Engendrement de connaissances à partir de données industrielles

Famili, Fazel; Dubé, François

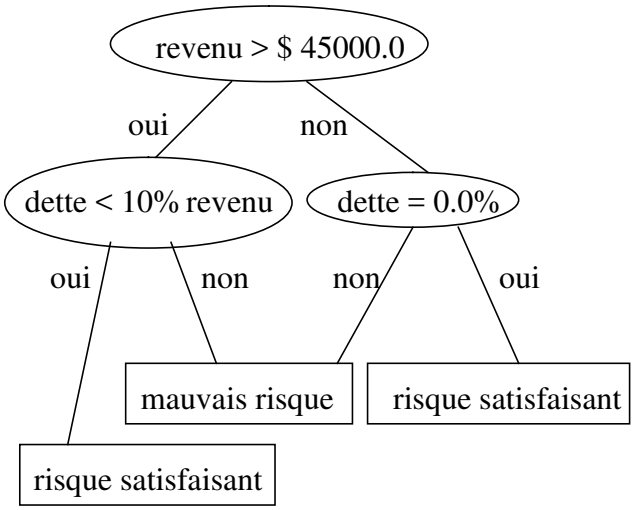

Figure 3: Example d'arbre de décision

Biographie

A. (Fazel) Famili est un agent de recherche senior à l'Institut de technologie de l'information (ITI) du Conseil national de recherches de Canada (CNRC). Ses recherches actuelles portent sur l'apprentissage automatique, l'acquisition automatique des données et l'engendrement de la connaissance. Il a obtenu sa maîtrise à l'Université de l'État d'Ohio et son doctorat à l' Université de l'État du Michigan. Il a oeuvré dans l'industrie durant trois années avant se joindre au CNRC en 1984. Il a publié plus de 20 articles dans les domaines de l'apprentissage automatique et de l'analyse des données. Il est aussi co-auteur du livre intitulée "*AI Applications in Manufacturing*" publié en 1992 par la maison d'édition AAAI/MIT.

François Dubé est agent de recherche à l'ITI/CNRC. Son domaine d'expertise est le suivi de la condition et le diagnostic de systèmes mécaniques complexes. Il est aussi intéressé par la gestion de projets impliquant le développement de nouvelles technologies. Il a obtenu une maîtrise en génie mécanique de l'Université Laval en 1984 et complète un programme de maîtrise en gestion de projets à l'Université du Québec à Hull, il s'est joint au CNRC en 1989 après une courte carrière dans les forces armées canadiennes comme officier de génie aérospatial.

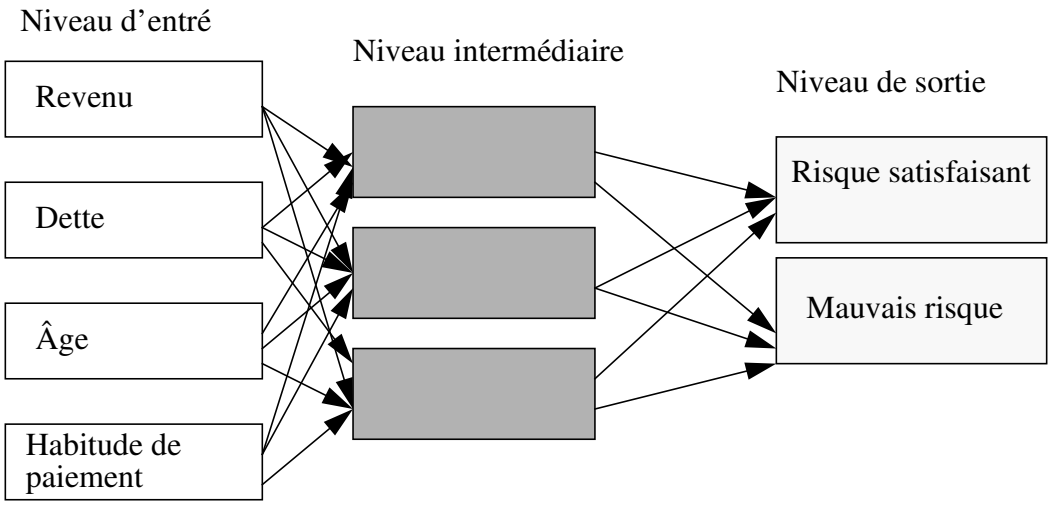

Figure 2: Example de réseau neuronique

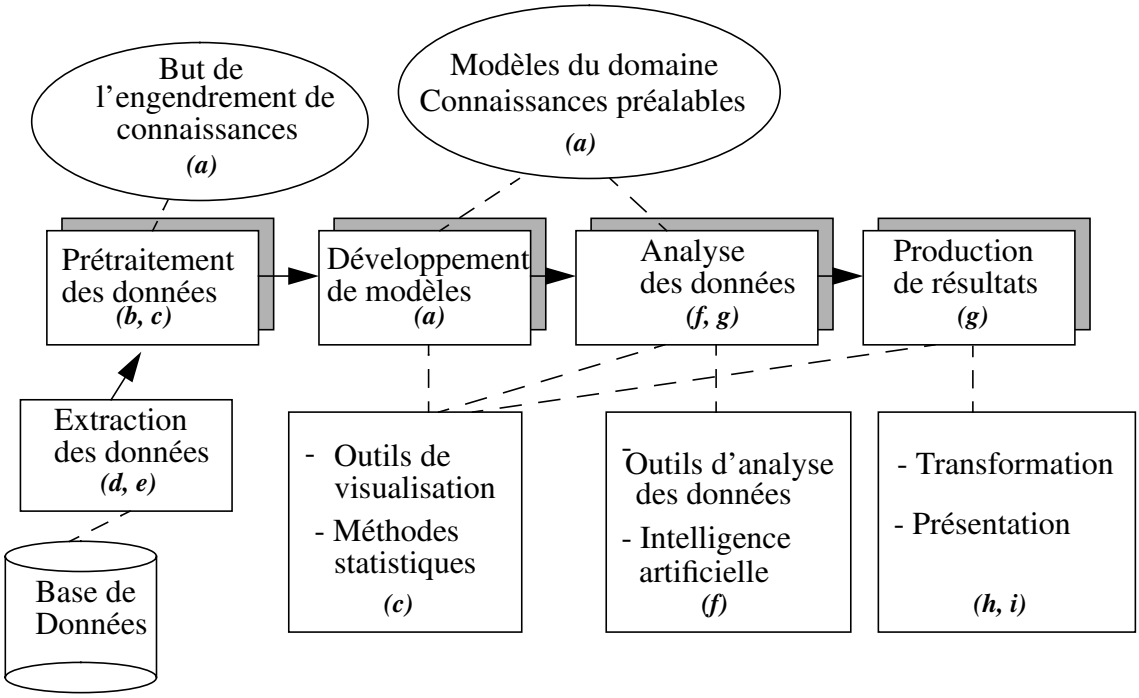

Figure 1: Le processus d'engendrement de connaissances

FOLEY J., *Data Dilema*, Information Week, June 1996, p. 14-16.

GLYMOUR Clark, *etal*, *Statistical Inference and Data Mining*, 1996, Communications of the ACM, vol. 39, no. 11, p. 35-41.

IMIELINSKI Tomasz et MANNILA Heikki, *A Database Perspective on Knowledge Discovery*, Communications of the ACM, vol. 39, no. 11, 1996, p. 58-64.

MENA Jenus, *Automatic Data Mining*, PC-AI, vol.10, no. 6, 1996, p. 16-20.

6 Conclusions et défis

Il est important de garder à l'esprit que l'engendrement de connaissances utiles n'est pas magique. L'achat du logiciel le plus dispendieux pour l'analyse de téraoctets de données sans préparation préalable ne produira propablement pas de résultats utiles. Cela prendra beaucoup de temps pour obtenir une réponse qui sera probablement sans valeur. Mais lorsque compris et appliqué convenablement, le processus d'engendrement de connaissances peut fournir des bénifices considérables à partir de l'utilisation judicieuse des données de l'industrie.

L'engendrement de connaissances est un processus complexe qui demande une implication humaine. Ce qui signifie que l'implication d'un expert humain est bénifique pour presque toutes les étapes du processus. Nous avons conclus que pour fournir un support efficient au client, le système logiciel utilisé pour l'engendrement de connaissances doit agir à titre d'assistant. Ceci principalement parce que le processus est très itératif et est complètement dépendant des buts de l'utilisateur.

Il est aussi important de noter que la plupart des clients sont indifférents au contenu (i.e. les algorithmes et méthodes) des systèmes d'engendrement de connaissances ou comment ils fonctionnent. Leurs intérêts sont plutôt: la facilité d'utilisation du système, la fiabilité des résultats produits, l'obtention de résultats faciles à interpréter et à utiliser et le moins de surcharge de travail possible généré par l'utilisation du système.

Le premier défi est donc de tenir compte des intérêts des clients mentionnés ci-haut dans toutes applications d'engendrement de connaissances. Un des défis les plus importants est le développement d'un outil qui serait capable de reconnaître la nature des données et de sélectionner un algorithme approprié et sous quelle forme pour l'engendrement de connaissances utiles. L'utilisation approprié des connaissances du domaine et l'application du processus d'engendrement de connaissances dans un environnement changeant sont aussi des défis importants.

Bibliographie

BRACHMAN Ronald. J. *etal*, "*Mining Business Databases"*, communications of the ACM, vol. 39, no. 11, 1996, pp. 42- 48.

BREIMAN Leo *etal*, "*Classification and Regression Trees"*, 1984, Wardsworth& Brooks.

CHARLES Darling, "*How to Integrate Your Data Warehouse*, Datamation", vol. 41, no. 23, 1996.

CHENG B. et TITTERINGTON D.M. "Neural Networks- a Review from a Statistical Perspective", 1994, Statistical Science, vol. 9 no 1, p. 2-30.

