives des marchés de technologies, souligne que plus de la moitié des entreprises américaines ont ou vont acheter un outil de datamining. Les différentes études de marché estiment que la taille du marché du datamining était d'environ 300 millions de francs (source : Meta Group) et tablent sur 5 milliards en l'an 2000, soit un taux de croissance annuelle de plus de 40 %.

4.2 La segmentation des outils

L'offre des logiciels de datamining est aujourd'hui encore largement atomisée. Il est impossible d'en dresser un panorama exhaustif. De nouveaux produits, toujours plus puissants, toujours plus innovants, sont annoncés régulièrement. L'étude du Gartner Group, dont nous avons reproduit une adaptation ci-après, propose pour sa part une segmentation du marché selon le prix et le niveau de compétence nécessaire. L'axe prix distingue *grosso modo* trois gammes principales :

1. **Les outils de micro-mining**, qui valent moins de 10 000 F (Alice, Scenario, DataMind, Solo,etc.), sont encore rares. Ils s'adressent en général à des utilisateurs finals. Ils sont issus des versions allégées ou des produits d'appel dans une gamme plus large. Outils dédiés à un type d'algorithme unique (arbres de décision ou réseaux de neurones), ils offrent la particularité d'être faciles à utiliser et conviviaux.

2. **Les outils intermédiaires**, qui tournent autour de 100 000 F (de 50 000 à 200 000 F), constituent le gros de l'offre à l'heure actuelle. Il s'agit d'une gamme composite qui comprend à la fois des versions évoluées de la gamme des « outils PC de bureau » (Clementine, Knowledge Seeker, DataMind Pro...), et des concurrents agressifs des « poids lourds », comme Clementine. Dans cette gamme, on trouve généralement des solutions qui fonctionnent sur PC sous Windows ou NT et sur Unix. Elles proposent souvent conjointement des assistants pour les néophytes et des fonctions avancées pour les experts.

3. **Les poids lourds du macro-mining** (IBM Intelligent Miner, NeoVista Decision Series ou Silicon Graphics Mineset) coûtent plus de 200 000 F, voire même un million ou plus, quand ils ne sont pas uniquement en location. Il s'agit en général de solutions fonctionnant sur des machines Unix, parfois sur des superordinateurs. Ils proposent un ensemble de modules déclinant plusieurs types d'algorithmes dans un ensemble intégré. Ce type de logiciel met avant tout l'accent sur la puissance de traitement et les algorithmes. La cible est résolument un marché de spécialistes.

4.3 Quelques outils

Intelligent Miner	d'IBM
Volumes	Pas de limites
Liens aux données	DB2, fichiers
Méthodes de modélisation	Multiples
Intégration des résultats	API
Catégorie	Poids lourd
Utilisateurs	Experts

Clementine	de SPSS
Volumes	Peu de limites
Liens aux données	SGBD et fichiers
Méthodes de modélisation	Multiples
Intégration des résultats	API
Catégorie	Intermédiaire intégré
Utilisateurs	Avertis

SAS Enterprise Miner	SAS
Volumes	Peu de limites
Liens aux données	SAS, SGBD et fichiers
Méthodes de modélisation	Multiples
Intégration des résultats	
Catégorie	Poids lourd
Utilisateurs	Avertis

4Thought	de Cognos
Volumes	Peu de limites
Liens aux données	SGBD
Méthodes de modélisation	Réseaux de neurones
Intégration des résultats	Programme Excel ou langage C
Catégorie	Intermédiaire spécialisé
Utilisateurs	Avertis

Predict	de NeuralWare
Volumes	Quelques milliers d'enregistrements
Liens aux données	SGBD
Méthodes de modélisation	Réseaux de neurones
Intégration des résultats	Sans objet
Catégorie	PC de bureau
Utilisateurs	Néophytes

Previa	de Elseware
Volumes	Quelques milliers de records
Liens aux données	Fichiers
Méthodes de modélisation	Réseaux de neurones
Intégration des résultats	
Catégorie	PC de bureau
Utilisateurs	Néophytes