FAYYAD Usama, PIATETSKY-SHAPIRO Gregory, et SMYTH Padhraic, "*Advances in Knowledge Discovery and Data Mining"*, 1996, AAAI/MIT Press.

FAYYAD Usama, HAUSSLER David, et STOLORZ Paul, *Mining Scientific Data*, Communications of the ACM, vol. 39, no. 11, 1996, p. 51-57.

Pour chacun des cas potentiels identifiés, les tâches suivantes ont été exécutées:

- Faire la sélection des cas (ensemble d'observations) pour lesquels il y a une quantité suffisante de données disponibles. Pour chaque cas, un ensemble de données est composé d'une matrice de valeurs d'attribut, préférablement chaque vecteur devrait contenir au moins 100 données. Donc l'ensemble de données utilisé était constitué d'environ 100 registres de mesures représentant les différents paramètres.
- Choisir les paramètres appropriés (paramètres dépendants) pour la définition du problème (e.g. *exhaust-temp or fuel-flow* parmis les paramètres reliés au moteur). L'information de base pour l'identification des paramètres dépendants provient habituellement des experts du domaine ou de la documentation des systèmes.
- Définir les seuils de détection pour la définition du problème (e.g. un problème relié à la température des gaz d'échappement, *high-egt: exhaust-temp > 650*). L'information de base pour l'identification des seuils limites provient aussi des experts du domaine ou de la documentation des systèmes. Toutefois, des techniques de visualisation des données et des histogrammes des paramètres dépendants ont été utilisées pour la sélection et la définition des seuils appropriés.
- Utilisation d'information additionnelle, comme l'élimination de paramètres particuliers n'ayant aucune influence sur le problème. Cette tâche a pour but d'obtenir des données plus précises pour le processus d'exploration des données. Par exemple, si il y a une corrélation fortuite entre un paramètre dépendant et un paramètre indépendant qui ne peut pas techniquement avoir d'incidence sur le problème, ce paramètre indépendant devrait être masqué durant le processus d'induction.

Voivi un exemple des résultats obtenus; le résultat est sous la forme d'une règle:

```
Définition du problème: p1 est au dessus de 75
Incapacité d'utiliser 0.0% (0 de 554) registres
Le problème est présent 21.3% (118 de 554) registres
Règle 1:
     Variable 1: paramètre x >= 1.75
          r-carré 0.9
          couverture 90.7%
          taux d'erreur 2.7%
          qualité 9.2
```

Cette règle fournit quatres informations: le nom du paramètre (x) parmis tous les paramètres des données qui influencent le problème, le seuil (1.75) au dela duquel cette influence est valide, la direction de dépassement du seuil (>=), et finallement des valeurs statistiques (r-carré, couverture, taux d'erreur, qualité) qui indique la fiabilité de la règle.

- Réduction des résultats, en éliminant les informations non-pertinentes ou moins fiables. Cette tâche fait intervenir la connaissance du domaine et les informations obtenues du processus. Les résultats de l'induction de règles peuvent être présentés aux experts du domaine sous différentes formes. Nous avons présenté les résultats de deux façons différentes: (i) un résumé des règles les plus fiables à être utilisées et (ii) la présentation d'un graphique causal montrant l'influence des paramètres indépendants sur le problème.

Les exemples fournis ont été choisis de façon arbitraire dans plusieurs domaines d'application. Toutefois, chaque méthode est applicable à tout domaines, en autant qu'il n'y ait pas de contraintes spécifiques.

5 Application dans le domaine du transport aérien

L'application particulière utilisée comme exemple dans cette section provient du domaine du transport aérien. C'est l'une des applications industrielles dans lesquelles de grandes quantités de données, sous formes numérique, symbolique et textuelle sont générées et emmagasinées pour utilisation subséquente. Le but de cette application est de découvrir des informations utiles à partir de données générées durant l'opération et les activités d'entretient d'avions commerciaux. Nous avons choisi la technique d'induction de règles comme étant la plus appropriée pour l'analyse des données de cette application. Le premier défi fut l'emmagasinage d'une quantité "de taille volumétriquement commensurable" de façon que: le contenu des bases de données soit sans parasites, pas d'information utile soit perdue, les données requises soient récupérées le plus efficacement possible et aucune données non pertinentes ou redontantes soient emmagasiner.

Cette application met en jeu les quatre défis du "data warehousing" mentionnés plus haut. Nos objectifs étaient: d'effectuer une vaste exploration des données pour trouver des explications aux problèmes identifiés par la ligne aérienne ou documentés dans une de leur bases de données; de développer une méthodologie pour appuyer le développement d'un environnement pour l'application de la technique de raisonnement "case-based" (à base de cas) contenant toute l'information reliée à l'entretien des avions qui a été effectué au cours des quelques dernières années (depuis octobre 1994). L'information sur chaque activité d'entretien inclut une description du problème, une liste de symptômes et de données générée par l'avion, et l'activité d'entretien appropriée pour résoudre le problème.

Nous avons noté plusieurs anomalies dans les données utilisées pour cette application dont: données manquantes pour certains paramètres, types de données inopportuns, nombres hors-limites, registres incomplets, données non-disponibles. Des problèmes similaires ont aussi été observés dans deux autres applications industrielles dans lesquelles nous avons utilisé les techniques d'engendrement de connaissances (usinage électrochimique et fabrication de semi-conducteur). Les données provenant d'applications réelles ne sont généralement pas parfaites.

Pour l'analyse et l'engendrement de connaissances, nous nous sommes limités sur une section particulière de la base de données contenant la description de l'exploitation technique de chaque avion et des systèmes ayant des problèmes techniques. En d'autres mots, les données étaient constituées de mesures représentant l'état de l'appareil durant différentes phases de vol, i.e au décollage, en assention, en vol stable, etc. Nous nous sommes aussi concentrés sur les messages produits automatiquement lorsqu'un problème se développe. Par exemple, lorsque le niveau de vibration de la soufflante d'un moteur dévie de ses limites acceptables, un rapport contenant les paramètres décrivant l'état de ce moteur est généré. De cette façon nous avons identifié plusieurs cas potentiels pour la recherche de données et l'engendrement de connaissances. Techniquement ces cas peuvent être classés comme étant soit *défectuosité des composantes* ou *écart de performance* dans l'opération de l'avion.

Les arbres de décision divisent les données en groupes basés sur la valeur des variables. Ils utilisent une méthode qui s'apparente au jeu des 20 questions. Le résultat est une hiérarchie d'énoncés "si-alors" qui classifie les données (BREIMAN Leo *etal*, 1984). Par example: si un client a effectué 25% moins d'appels de dépannage à chaque mois pour les six derniers mois, alors il y a 70% de probabilité que ce client a une connaissance satisfaisante du système. Il y a eu un intérêt marqué pour les logiciels basés sur la méthode des arbres de décision (Mena, 1996) principalement parce que ceux-ci sont plus rapides que les réseaux neuroniques pour résoudre plusieurs problèmes reliés aux affaires et ils sont plus faciles à comprendre. Toutefois les arbres de décision ne fournissent pas la solution à tous les problèmes. Ils peuvent avoir des difficultés à fournir une solution à partir d'un ensemble de données représentant des variables continues et demandent que les données soient regroupées en intervalles. Les intervalles choisies peuvent être telles que certaines formes soient indétectables. Par example, si une intervalle de 25 à 34 ans est choisie pour classer des groupes d'âges, il ne sera pas possible de détecter un comportement significatif relié aux personnes agées de 30 ans. Un example d'arbre de décision est présenté à la figure 3.

L'induction de règles s'apparente étroitement aux arbres de décision. Pour une meilleure compréhension, interprétation et application des résultats, les arbres de décisions sont convertis en ensemble de règles compréhensibles pour un humain. L'induction de règles est une technique courante pour la construction de bases de connaissances pour les systèmes experts (IMIELINSKI et MANNILA, 1996). Une règle type, générée par induction, consiste en deux parties: la partie gauche, appelée la condition, et la partie droite, appelée l'action. Chacune des parties d'une règle peut être composée de plusieurs composantes. L'example suivant est typique d'une règle pouvant être utilisée dans un système expert.

```
Si:    Température d'échappement des gaz > 600 C et
       temps de démarage < 20 secondes
Alors: Le problème "Hot-fast" existe et les volets
       d'entrée d'air devraient être ajustés
```

La visualisation des données a progressivement évolué de techniques pour la visualisation de données expérimentales à des représentations plus abstraites des données. Il y a deux raisons principales pour cette évolution. Premièrement, plusieurs paramètres quantitatifs importants ne sont pas mesurables directement. Deuxièmement, la représentation appropriée de données hautement multi-variées contenues dans des champs de vecteurs de paramètres, tout en évitant l'engorgement visuel, nécessite une simplification de la représentation visuelle par l'extraction et la fonte des paramètres caractéristiques. Pour une tâche d'analyse courante, on peut diviser le processus d'engendrement de connaissances en trois étapes: *le prétraitement des données, l'analyse, et le post-traitement.* La visualisation des données peut être utile à chacune des étapes. Lors du prétraitement, la visualisation peut être bénifique pour la compréhension de la nature des données, ce qui peut permettre le choix des outils d'analyse les plus appropriés et de décider de la meilleure stratégie de leur utilisation. À l'étape d'analyse, la visualisation est utilisée pour présenter les résultats. Les surfaces tri-dimentionnelles avec facade de projection sont souvent utilisées pour présenter les résultats d'induction par arbre de décision. De façon similaire, lors du post-traitement, la visualisation est utilisée pour récapituler les résultats.

e. **Le choix de la méthode d'analyse:** cette tâche est crucialle parce que les résultats générés en dépendent directement et la méthode choisie dépend aussi du but poursuivi par l'analyse. La classification et le regroupement sont deux examples de buts d'analyse.

f. **Le choix de l'outil d'analyse:** ceci consiste à choisir un ou plusieurs outils, méthodes, ou algorithmes à utiliser pour l'analyse. Cette tâche dépend de la nature des données, qui a été préalablement identifiée durant le prétraitement, et du but de l'analyse. Des études ont été effectuées pour automatiser ce processus et des boîtes à outils logiciels ont été développées permettant un choix parmi plusieurs outils d'analyse.

g. **Forage de données**: Cette étape est la principale de tout le processus et elle consiste à chercher des formes significatives parmis un ou plusieurs ensembles de données choisis pour l'analyse. Les arbres de décision, les règles de classification, ou le regroupement d'évènements sont des exemples de résultats.

h. **Interprétation des résultats:** la plupart des cas, cette étape requiert la suppression des informations non pertinentes obtenues de l'étape de forage de données. Cette tâche fait appel à la participation d'experts du domaine et peut nécessiter de revenir sur des étapes précédentes pour une analyse plus pointue. Les itérations peuvent impliquer un changement des ensembles de données principaux originalement sélectionnés pour l'analyse.

i. **Intégration des résultats:** ceci consiste à incorporer les nouvelles connaissances dans un mécanisme qui sera utiliser par plusieurs membres de l'entreprise. Ceci donne lieu à la documentation des résultats, préférablement sous forme électronique, ou à l'incorporation à une base de connaissances. Si cette connaissance est utilisé pour construire une base de connaissances, elle est combinée avec d'autre connaissances obtenues de personnes ou de documents, et souvent de résultats empiriques obtenus précédemment.

4 Les méthodes d'engendrement de connaissances

Il y a quatre méthodes principales pour l'engendrement de connaissances: *les réseaux neuroniques, les arbres de décisions, l'induction de règles* et *la visualisation des données*. Quelques outils commerciaux disponibles sont basés sur une combinaison de ces méthodes.

Les réseaux neuroniques sont essentiellement des collections de noeuds de traitement avec des entrés et sorties. Entre les niveaux visibles d'entrée et de sortie, il peut y avoir plusieurs niveaux cachés de traitement (CHENG B. et TITTERINGTON D.M. 1994). Le réseau est capable d'apprentissage; pour l'apprentissage on utilise un ensemble de données pour lesquelles les résultats (sorties) sont connus. Chaque cas de l'ensemble d'apprentissage est comparé avec le résultat connu; si le résultat du réseau est différent, une correction est calculée et appliquée aux noeuds de traitement du réseau. Ces étapes sont répétées jusqu'à ce que la condition d'arrêt d'apprentissage soit satisfaite. Les réseaux sont un processus opaque, i.e. le modèle obtenu ne se prête pas à une interprétation claire des résultats. Les réseaux neuroniques sont souvent utilisés pour la reconnaissance des formes comme l'interprétation d'électrocardiogrammes ou la reconnaissance de l'écriture (figure 2).

L'objectif de l'engendrement de connaissances est d'aider les organisations à trouver des formes et interdépendances entre paramètres qui sont dans leur données. Des algorithmes spécifiques sont utilisés pour l'extraction de résultats potentiellement utiles. En général, l'extraction de données et l'engendrement de connaissances sont différents de OLAP (on-line analytical processing). Ce sont des approches très différentes mais elles se complètent l'une l'autre.

Les requêtes d'informations et les outils traditionnels sont utilisés pour décrire et extraire l'information d'une base de données. OLAP est utilisé pour l'analyse des données en identifiant pourquoi certaines choses sont vraies. L'utilisateur forme une hypothèse à propos d'une relation et vérifie cette hypothèse à l'aide d'une série de requêtes sur les données. Par exemple, un analyste pourrait faire l'hypothèse que les personnes avec peu de revenus et ayant beaucoup de dettes constituent un risque élevé de crédit, cette hypothèse serait confirmer ou infirmer à l'aide de requêtes d'information utilisant une base d'information particulière.

3 Le processus d'engendrement de connaissances

L'engendrement de connaissances est un processus typiquement itératif qui implique la prise en considération de plusieurs critères, contraintes, et composantes. Ce processus nécessite l'implication directe d'experts du domaine qui peuvent fournir de la rétroaction à toutes les étapes du processus. La figure 1 montre les étapes préliminaires du processus. Ce diagramme n'est pas restreint à un domaine d'application particulier. Pour être efficace, le processus se doit d'être à la fois interactif et itératif. Il met en cause plusieurs étapes avec plusieurs décisions à être prises à chaque étape.

Les étapes de base de l'engendrement de connaissances sont:

a. **Compréhension du domaine d'application:** ceci implique l'acquisition et l'intégration de toutes les connaissances disponibles, de sources variées, et l'identification des buts pour l'engendrement de nouvelles connaissances.

b. **Acquérir et organiser les données nécessaires:** cette étape donne lieu à l'extraction des données de toutes les sources de données, à la sélection des données appropriées et à leur organisation pour fins d'analyse.

c. **Prétraitement des données:** ceci comprend toutes les actions requises avant de débuter le processus d'analyse. Il se matérialise par de simples tâches effectuées pour améliorer la qualité des données comme s'occuper des données manquantes ou non pertinentes, et par des tâches complexes comme la fusion de données où certaines données de sources différentes sont combinées pour n'en former qu'une. Il y a trois raisons principales pour effectuer le prétraitement des données: résoudre les problèmes reliés aux données, comprendre la nature des données et générer des connaissances utiles à partir des mêmes données. Dans la plupart des cas, la présence d'imperfections dans les données n'est pas détectée avant le début le l'analyse.

d. **Sélection et simplification des données:** ceci consiste à choisir les sous-ensembles voulus de l'ensemble des données à partir des paramètres utiles. Cette tâche est reliée au but fixé pour l'analyse afin d'engendrer de la connaissance utile.

nécessaire d'accumuler les données, de les filtrer, et de les combiner manuellement. Maintenant, dans plusieurs cas, ces tâches ont déjà été accomplies et les données sont prêtes à être analysées directement à partir des bases de données (CHARLES, 1996).

Dans cette article, nous définissons premièrement ce qu'est la génération de nouvelles informations et expliquons le processus d'engendrement de connaissances. Aux sections trois et quatre, nous fournissons un survol des techniques reliées au processus et à la section cinq nous discutons de l'application des techniques à partir d'une application réelle. Nous concluons à la section six en décrivant des défis qui devraient occuper les chercheurs pour quelques années à venir.

2 Génération d'informations nouvelles

Il y a cinq types usuels d'information qui peuvent être générés par engendrement de connaissances. Ces types sont: (i) *les associations*, (ii) *les séquences*, (iii) *les classes*, (iv) *les groupes*, et (v) *les prévisions*.

(i) **Les associations** se produisent quand les faits sont liés par un seul évènement. Par exemple, l'étude du contenu d'un panier de provisions pourait révéler que lorsque des croustilles au maïs sont achetés, 65% du temps le consommateur achète aussi du cola, à moins qu'il y ait une promotion, dans lequel cas le cola est acheté dans 85% des cas. Ces connaissances permettent au gérant d'évaluer l'efficacité de la promotion.

(ii) **Le groupement séquentiel** permet de lier les faits pour une période donnée. Suite à l'achat d'une maison, un nouvelle cuisinière sera achetée dans le mois suivant dans 45% des cas et un nouveau réfrigérateur sera acheté au cours des deux semaines suivantes dans 60% des cas.

(iii) **La classification** est l'activité la plus commune de l'extraction des données. Cette activité reconnait les formes qui décrivent le groupe de données auquel un cas ou une observation appartient. Ceci est effectué par l'examen de cas existants qui ont déjà été classifiés et desquels un ensemble de règles ont été déduites. Par exemple, la perte de clients habituels est un problème qui affecte plusieurs entreprises. La classification peut aider à la découverte d'attributs qui caractérisent le client type qui va probablement faire défection. Ces attributs peuvent constituer un modèle qui pourrait être utilisé pour identifier ces clients et déterminer le type de promotions qui est le plus efficace pour retenir la clientèle cible.

(iv) **Le regroupement** est relié à la classification mais il diffère par le fait qu'aucun groupe n'a encore été défini. En utilisant la technique de regroupement, l'outil d'extraction des données fait resortir les différents groupes de données. Ceci peut être appliqué à des problèmes aussi diversifiés que la détection de défauts pour des produits manufacturés ou la recherche de groupes distinctifs pour les cartes bancaires. Toutes ces applications peuvent être associées à des prédictions comme prédire si un client va renouveler son abonnement ou non.

(v) **Les prévisions** sont obtenus par régression. Cette activité consiste à estimer la valeur future d'une variable continue en se basant sur les formes et tendances des données.

Engendrement de connaissances à partir de données industrielles

A. (Fazel) Famili et François Dubé

Institut de technologie de l'information
Conseil national de recherches Canada
Ottawa, Canada
K1A 0R6
(famili@ai.iit.nrc.ca, francois@ai.iit.nrc.ca)

Résumé

Cette article présente une perspective de l'approche pour l'engendrement de connaissances à partir de bases de données. Tout particulièrement, nous proposons une vue d'ensemble des raisons de l'importance de la découverte de connaissances à partir de données industrielles, qui sont les bénificiaires des ces connaissances nouvellement acquises, et comment ceci peut être accompli. Les étapes requises pour l'engendrement de connaissances sont sommairements décrites et quelques méthodes sont commentées. Tout particulièrement nous expliquons le pré-traitement des données, l'utilisation des techniques d'apprentissage automatique pour l'analyse de données industrielles, et le processus de récapitulation des résultats générés par le processus d'analyse. Finallement, nous fournissons un example d'application des techniques d'engendrement de connaissances appliquées à des données provenant du dommaine de l'industrie du transport aérien. En fin d'article, nous nous permettons d'émettres des remarques stimulantes pour générer de nouvelles idées.