Saxon	de Pmsi
Volumes	Peu de limites
Liens aux données	Fichiers
Méthodes de modélisation	Réseaux de neurones
Intégration des résultats	Programme C
Catégorie	Intermédiaire spécialisé
Utilisateurs	Experts

Strada	Complex System
Volumes	Quelques milliers de records
Liens aux données	Fichiers
Méthodes de modélisation	Réseaux de neurones Algorithmes génétiques
Intégration des résultats	
Catégorie	Intermédiaire spécialisé
Utilisateurs	Avertis

Scenario	de Cognos
Volumes	Quelques milliers d'enregistrements
Liens aux données	SGBD, Fichiers
Méthodes de modélisation	Arbres de décision
Intégration des résultats	
Catégorie	PC de bureau
Utilisateurs	Néophytes

Alice	de Isoft
Volumes	Quelques milliers d'enregistrements
Liens aux données	SGBD, Fichiers
Méthodes de modélisation	Arbres de décision
Intégration des résultats	
Catégorie	PC de bureau
Utilisateurs	Néophytes

Knowledge Seeker	Angoss
Volumes	Quelques milliers d'enregistrements
Liens aux données	SGBD, Fichiers
Méthodes de modélisation	Arbres de décision
Intégration des résultats	SQL
Catégorie	Intermédiaire spécialisé
Utilisateurs	Néophytes

Datamind	D'Epiphany

Volumes	Quelques milliers d'enregistrements
Liens aux données	SGBD, Fichiers
Méthodes de modélisation	Propriétaire (proche de Bayes)
Intégration des résultats	
Catégorie	PC de bureau/Intermédiaire
Utilisateurs	Néophytes

Wizwhy	de Wizsoft
Volumes	Quelques milliers d'enregistrements
Liens aux données	SGBD, Fichiers
Méthodes de modélisation	Associations
Intégration des résultats	
Catégorie	PC de bureau
Utilisateurs	Néophytes

SPAD	de CISIA
Volumes	Quelques milliers d'enregistrements
Liens aux données	SGBD, Fichiers
Méthodes de modélisation	Multiples
Intégration des résultats	Fichiers
Catégorie	PC de bureau/Intermédiaire
Utilisateurs	Avertis

5 Mise en œuvre de la technologie

5.1 Quelques pièges à éviter

L'introduction de nouvelles méthodes et technologies comprend toujours une dose de risque pour l'entreprise. Sans prétendre être exhaustifs, voici quelques pièges à éviter :

- **Attention à la qualité des données**: avant de promettre des retours mirobolants du datamining sur un domaine particulier, assurez-vous que les données dont vous aurez besoin sont suffisamment fiables. A l'usage, il s'avère qu'une forte proportion de données d'un système d'informations est entachée d'erreurs !
- **Evitez une démarche centrée sur les outils** : les techniques de modélisation, et donc les logiciels, ne peuvent être sélectionnés qu'une fois le problème à traiter correctement formulé. Ne vous laissez pas tenter par tel ou tel fournisseur qui vous propose un outil miracle ; restez concentré sur le processus de datamining plutôt que focalisé sur des outils.
- **Ne substituez pas le Data Mining aux statistiques** : une erreur grave consiste à remplacer les outils (et les équipes) statistiques par du datamining. Cette position risque de créer des conflits internes en mettant en compétition des techniques et des hommes. En définitive, le datamining et les statistiques sont complémentaires et doivent nécessairement collaborer.
- **N'oubliez pas l'intégration dans le système d'informations** : la construction d'un modèle a souvent un objectif opérationnel qui passe par une application du modèle à des données de l'entreprise (affectation d'un score à des clients, calcul quotidien des prévisions de stocks…). Cette phase d'application doit être considérée dès le démarrage d'une opération de datamining, tant sur le plan des données (un modèle parfait mais qui travaille sur des données inexistantes dans l'entreprise n'est d'aucune utilité) que sur le plan des technologies (un réseau de neurones sur PC parfaitement apte à prédire le risque d'un client sera très difficilement intégrable dans une transaction de saisie de dossiers de crédit sur site central).
- **Ne négligez pas la communication et la mise en application** : le processus de datamining étant arrivé à son terme, il reste encore à en communiquer les résultats et à en assurer la mise en application. Ces deux étapes sont fondamentales pour que les promesses de retour sur investissement se concrétisent effectivement et ne restent pas de simples chiffres abstraits posés sur un transparent ou dans un rapport.
- **Anticipez la résistance au changement** : le datamining ne révolutionne pas l'entreprise. Cependant, il rend certains changements nécessaires. Or, les organisations présentent toutes des résistances au changement. Une communication adéquate autour du datamining et une transparence des objectifs visés peuvent contribuer à créer un consensus autour du changement, plutôt qu'une levée de boucliers.
- **Faites participer les utilisateurs** : les connaissances contenues dans les données ne sont finalement qu'une partie de la connaissance de l'entreprise. Les expertises internes, les procédures et les orientations stratégiques sont autant de sources qu'il faut assembler avec les résultats du datamining pour obtenir des modèles probants. Dans ces conditions, les utilisateurs, c'est-à-dire les directions fonctionnelles concernées, doivent être impliqués dans le processus de datamining dans le cadre d'un contrôle continu. Le challenge n'est pas de construire un modèle à partir des données, mais de pouvoir prendre en compte le maximum de connaissances qui sont externes aux données.
- **Démystifiez le datamining**: certaines techniques, on l'a vu, cultivent un certain ésotérisme, tant dans les termes utilisés que dans la transparence des résultats. Pour éviter de positionner le datamining comme une technique d'hyperspécialistes, il est souvent nécessaire d'accompagner sa mise en place d'une communication et de formations sur les concepts.