Mots clés

Analyse des données, engendrement de connaissances, apprentissage automatique.

1 Introduction

L'engendrement de connaissances est la base pour la conduites d'enquêtes dans plusieurs champs d'expertise, de la science au génie ou de la gestion au contrôle des processus (BRACHMAN *etal.*1996). Des données sur un sujet particulier sont acquises sous forme d'attributs symboliques ou numériques. La provenance de ces données peut être multiple, i.e. les données peuvent être fournies par un humain ou par des capteurs ayant différents degrés de complexité et de fiabilité. L'analyse de ces données fournit une meilleure compréhension du phénomène étudié. L'objectif principal de l'engendrement de connaissances est donc de produire des connaissances nouvelles et utiles pour la résolution de problèmes et la prise de meilleures décisions (GLYMOUR *etal,* 1996).

L'engendrement de connaissances utiles à partir de bases de données constitue un défi de taille, c'est un processus non trivial et dans certains cas très complexe (FOLEY, 1996). Le processus est défini comme l'extraction implicite d'information potentiellement utile et étant préalablement inconnues à partir de données (FAYYAD *et al*, 1996). La popularité du "data warehousing" et des outils associés pour le suivi, l'acquisition, l'organisation, et l'emmagasinage des données a grandement réduit les barrières à l'engendrement de connaissances. Par le passé, il était souvent

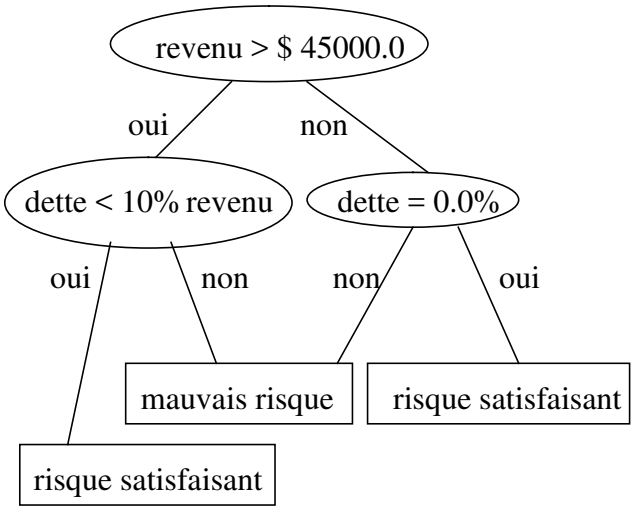

Figure 3: Example d'arbre de décision

Biographie

A. (Fazel) Famili est un agent de recherche senior à l'Institut de technologie de l'information (ITI) du Conseil national de recherches de Canada (CNRC). Ses recherches actuelles portent sur l'apprentissage automatique, l'acquisition automatique des données et l'engendrement de la connaissance. Il a obtenu sa maîtrise à l'Université de l'État d'Ohio et son doctorat à l' Université de l'État du Michigan. Il a oeuvré dans l'industrie durant trois années avant se joindre au CNRC en 1984. Il a publié plus de 20 articles dans les domaines de l'apprentissage automatique et de l'analyse des données. Il est aussi co-auteur du livre intitulée "*AI Applications in Manufacturing*" publié en 1992 par la maison d'édition AAAI/MIT.

François Dubé est agent de recherche à l'ITI/CNRC. Son domaine d'expertise est le suivi de la condition et le diagnostic de systèmes mécaniques complexes. Il est aussi intéressé par la gestion de projets impliquant le développement de nouvelles technologies. Il a obtenu une maîtrise en génie mécanique de l'Université Laval en 1984 et complète un programme de maîtrise en gestion de projets à l'Université du Québec à Hull, il s'est joint au CNRC en 1989 après une courte carrière dans les forces armées canadiennes comme officier de génie aérospatial.

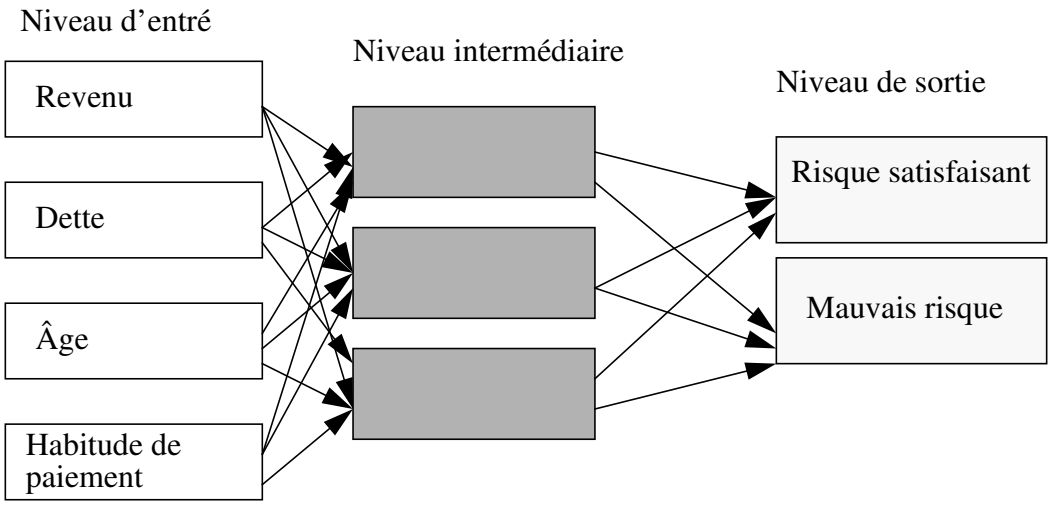

Figure 2: Example de réseau neuronique

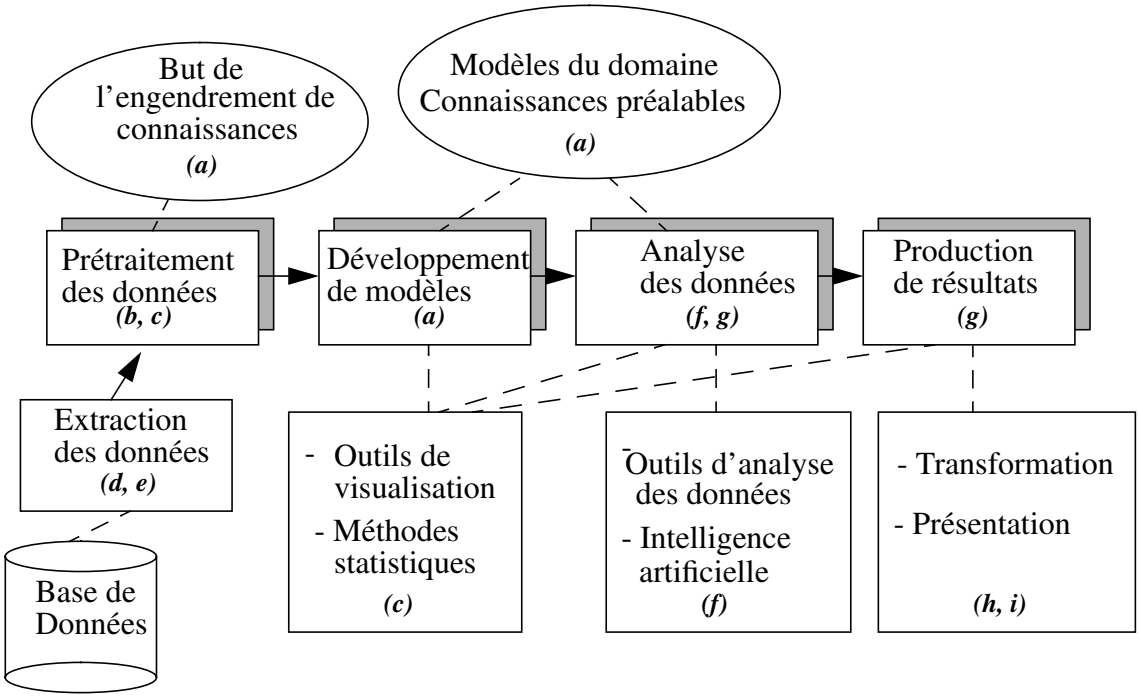

Figure 1: Le processus d'engendrement de connaissances

FOLEY J., *Data Dilema*, Information Week, June 1996, p. 14-16.

GLYMOUR Clark, *etal*, *Statistical Inference and Data Mining*, 1996, Communications of the ACM, vol. 39, no. 11, p. 35-41.

IMIELINSKI Tomasz et MANNILA Heikki, *A Database Perspective on Knowledge Discovery*, Communications of the ACM, vol. 39, no. 11, 1996, p. 58-64.

MENA Jenus, *Automatic Data Mining*, PC-AI, vol.10, no. 6, 1996, p. 16-20.

6 Conclusions et défis

Il est important de garder à l'esprit que l'engendrement de connaissances utiles n'est pas magique. L'achat du logiciel le plus dispendieux pour l'analyse de téraoctets de données sans préparation préalable ne produira propablement pas de résultats utiles. Cela prendra beaucoup de temps pour obtenir une réponse qui sera probablement sans valeur. Mais lorsque compris et appliqué convenablement, le processus d'engendrement de connaissances peut fournir des bénifices considérables à partir de l'utilisation judicieuse des données de l'industrie.

L'engendrement de connaissances est un processus complexe qui demande une implication humaine. Ce qui signifie que l'implication d'un expert humain est bénifique pour presque toutes les étapes du processus. Nous avons conclus que pour fournir un support efficient au client, le système logiciel utilisé pour l'engendrement de connaissances doit agir à titre d'assistant. Ceci principalement parce que le processus est très itératif et est complètement dépendant des buts de l'utilisateur.

Il est aussi important de noter que la plupart des clients sont indifférents au contenu (i.e. les algorithmes et méthodes) des systèmes d'engendrement de connaissances ou comment ils fonctionnent. Leurs intérêts sont plutôt: la facilité d'utilisation du système, la fiabilité des résultats produits, l'obtention de résultats faciles à interpréter et à utiliser et le moins de surcharge de travail possible généré par l'utilisation du système.