5.2 Démystification du datamining

- **Question** : le datamining produit des résultats si surprenants qu'il va profondément révolutionner votre métier.

Réponse : certains phénomènes décelés dans les données peuvent effectivement remettre partiellement en cause l'organisation d'une entreprise, mais nous n'avons jamais vu de révolutions organisationnelles déclenchées par le datamining.

- **Question** : le datamining est si sophistiqué qu'il se substitue à la connaissance et à l'expérience des experts pour la construction des modèles.

Réponse : aucune technique d'analyse de données ne remplacera l'expertise humaine. Le datamining se marie parfaitement bien avec des techniques de recueil de connaissance soit en parallèle, soit en tant que catalyseur de la réflexion pour édicter des règles d'experts. Qui plus est la qualité de l'interprétation des résultats du datamining dépendra avant tout de la capacité de l'analyste à comprendre le problème dans son contexte métier.

- **Question** : les outils de datamining trouveront automatiquement les « formes » que vous cherchez sans qu'il soit nécessaire de les leur préciser.

Réponse : le datamining devient d'autant plus efficace que le problème est bien posé. Si les outils actuels peuvent effectivement explorer de manière complètement autonome des bases, la plupart des utilisations constatées sont liées à des objectifs clairement énoncés.

- **Question** : le datamining n'est utile que pour le marketing, les ventes et la détection de fraude.
-

Réponse : ces domaines sont effectivement les plus porteurs actuellement, compte tenu des marges de progrès qu'ils recèlent et de la tangibilité des résultats obtenus. Ils n'en sont pas pour autant les domaines exclusifs d'application : nous travaillons actuellement sur des applications portant sur l'aide à la navigation sur Internet, l'audit de comptes, le contrôle de qualité ou l'optimisation de processus organisationnels, et nous découvrons tous les jours de nouvelles applications. Globalement défini, le datamining peut s'avérer pertinent dans tous les domaines où le volume d'informations sur un sujet est important.

- **Question** : le datamining est une révolution par rapport aux statistiques « traditionnelles ».

Réponse : les méthodes proposées par la génération actuelle des outils de datamining sont des extensions de méthodes qui datent parfois de plusieurs dizaines d'années. Les premiers réseaux de neurones ont vu le jour dans les années 40, les algorithmes de création d'arbres (CART, Chaid) étaient utilisés par les démographes dans les années 60 et étaient proposés depuis longtemps dans des outils statistiques tels SAS ou SPSS. En outre, certaines techniques statistiques « traditionnelles », comme les clusters, relèvent parfaitement de la définition que l'on peut faire du datamining comme étant une technique exploratoire plutôt que confirmative.