Le premier défi est donc de tenir compte des intérêts des clients mentionnés ci-haut dans toutes applications d'engendrement de connaissances. Un des défis les plus importants est le développement d'un outil qui serait capable de reconnaître la nature des données et de sélectionner un algorithme approprié et sous quelle forme pour l'engendrement de connaissances utiles. L'utilisation approprié des connaissances du domaine et l'application du processus d'engendrement de connaissances dans un environnement changeant sont aussi des défis importants.

Bibliographie

BRACHMAN Ronald. J. *etal*, "*Mining Business Databases"*, communications of the ACM, vol. 39, no. 11, 1996, pp. 42- 48.

BREIMAN Leo *etal*, "*Classification and Regression Trees"*, 1984, Wardsworth& Brooks.

CHARLES Darling, "*How to Integrate Your Data Warehouse*, Datamation", vol. 41, no. 23, 1996.

CHENG B. et TITTERINGTON D.M. "Neural Networks- a Review from a Statistical Perspective", 1994, Statistical Science, vol. 9 no 1, p. 2-30.

FAYYAD Usama, PIATETSKY-SHAPIRO Gregory, et SMYTH Padhraic, "*Advances in Knowledge Discovery and Data Mining"*, 1996, AAAI/MIT Press.

FAYYAD Usama, HAUSSLER David, et STOLORZ Paul, *Mining Scientific Data*, Communications of the ACM, vol. 39, no. 11, 1996, p. 51-57.

Pour chacun des cas potentiels identifiés, les tâches suivantes ont été exécutées:

- Faire la sélection des cas (ensemble d'observations) pour lesquels il y a une quantité suffisante de données disponibles. Pour chaque cas, un ensemble de données est composé d'une matrice de valeurs d'attribut, préférablement chaque vecteur devrait contenir au moins 100 données. Donc l'ensemble de données utilisé était constitué d'environ 100 registres de mesures représentant les différents paramètres.
- Choisir les paramètres appropriés (paramètres dépendants) pour la définition du problème (e.g. *exhaust-temp or fuel-flow* parmis les paramètres reliés au moteur). L'information de base pour l'identification des paramètres dépendants provient habituellement des experts du domaine ou de la documentation des systèmes.
- Définir les seuils de détection pour la définition du problème (e.g. un problème relié à la température des gaz d'échappement, *high-egt: exhaust-temp > 650*). L'information de base pour l'identification des seuils limites provient aussi des experts du domaine ou de la documentation des systèmes. Toutefois, des techniques de visualisation des données et des histogrammes des paramètres dépendants ont été utilisées pour la sélection et la définition des seuils appropriés.
- Utilisation d'information additionnelle, comme l'élimination de paramètres particuliers n'ayant aucune influence sur le problème. Cette tâche a pour but d'obtenir des données plus précises pour le processus d'exploration des données. Par exemple, si il y a une corrélation fortuite entre un paramètre dépendant et un paramètre indépendant qui ne peut pas techniquement avoir d'incidence sur le problème, ce paramètre indépendant devrait être masqué durant le processus d'induction.

Voivi un exemple des résultats obtenus; le résultat est sous la forme d'une règle:

```
Définition du problème: p1 est au dessus de 75
Incapacité d'utiliser 0.0% (0 de 554) registres
Le problème est présent 21.3% (118 de 554) registres
Règle 1:
     Variable 1: paramètre x >= 1.75
          r-carré 0.9
          couverture 90.7%
          taux d'erreur 2.7%
          qualité 9.2
```

Cette règle fournit quatres informations: le nom du paramètre (x) parmis tous les paramètres des données qui influencent le problème, le seuil (1.75) au dela duquel cette influence est valide, la direction de dépassement du seuil (>=), et finallement des valeurs statistiques (r-carré, couverture, taux d'erreur, qualité) qui indique la fiabilité de la règle.

- Réduction des résultats, en éliminant les informations non-pertinentes ou moins fiables. Cette tâche fait intervenir la connaissance du domaine et les informations obtenues du processus. Les résultats de l'induction de règles peuvent être présentés aux experts du domaine sous différentes formes. Nous avons présenté les résultats de deux façons différentes: (i) un résumé des règles les plus fiables à être utilisées et (ii) la présentation d'un graphique causal montrant l'influence des paramètres indépendants sur le problème.

Les exemples fournis ont été choisis de façon arbitraire dans plusieurs domaines d'application. Toutefois, chaque méthode est applicable à tout domaines, en autant qu'il n'y ait pas de contraintes spécifiques.

5 Application dans le domaine du transport aérien

L'application particulière utilisée comme exemple dans cette section provient du domaine du transport aérien. C'est l'une des applications industrielles dans lesquelles de grandes quantités de données, sous formes numérique, symbolique et textuelle sont générées et emmagasinées pour utilisation subséquente. Le but de cette application est de découvrir des informations utiles à partir de données générées durant l'opération et les activités d'entretient d'avions commerciaux. Nous avons choisi la technique d'induction de règles comme étant la plus appropriée pour l'analyse des données de cette application. Le premier défi fut l'emmagasinage d'une quantité "de taille volumétriquement commensurable" de façon que: le contenu des bases de données soit sans parasites, pas d'information utile soit perdue, les données requises soient récupérées le plus efficacement possible et aucune données non pertinentes ou redontantes soient emmagasiner.

Cette application met en jeu les quatre défis du "data warehousing" mentionnés plus haut. Nos objectifs étaient: d'effectuer une vaste exploration des données pour trouver des explications aux problèmes identifiés par la ligne aérienne ou documentés dans une de leur bases de données; de développer une méthodologie pour appuyer le développement d'un environnement pour l'application de la technique de raisonnement "case-based" (à base de cas) contenant toute l'information reliée à l'entretien des avions qui a été effectué au cours des quelques dernières années (depuis octobre 1994). L'information sur chaque activité d'entretien inclut une description du problème, une liste de symptômes et de données générée par l'avion, et l'activité d'entretien appropriée pour résoudre le problème.

Nous avons noté plusieurs anomalies dans les données utilisées pour cette application dont: données manquantes pour certains paramètres, types de données inopportuns, nombres hors-limites, registres incomplets, données non-disponibles. Des problèmes similaires ont aussi été observés dans deux autres applications industrielles dans lesquelles nous avons utilisé les techniques d'engendrement de connaissances (usinage électrochimique et fabrication de semi-conducteur). Les données provenant d'applications réelles ne sont généralement pas parfaites.

Pour l'analyse et l'engendrement de connaissances, nous nous sommes limités sur une section particulière de la base de données contenant la description de l'exploitation technique de chaque avion et des systèmes ayant des problèmes techniques. En d'autres mots, les données étaient constituées de mesures représentant l'état de l'appareil durant différentes phases de vol, i.e au décollage, en assention, en vol stable, etc. Nous nous sommes aussi concentrés sur les messages produits automatiquement lorsqu'un problème se développe. Par exemple, lorsque le niveau de vibration de la soufflante d'un moteur dévie de ses limites acceptables, un rapport contenant les paramètres décrivant l'état de ce moteur est généré. De cette façon nous avons identifié plusieurs cas potentiels pour la recherche de données et l'engendrement de connaissances. Techniquement ces cas peuvent être classés comme étant soit *défectuosité des composantes* ou *écart de performance* dans l'opération de l'avion.

Les arbres de décision divisent les données en groupes basés sur la valeur des variables. Ils utilisent une méthode qui s'apparente au jeu des 20 questions. Le résultat est une hiérarchie d'énoncés "si-alors" qui classifie les données (BREIMAN Leo *etal*, 1984). Par example: si un client a effectué 25% moins d'appels de dépannage à chaque mois pour les six derniers mois, alors il y a 70% de probabilité que ce client a une connaissance satisfaisante du système. Il y a eu un intérêt marqué pour les logiciels basés sur la méthode des arbres de décision (Mena, 1996) principalement parce que ceux-ci sont plus rapides que les réseaux neuroniques pour résoudre plusieurs problèmes reliés aux affaires et ils sont plus faciles à comprendre. Toutefois les arbres de décision ne fournissent pas la solution à tous les problèmes. Ils peuvent avoir des difficultés à fournir une solution à partir d'un ensemble de données représentant des variables continues et demandent que les données soient regroupées en intervalles. Les intervalles choisies peuvent être telles que certaines formes soient indétectables. Par example, si une intervalle de 25 à 34 ans est choisie pour classer des groupes d'âges, il ne sera pas possible de détecter un comportement significatif relié aux personnes agées de 30 ans. Un example d'arbre de décision est présenté à la figure 3.

L'induction de règles s'apparente étroitement aux arbres de décision. Pour une meilleure compréhension, interprétation et application des résultats, les arbres de décisions sont convertis en ensemble de règles compréhensibles pour un humain. L'induction de règles est une technique courante pour la construction de bases de connaissances pour les systèmes experts (IMIELINSKI et MANNILA, 1996). Une règle type, générée par induction, consiste en deux parties: la partie gauche, appelée la condition, et la partie droite, appelée l'action. Chacune des parties d'une règle peut être composée de plusieurs composantes. L'example suivant est typique d'une règle pouvant être utilisée dans un système expert.

```
Si:    Température d'échappement des gaz > 600 C et
       temps de démarage < 20 secondes
Alors: Le problème "Hot-fast" existe et les volets
       d'entrée d'air devraient être ajustés
```

La visualisation des données a progressivement évolué de techniques pour la visualisation de données expérimentales à des représentations plus abstraites des données. Il y a deux raisons principales pour cette évolution. Premièrement, plusieurs paramètres quantitatifs importants ne sont pas mesurables directement. Deuxièmement, la représentation appropriée de données hautement multi-variées contenues dans des champs de vecteurs de paramètres, tout en évitant l'engorgement visuel, nécessite une simplification de la représentation visuelle par l'extraction et la fonte des paramètres caractéristiques. Pour une tâche d'analyse courante, on peut diviser le processus d'engendrement de connaissances en trois étapes: *le prétraitement des données, l'analyse, et le post-traitement.* La visualisation des données peut être utile à chacune des étapes. Lors du prétraitement, la visualisation peut être bénifique pour la compréhension de la nature des données, ce qui peut permettre le choix des outils d'analyse les plus appropriés et de décider de la meilleure stratégie de leur utilisation. À l'étape d'analyse, la visualisation est utilisée pour présenter les résultats. Les surfaces tri-dimentionnelles avec facade de projection sont souvent utilisées pour présenter les résultats d'induction par arbre de décision. De façon similaire, lors du post-traitement, la visualisation est utilisée pour récapituler les résultats.

e. **Le choix de la méthode d'analyse:** cette tâche est crucialle parce que les résultats générés en dépendent directement et la méthode choisie dépend aussi du but poursuivi par l'analyse. La classification et le regroupement sont deux examples de buts d'analyse.

f. **Le choix de l'outil d'analyse:** ceci consiste à choisir un ou plusieurs outils, méthodes, ou algorithmes à utiliser pour l'analyse. Cette tâche dépend de la nature des données, qui a été préalablement identifiée durant le prétraitement, et du but de l'analyse. Des études ont été effectuées pour automatiser ce processus et des boîtes à outils logiciels ont été développées permettant un choix parmi plusieurs outils d'analyse.

g. **Forage de données**: Cette étape est la principale de tout le processus et elle consiste à chercher des formes significatives parmis un ou plusieurs ensembles de données choisis pour l'analyse. Les arbres de décision, les règles de classification, ou le regroupement d'évènements sont des exemples de résultats.

h. **Interprétation des résultats:** la plupart des cas, cette étape requiert la suppression des informations non pertinentes obtenues de l'étape de forage de données. Cette tâche fait appel à la participation d'experts du domaine et peut nécessiter de revenir sur des étapes précédentes pour une analyse plus pointue. Les itérations peuvent impliquer un changement des ensembles de données principaux originalement sélectionnés pour l'analyse.

i. **Intégration des résultats:** ceci consiste à incorporer les nouvelles connaissances dans un mécanisme qui sera utiliser par plusieurs membres de l'entreprise. Ceci donne lieu à la documentation des résultats, préférablement sous forme électronique, ou à l'incorporation à une base de connaissances. Si cette connaissance est utilisé pour construire une base de connaissances, elle est combinée avec d'autre connaissances obtenues de personnes ou de documents, et souvent de résultats empiriques obtenus précédemment.

4 Les méthodes d'engendrement de connaissances

Il y a quatre méthodes principales pour l'engendrement de connaissances: *les réseaux neuroniques, les arbres de décisions, l'induction de règles* et *la visualisation des données*. Quelques outils commerciaux disponibles sont basés sur une combinaison de ces méthodes.

Les réseaux neuroniques sont essentiellement des collections de noeuds de traitement avec des entrés et sorties. Entre les niveaux visibles d'entrée et de sortie, il peut y avoir plusieurs niveaux cachés de traitement (CHENG B. et TITTERINGTON D.M. 1994). Le réseau est capable d'apprentissage; pour l'apprentissage on utilise un ensemble de données pour lesquelles les résultats (sorties) sont connus. Chaque cas de l'ensemble d'apprentissage est comparé avec le résultat connu; si le résultat du réseau est différent, une correction est calculée et appliquée aux noeuds de traitement du réseau. Ces étapes sont répétées jusqu'à ce que la condition d'arrêt d'apprentissage soit satisfaite. Les réseaux sont un processus opaque, i.e. le modèle obtenu ne se prête pas à une interprétation claire des résultats. Les réseaux neuroniques sont souvent utilisés pour la reconnaissance des formes comme l'interprétation d'électrocardiogrammes ou la reconnaissance de l'écriture (figure 2).

L'objectif de l'engendrement de connaissances est d'aider les organisations à trouver des formes et interdépendances entre paramètres qui sont dans leur données. Des algorithmes spécifiques sont utilisés pour l'extraction de résultats potentiellement utiles. En général, l'extraction de données et l'engendrement de connaissances sont différents de OLAP (on-line analytical processing). Ce sont des approches très différentes mais elles se complètent l'une l'autre.

Les requêtes d'informations et les outils traditionnels sont utilisés pour décrire et extraire l'information d'une base de données. OLAP est utilisé pour l'analyse des données en identifiant pourquoi certaines choses sont vraies. L'utilisateur forme une hypothèse à propos d'une relation et vérifie cette hypothèse à l'aide d'une série de requêtes sur les données. Par exemple, un analyste pourrait faire l'hypothèse que les personnes avec peu de revenus et ayant beaucoup de dettes constituent un risque élevé de crédit, cette hypothèse serait confirmer ou infirmer à l'aide de requêtes d'information utilisant une base d'information particulière.

3 Le processus d'engendrement de connaissances

L'engendrement de connaissances est un processus typiquement itératif qui implique la prise en considération de plusieurs critères, contraintes, et composantes. Ce processus nécessite l'implication directe d'experts du domaine qui peuvent fournir de la rétroaction à toutes les étapes du processus. La figure 1 montre les étapes préliminaires du processus. Ce diagramme n'est pas restreint à un domaine d'application particulier. Pour être efficace, le processus se doit d'être à la fois interactif et itératif. Il met en cause plusieurs étapes avec plusieurs décisions à être prises à chaque étape.

Les étapes de base de l'engendrement de connaissances sont:

a. **Compréhension du domaine d'application:** ceci implique l'acquisition et l'intégration de toutes les connaissances disponibles, de sources variées, et l'identification des buts pour l'engendrement de nouvelles connaissances.

b. **Acquérir et organiser les données nécessaires:** cette étape donne lieu à l'extraction des données de toutes les sources de données, à la sélection des données appropriées et à leur organisation pour fins d'analyse.

c. **Prétraitement des données:** ceci comprend toutes les actions requises avant de débuter le processus d'analyse. Il se matérialise par de simples tâches effectuées pour améliorer la qualité des données comme s'occuper des données manquantes ou non pertinentes, et par des tâches complexes comme la fusion de données où certaines données de sources différentes sont combinées pour n'en former qu'une. Il y a trois raisons principales pour effectuer le prétraitement des données: résoudre les problèmes reliés aux données, comprendre la nature des données et générer des connaissances utiles à partir des mêmes données. Dans la plupart des cas, la présence d'imperfections dans les données n'est pas détectée avant le début le l'analyse.

d. **Sélection et simplification des données:** ceci consiste à choisir les sous-ensembles voulus de l'ensemble des données à partir des paramètres utiles. Cette tâche est reliée au but fixé pour l'analyse afin d'engendrer de la connaissance utile.

nécessaire d'accumuler les données, de les filtrer, et de les combiner manuellement. Maintenant, dans plusieurs cas, ces tâches ont déjà été accomplies et les données sont prêtes à être analysées directement à partir des bases de données (CHARLES, 1996).

Dans cette article, nous définissons premièrement ce qu'est la génération de nouvelles informations et expliquons le processus d'engendrement de connaissances. Aux sections trois et quatre, nous fournissons un survol des techniques reliées au processus et à la section cinq nous discutons de l'application des techniques à partir d'une application réelle. Nous concluons à la section six en décrivant des défis qui devraient occuper les chercheurs pour quelques années à venir.

2 Génération d'informations nouvelles

Il y a cinq types usuels d'information qui peuvent être générés par engendrement de connaissances. Ces types sont: (i) *les associations*, (ii) *les séquences*, (iii) *les classes*, (iv) *les groupes*, et (v) *les prévisions*.

(i) **Les associations** se produisent quand les faits sont liés par un seul évènement. Par exemple, l'étude du contenu d'un panier de provisions pourait révéler que lorsque des croustilles au maïs sont achetés, 65% du temps le consommateur achète aussi du cola, à moins qu'il y ait une promotion, dans lequel cas le cola est acheté dans 85% des cas. Ces connaissances permettent au gérant d'évaluer l'efficacité de la promotion.

(ii) **Le groupement séquentiel** permet de lier les faits pour une période donnée. Suite à l'achat d'une maison, un nouvelle cuisinière sera achetée dans le mois suivant dans 45% des cas et un nouveau réfrigérateur sera acheté au cours des deux semaines suivantes dans 60% des cas.

(iii) **La classification** est l'activité la plus commune de l'extraction des données. Cette activité reconnait les formes qui décrivent le groupe de données auquel un cas ou une observation appartient. Ceci est effectué par l'examen de cas existants qui ont déjà été classifiés et desquels un ensemble de règles ont été déduites. Par exemple, la perte de clients habituels est un problème qui affecte plusieurs entreprises. La classification peut aider à la découverte d'attributs qui caractérisent le client type qui va probablement faire défection. Ces attributs peuvent constituer un modèle qui pourrait être utilisé pour identifier ces clients et déterminer le type de promotions qui est le plus efficace pour retenir la clientèle cible.

(iv) **Le regroupement** est relié à la classification mais il diffère par le fait qu'aucun groupe n'a encore été défini. En utilisant la technique de regroupement, l'outil d'extraction des données fait resortir les différents groupes de données. Ceci peut être appliqué à des problèmes aussi diversifiés que la détection de défauts pour des produits manufacturés ou la recherche de groupes distinctifs pour les cartes bancaires. Toutes ces applications peuvent être associées à des prédictions comme prédire si un client va renouveler son abonnement ou non.

(v) **Les prévisions** sont obtenus par régression. Cette activité consiste à estimer la valeur future d'une variable continue en se basant sur les formes et tendances des données.