- **Question** : le datamining est un processus très complexe.

Réponse : les algorithmes de datamining peuvent être complexes mais la caractéristique commune à tous les nouveaux outils est de contribuer à masquer cette complexité par des assistants à l'utilisation et une interface utilisateur conviviale. La tâche en général la plus complexe sur le plan technique sera la préparation des données, qui n'est, en aucun cas, spécifique au datamining. Sur le plan fonctionnel, il s'agira d'être pertinent dans l'interprétation des résultats, ce qui, au final, reste avant tout une question de bon sens et de connaissance du métier.

- **Question** : il faut un entrepôt de données avant de se lancer dans le datamining.

Réponse : si c'est en effet une condition souhaitable, ce n'est nullement un prérequis nécessaire. Au contraire, il arrive souvent qu'une entreprise utilise des techniques du datamining en se fondant sur des extractions de données ponctuelles, voire sur l'acquisition de données externes. Cela lui permet de dégager des marges financières à court terme, lesquelles peuvent ensuite contribuer au financement d'une démarche plus globale de mise en place d'un entrepôt de données.

- **Question** : le Data Mining est d'autant plus efficace qu'il travaille sur un gros volume de données.

Réponse : accroître le nombre de données n'a de sens dans un processus de datamining que dans la mesure où les données ajoutées augmentent la précision ou la puissance du modèle. A l'extrême, utiliser trop de données au départ peut aboutir à extraire de la connaissance inutile et à masquer des relations essentielles.

- **Question** : développer un modèle sur un échantillon d'une base de données est inefficace car l'échantillonnage tend à biaiser le modèle.

Réponse : il s'agit en réalité de trouver un optimum entre la performance du modèle et les efforts nécessaires pour le bâtir. En d'autres termes, votre problème justifie-t-il que, pour augmenter de 1 % votre taux de prédiction, vous multipliiez par 10 la taille de votre échantillon et, par conséquent, les temps de traitements et de préparation ainsi que les risques d'erreurs ? En outre, les sondages portant sur 1 000 personnes ne sont-ils pas communément acceptés comme représentatifs d'une population de plusieurs dizaines de millions d'habitants ? Par ailleurs, il arrive fréquemment que le datamining appliqué à une base complète aboutisse rapidement à la définition de sous-ensembles homogènes constituant autant d'ensembles qui feront l'objet d'analyses distinctes.

- **Question** : le datamining n'est qu'un phénomène de mode qui disparaîtra aussi vite qu'il est apparu.

Réponse : certainement amené à évoluer dans ses offres et ses applications, le datamining, en tant que technologie, est appelé à se développer et à perdurer. Comme telle, il s'insère, en effet, totalement dans l'orientation globale de l'informatique qui tend à engranger de plus en plus d'informations à partir desquelles il est

6 L'évolution prévisible

6.1 Performance et accessibilité

La tendance générale des outils de datamining se profile dans deux directions opposées : l'accessibilité et la performance.

- **L'accessibilité** : les outils masquent de plus en plus la complexité des modèles. Ils offrent des assistants méthodologiques couvrant l'ensemble du processus. Le datamining se démocratise au fur et à mesure que les outils deviennent plus conviviaux et proposent des assistants sophistiqués pour prendre en charge la modélisation.

- **La performance** : les algorithmes de prédiction sont toujours plus élaborés et performants. Ils évoluent vers des prévisions de plus en plus précises. Ils prennent en compte de mieux en mieux des données bruitées ou incomplètes. Cette tendance est encore accentuée par l'augmentation constante de la puissance machine, qui rend aujourd'hui accessibles des analyses encore inconcevables il y a quelques années.

Cette double tendance a pour conséquence paradoxale, d'une part de remettre entre les mains des utilisateurs finals des techniques jusqu'alors réservées à des spécialistes de la statistique et, d'autre part de contribuer à créer une nouvelle caste de spécialistes pour piloter et paramétrer des algorithmes toujours plus puissants. Les évolutions récentes des logiciels confirment que le datamining se dirige clairement vers une technologie à deux vitesses qui correspond à deux cibles : des spécialistes pour la mise en œuvre de modélisations sophistiquées et des utilisateurs finals pour l'exploration guidée de données.