Engendrement de connaissances à partir de données industrielles

A. (Fazel) Famili et François Dubé

Institut de technologie de l'information
Conseil national de recherches Canada
Ottawa, Canada
K1A 0R6
(famili@ai.iit.nrc.ca, francois@ai.iit.nrc.ca)

Résumé

Cette article présente une perspective de l'approche pour l'engendrement de connaissances à partir de bases de données. Tout particulièrement, nous proposons une vue d'ensemble des raisons de l'importance de la découverte de connaissances à partir de données industrielles, qui sont les bénificiaires des ces connaissances nouvellement acquises, et comment ceci peut être accompli. Les étapes requises pour l'engendrement de connaissances sont sommairements décrites et quelques méthodes sont commentées. Tout particulièrement nous expliquons le pré-traitement des données, l'utilisation des techniques d'apprentissage automatique pour l'analyse de données industrielles, et le processus de récapitulation des résultats générés par le processus d'analyse. Finallement, nous fournissons un example d'application des techniques d'engendrement de connaissances appliquées à des données provenant du dommaine de l'industrie du transport aérien. En fin d'article, nous nous permettons d'émettres des remarques stimulantes pour générer de nouvelles idées.

Mots clés

Analyse des données, engendrement de connaissances, apprentissage automatique.

1 Introduction

L'engendrement de connaissances est la base pour la conduites d'enquêtes dans plusieurs champs d'expertise, de la science au génie ou de la gestion au contrôle des processus (BRACHMAN *etal.*1996). Des données sur un sujet particulier sont acquises sous forme d'attributs symboliques ou numériques. La provenance de ces données peut être multiple, i.e. les données peuvent être fournies par un humain ou par des capteurs ayant différents degrés de complexité et de fiabilité. L'analyse de ces données fournit une meilleure compréhension du phénomène étudié. L'objectif principal de l'engendrement de connaissances est donc de produire des connaissances nouvelles et utiles pour la résolution de problèmes et la prise de meilleures décisions (GLYMOUR *etal,* 1996).

L'engendrement de connaissances utiles à partir de bases de données constitue un défi de taille, c'est un processus non trivial et dans certains cas très complexe (FOLEY, 1996). Le processus est défini comme l'extraction implicite d'information potentiellement utile et étant préalablement inconnues à partir de données (FAYYAD *et al*, 1996). La popularité du "data warehousing" et des outils associés pour le suivi, l'acquisition, l'organisation, et l'emmagasinage des données a grandement réduit les barrières à l'engendrement de connaissances. Par le passé, il était souvent

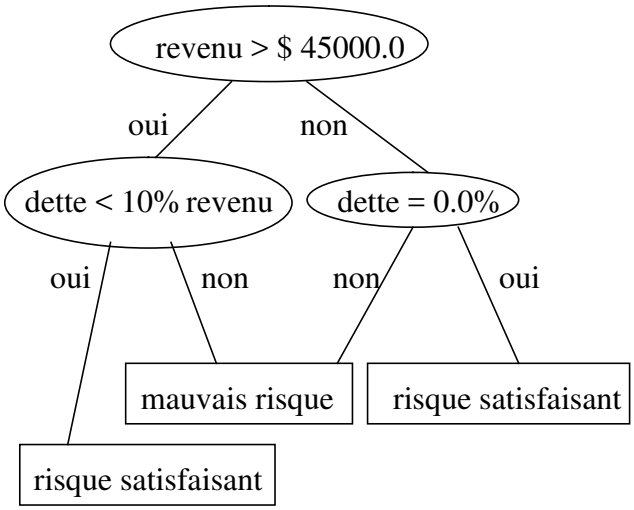

Figure 3: Example d'arbre de décision

Biographie

A. (Fazel) Famili est un agent de recherche senior à l'Institut de technologie de l'information (ITI) du Conseil national de recherches de Canada (CNRC). Ses recherches actuelles portent sur l'apprentissage automatique, l'acquisition automatique des données et l'engendrement de la connaissance. Il a obtenu sa maîtrise à l'Université de l'État d'Ohio et son doctorat à l' Université de l'État du Michigan. Il a oeuvré dans l'industrie durant trois années avant se joindre au CNRC en 1984. Il a publié plus de 20 articles dans les domaines de l'apprentissage automatique et de l'analyse des données. Il est aussi co-auteur du livre intitulée "*AI Applications in Manufacturing*" publié en 1992 par la maison d'édition AAAI/MIT.

François Dubé est agent de recherche à l'ITI/CNRC. Son domaine d'expertise est le suivi de la condition et le diagnostic de systèmes mécaniques complexes. Il est aussi intéressé par la gestion de projets impliquant le développement de nouvelles technologies. Il a obtenu une maîtrise en génie mécanique de l'Université Laval en 1984 et complète un programme de maîtrise en gestion de projets à l'Université du Québec à Hull, il s'est joint au CNRC en 1989 après une courte carrière dans les forces armées canadiennes comme officier de génie aérospatial.

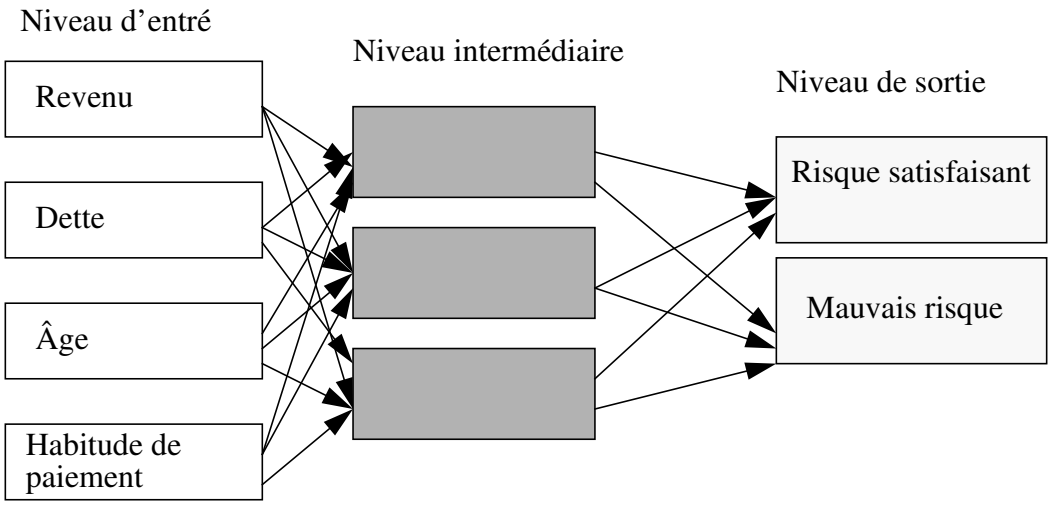

Figure 2: Example de réseau neuronique

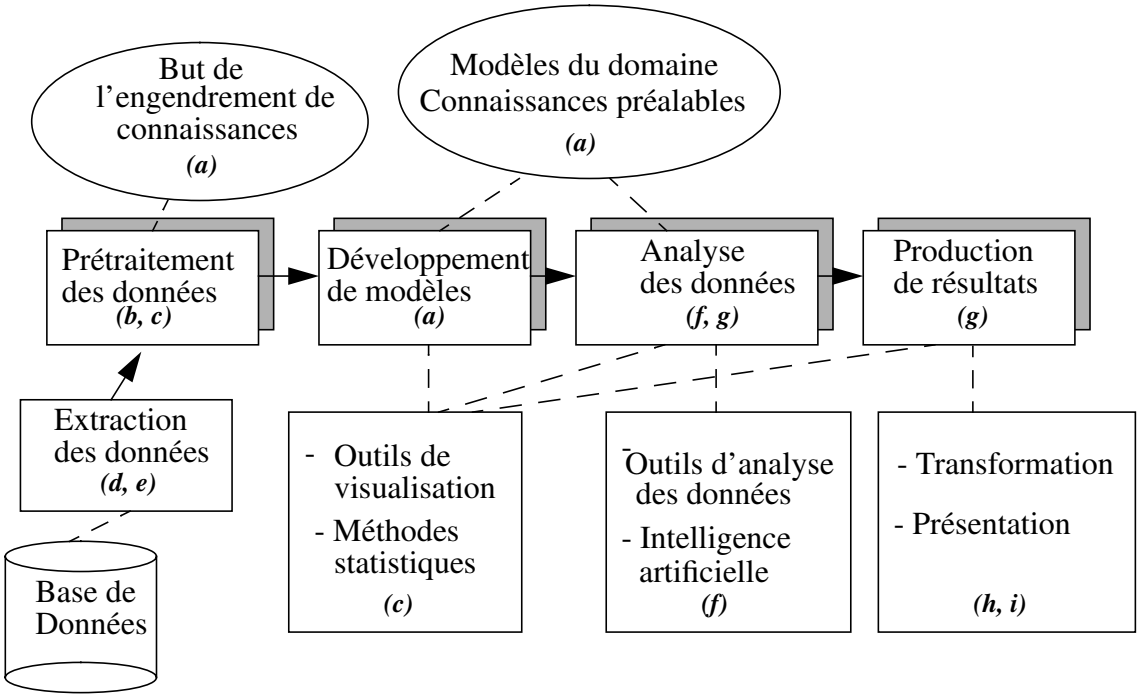

Figure 1: Le processus d'engendrement de connaissances

FOLEY J., *Data Dilema*, Information Week, June 1996, p. 14-16.

GLYMOUR Clark, *etal*, *Statistical Inference and Data Mining*, 1996, Communications of the ACM, vol. 39, no. 11, p. 35-41.

IMIELINSKI Tomasz et MANNILA Heikki, *A Database Perspective on Knowledge Discovery*, Communications of the ACM, vol. 39, no. 11, 1996, p. 58-64.

MENA Jenus, *Automatic Data Mining*, PC-AI, vol.10, no. 6, 1996, p. 16-20.