6.2 Le rapprochement de l'OLAP et du datamining

Une tendance des technologies de datamining est à l'intégration des algorithmes dans les outils d'interrogation et de visualisation. Cette alternative est technique car elle aura les mêmes résultats pour l'utilisateur final. Des éditeurs d'outils de requêtes ou d'OLAP tels que Business Objects, avec les accords signés avec Isoft pour son produit Alice, ou Cognos, dans le cadre de son partenariat avec Angoss pour le développement de son offre Scenario, illustrent cette tendance.
Pour apporter plus de valeur ajoutée et se distinguer les uns des autres, il n'y a aucun doute sur le fait que les acteurs du marché des outils de reporting chercheront, tout comme les fournisseurs de SGBD/R, à intégrer des capacités de datamining dans leurs outils. Cette tendance semble d'autant plus inéluctable que la cible, pour ces outils, est l'utilisateur final, et que celui-ci cherche toujours plus de facilité, de souplesse et de convivialité dans l'accès à l'information.

6.3 Le datamining et le multimédia

L'essentiel des outils et des applications du datamining cible le domaine de l'exploitation de données structurées. Cependant, les réseaux de neurones, par exemple, ont de longue date été utilisés dans le cadre de la reconnaissance d'image ou d'écriture. Un certain nombre de recherches ou de produits émergents traitent du Data Mining appliqué à des informations multimédias :

- **Text mining** : l'analyse de documents pour la recherche d'associations de mots ou de concepts. Ces techniques sont utilisées pour mettre en relation des cibles et le vocabulaire utilisé dans la communication, ou pour prendre en considération les commentaires des clients sur des enquêtes de qualité.

- **Image mining** : il s'agit de rechercher des relations entre des images ou des séquences d'images. Ainsi, l'image mining peut, par exemple, contribuer à rechercher des similarités entre des images médicales pour trouver une pathologie semblable.

- **Vidéo Mining** : le vidéo mining est une extension de l'image mining dans le domaine de la vidéo. Très théorique pour l'instant, compte tenu de la puissance machine qu'il nécessite, le video mining consiste à rechercher des éléments communs ou à classer des vidéos en fonction de leur contenu. Les applications potentielles sont l'indexation de banques de films ou l'optimisation des grilles de programmes des opérateurs de télévision.

Il convient de souligner, en conclusion sur le multimédia mining, que seule l'exploration de texte peut aujourd'hui être considérée comme industrialisable à court terme, avec des produits d'ores et déjà disponibles chez IBM, SLP Infoware ou Le Sphinx-Lexica.

6.4 Datamining et Internet

Actuellement, les principaux champs d'interactions entre Internet et le datamining se situent à trois niveaux :

- D'une part, Internet bouleverse les facteurs de coûts de collecte de l'information et de création de bases de données sur les comportements des clients. Il contribue ainsi à constituer des bases de données importantes sur lesquelles le datamining s'avère intéressant, voire indispensable.
- D'autre part, le datamining apporte des solutions innovantes pour guider l'utilisateur d'Internet dans ses recherches et dans sa navigation sur le Web.
- Enfin, l'Internet constitue pour les outils de datamining une interface utilisateur de prédilection avec l'éventail d'avantages qui en découle généralement : centralisation de l'administration, postes clients plus légers, ubiquité de l'accès au traitement, portabilité entre systèmes d'exploitation des interfaces utilisateur, compatibilité avec les Network Computers...

Ces trois intersections sont les plus évidentes. Elles font déjà l'objet d'offres plus ou moins développées, pour la plupart affublées du qualificatif webmining (Cf. mémo technique de Soft Computing sur le Webmining) sur le marché. Mais on peut faire confiance à la créativité sans limite des startups de la Silicon Valley et à l'excitation des investisseurs dès qu'ils entendent les mots d'Internet ou de datamining, pour que de nouvelles formes de collaboration entre ces deux technologies émergent.

6.5 Vers une verticalisation du datamining?

On assiste depuis peu de temps à l'émergence de solutions verticales dans le domaine du datamining. Ces offres comprennent non plus seulement un outil d'analyse mais également un ensemble de paramétrages pour son utilisation sur un domaine particulier. L'idée est séduisante car elle nous promet de profiter de la puissance du datamining sans investir en compétences spécifiques. Citons, par exemple, SLP qui propose un outil d'analyse des départs des clients dans le domaine des télécommunications, ou VM Data avec un outil prêt à l'emploi pour construire des ciblages. La question se pose de savoir s'il s'agit de « coups marketing » ou d'une tendance lourde du marché.

6.6 Data Mining et liberté du citoyen

Les champs d'application du datamining sont multiples, mais un domaine de prédilection reste la connaissance du client et ses applications dans le marketing direct. Tôt ou tard, une exploitation trop efficace des données disponibles pour connaître ou prédire les comportements individuels et s'immiscer toujours plus dans l'intimité du consommateur posera des problèmes éthiques de respect de la vie privée des individus.

La France s'est dotée très tôt d'un organisme, la CNIL, dont la vocation est d'éviter que l'informatique n'empiète sur la vie privée des citoyens, essentiellement en délimitant ce qui est autorisé en matière de stockage et d'exploitation des données nominatives. Elle est garante d'un texte dont le premier article stipule : « L'informatique doit être au service de chaque citoyen. Son développement doit s'opérer dans le cadre de la coopération internationale. Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques. »

Dans le domaine du datamining et du Data Warehouse, la CNIL s'est récemment positionnée en affirmant, en réponse à la plainte dont elle avait été saisie à l'encontre d'une banque, que « comme tout traitement portant sur

des données relatives à des personnes physiques, les méthodes de ciblage de la clientèle doivent être conformes à la loi sur l'Informatique et les Libertés ». Confirmant ainsi la légalité du datamining, elle appelle néanmoins les entreprises à déclarer leurs opérations de datamining. Ainsi, la segmentation et le ciblage de clientèle ne doivent pas prendre en considération des critères de race, de religion ou d'opinion politique, ni aboutir à des qualificatifs péjoratifs ou défavorables.

L'avenir nous dira comment les consommateurs accepteront le comportement de « Big Brother » induit par le datamining. L'expérience passée du marketing direct laisse à penser qu'il est finalement probable que la personnalisation à outrance entrera dans les mœurs et que, d'ici quelques années, le marketing hyperpersonnalisé ne nous choquera pas plus que les dizaines de mailings banalisés que nous recevons chaque semaine dans nos boîtes aux lettres.

7 Fiche d'identité

Raison sociale :	**Soft Computing**
Forme juridique :	SA au K de 635 367,20 (RCS : B 330 076 159 : NAF : 721 Z, Conseil en systèmes informatiques)

Société cotée au nouveau marché de la bourse de Paris depuis le 26 avril 2000 (Code ISIN FR 0000075517) (Reuters : SFTC.LN, Bloomberg : SOFT NM)

Date de création :	1984
Effectif :	280 personnes (prévision de recrutements pour 2004 : 50 personnes)
Activité :	Conseil et solutions pour transformer l'information en performance notamment appliqués à la Gestion de la Relation Client et du Risque.

Dirigeants :

Président directeur général :	Eric FISCHMEISTER
Directeur Général :	Gilles VENTURI
Secrétaire Général :	Arnaud MALLAT DESMORTIERS
Directeur Général Adjoint :	René LEFEBURE
Directeur d'activité Conseil :	Fabrice OTANO
Directeur d'activité Datamining :	Didier RICHAUDEAU
Directeur d'activité Ingénierie :	Jean-François KLEINFINGER

Contacts presse :	Soft Computing
	Valérie Parent
	vpr@softcomputing.com
Adresse :	46, rue de la Tour
	75116 Paris
Téléphone :	01 73 00 55 00
	01 73 00 55 01
Site Web :	http://www.softcomputing.com

Ouvrages de références :

Titre :	« Gestion de la relation Client » Panorama des produits et conduite de projets	« Le datamining »
Auteurs :	Gilles Venturi et René Lefébure	Gilles Venturi et René Lefébure
Collection :	EYROLLES – 2000- réédition en 2004	EYROLLES – 1998- 2001