6 Conclusions et défis

Il est important de garder à l'esprit que l'engendrement de connaissances utiles n'est pas magique. L'achat du logiciel le plus dispendieux pour l'analyse de téraoctets de données sans préparation préalable ne produira propablement pas de résultats utiles. Cela prendra beaucoup de temps pour obtenir une réponse qui sera probablement sans valeur. Mais lorsque compris et appliqué convenablement, le processus d'engendrement de connaissances peut fournir des bénifices considérables à partir de l'utilisation judicieuse des données de l'industrie.

L'engendrement de connaissances est un processus complexe qui demande une implication humaine. Ce qui signifie que l'implication d'un expert humain est bénifique pour presque toutes les étapes du processus. Nous avons conclus que pour fournir un support efficient au client, le système logiciel utilisé pour l'engendrement de connaissances doit agir à titre d'assistant. Ceci principalement parce que le processus est très itératif et est complètement dépendant des buts de l'utilisateur.

Il est aussi important de noter que la plupart des clients sont indifférents au contenu (i.e. les algorithmes et méthodes) des systèmes d'engendrement de connaissances ou comment ils fonctionnent. Leurs intérêts sont plutôt: la facilité d'utilisation du système, la fiabilité des résultats produits, l'obtention de résultats faciles à interpréter et à utiliser et le moins de surcharge de travail possible généré par l'utilisation du système.

Le premier défi est donc de tenir compte des intérêts des clients mentionnés ci-haut dans toutes applications d'engendrement de connaissances. Un des défis les plus importants est le développement d'un outil qui serait capable de reconnaître la nature des données et de sélectionner un algorithme approprié et sous quelle forme pour l'engendrement de connaissances utiles. L'utilisation approprié des connaissances du domaine et l'application du processus d'engendrement de connaissances dans un environnement changeant sont aussi des défis importants.

Bibliographie

BRACHMAN Ronald. J. *etal*, "*Mining Business Databases"*, communications of the ACM, vol. 39, no. 11, 1996, pp. 42- 48.

BREIMAN Leo *etal*, "*Classification and Regression Trees"*, 1984, Wardsworth& Brooks.

CHARLES Darling, "*How to Integrate Your Data Warehouse*, Datamation", vol. 41, no. 23, 1996.

CHENG B. et TITTERINGTON D.M. "Neural Networks- a Review from a Statistical Perspective", 1994, Statistical Science, vol. 9 no 1, p. 2-30.

FAYYAD Usama, PIATETSKY-SHAPIRO Gregory, et SMYTH Padhraic, "*Advances in Knowledge Discovery and Data Mining"*, 1996, AAAI/MIT Press.

FAYYAD Usama, HAUSSLER David, et STOLORZ Paul, *Mining Scientific Data*, Communications of the ACM, vol. 39, no. 11, 1996, p. 51-57.

Pour chacun des cas potentiels identifiés, les tâches suivantes ont été exécutées:

- Faire la sélection des cas (ensemble d'observations) pour lesquels il y a une quantité suffisante de données disponibles. Pour chaque cas, un ensemble de données est composé d'une matrice de valeurs d'attribut, préférablement chaque vecteur devrait contenir au moins 100 données. Donc l'ensemble de données utilisé était constitué d'environ 100 registres de mesures représentant les différents paramètres.
- Choisir les paramètres appropriés (paramètres dépendants) pour la définition du problème (e.g. *exhaust-temp or fuel-flow* parmis les paramètres reliés au moteur). L'information de base pour l'identification des paramètres dépendants provient habituellement des experts du domaine ou de la documentation des systèmes.
- Définir les seuils de détection pour la définition du problème (e.g. un problème relié à la température des gaz d'échappement, *high-egt: exhaust-temp > 650*). L'information de base pour l'identification des seuils limites provient aussi des experts du domaine ou de la documentation des systèmes. Toutefois, des techniques de visualisation des données et des histogrammes des paramètres dépendants ont été utilisées pour la sélection et la définition des seuils appropriés.
- Utilisation d'information additionnelle, comme l'élimination de paramètres particuliers n'ayant aucune influence sur le problème. Cette tâche a pour but d'obtenir des données plus précises pour le processus d'exploration des données. Par exemple, si il y a une corrélation fortuite entre un paramètre dépendant et un paramètre indépendant qui ne peut pas techniquement avoir d'incidence sur le problème, ce paramètre indépendant devrait être masqué durant le processus d'induction.

Voivi un exemple des résultats obtenus; le résultat est sous la forme d'une règle:

```
Définition du problème: p1 est au dessus de 75
Incapacité d'utiliser 0.0% (0 de 554) registres
Le problème est présent 21.3% (118 de 554) registres
Règle 1:
     Variable 1: paramètre x >= 1.75
          r-carré 0.9
          couverture 90.7%
          taux d'erreur 2.7%
          qualité 9.2
```

Cette règle fournit quatres informations: le nom du paramètre (x) parmis tous les paramètres des données qui influencent le problème, le seuil (1.75) au dela duquel cette influence est valide, la direction de dépassement du seuil (>=), et finallement des valeurs statistiques (r-carré, couverture, taux d'erreur, qualité) qui indique la fiabilité de la règle.

- Réduction des résultats, en éliminant les informations non-pertinentes ou moins fiables. Cette tâche fait intervenir la connaissance du domaine et les informations obtenues du processus. Les résultats de l'induction de règles peuvent être présentés aux experts du domaine sous différentes formes. Nous avons présenté les résultats de deux façons différentes: (i) un résumé des règles les plus fiables à être utilisées et (ii) la présentation d'un graphique causal montrant l'influence des paramètres indépendants sur le problème.

Les exemples fournis ont été choisis de façon arbitraire dans plusieurs domaines d'application. Toutefois, chaque méthode est applicable à tout domaines, en autant qu'il n'y ait pas de contraintes spécifiques.

5 Application dans le domaine du transport aérien

L'application particulière utilisée comme exemple dans cette section provient du domaine du transport aérien. C'est l'une des applications industrielles dans lesquelles de grandes quantités de données, sous formes numérique, symbolique et textuelle sont générées et emmagasinées pour utilisation subséquente. Le but de cette application est de découvrir des informations utiles à partir de données générées durant l'opération et les activités d'entretient d'avions commerciaux. Nous avons choisi la technique d'induction de règles comme étant la plus appropriée pour l'analyse des données de cette application. Le premier défi fut l'emmagasinage d'une quantité "de taille volumétriquement commensurable" de façon que: le contenu des bases de données soit sans parasites, pas d'information utile soit perdue, les données requises soient récupérées le plus efficacement possible et aucune données non pertinentes ou redontantes soient emmagasiner.

Cette application met en jeu les quatre défis du "data warehousing" mentionnés plus haut. Nos objectifs étaient: d'effectuer une vaste exploration des données pour trouver des explications aux problèmes identifiés par la ligne aérienne ou documentés dans une de leur bases de données; de développer une méthodologie pour appuyer le développement d'un environnement pour l'application de la technique de raisonnement "case-based" (à base de cas) contenant toute l'information reliée à l'entretien des avions qui a été effectué au cours des quelques dernières années (depuis octobre 1994). L'information sur chaque activité d'entretien inclut une description du problème, une liste de symptômes et de données générée par l'avion, et l'activité d'entretien appropriée pour résoudre le problème.

Nous avons noté plusieurs anomalies dans les données utilisées pour cette application dont: données manquantes pour certains paramètres, types de données inopportuns, nombres hors-limites, registres incomplets, données non-disponibles. Des problèmes similaires ont aussi été observés dans deux autres applications industrielles dans lesquelles nous avons utilisé les techniques d'engendrement de connaissances (usinage électrochimique et fabrication de semi-conducteur). Les données provenant d'applications réelles ne sont généralement pas parfaites.

Pour l'analyse et l'engendrement de connaissances, nous nous sommes limités sur une section particulière de la base de données contenant la description de l'exploitation technique de chaque avion et des systèmes ayant des problèmes techniques. En d'autres mots, les données étaient constituées de mesures représentant l'état de l'appareil durant différentes phases de vol, i.e au décollage, en assention, en vol stable, etc. Nous nous sommes aussi concentrés sur les messages produits automatiquement lorsqu'un problème se développe. Par exemple, lorsque le niveau de vibration de la soufflante d'un moteur dévie de ses limites acceptables, un rapport contenant les paramètres décrivant l'état de ce moteur est généré. De cette façon nous avons identifié plusieurs cas potentiels pour la recherche de données et l'engendrement de connaissances. Techniquement ces cas peuvent être classés comme étant soit *défectuosité des composantes* ou *écart de performance* dans l'opération de l'avion.

Les arbres de décision divisent les données en groupes basés sur la valeur des variables. Ils utilisent une méthode qui s'apparente au jeu des 20 questions. Le résultat est une hiérarchie d'énoncés "si-alors" qui classifie les données (BREIMAN Leo *etal*, 1984). Par example: si un client a effectué 25% moins d'appels de dépannage à chaque mois pour les six derniers mois, alors il y a 70% de probabilité que ce client a une connaissance satisfaisante du système. Il y a eu un intérêt marqué pour les logiciels basés sur la méthode des arbres de décision (Mena, 1996) principalement parce que ceux-ci sont plus rapides que les réseaux neuroniques pour résoudre plusieurs problèmes reliés aux affaires et ils sont plus faciles à comprendre. Toutefois les arbres de décision ne fournissent pas la solution à tous les problèmes. Ils peuvent avoir des difficultés à fournir une solution à partir d'un ensemble de données représentant des variables continues et demandent que les données soient regroupées en intervalles. Les intervalles choisies peuvent être telles que certaines formes soient indétectables. Par example, si une intervalle de 25 à 34 ans est choisie pour classer des groupes d'âges, il ne sera pas possible de détecter un comportement significatif relié aux personnes agées de 30 ans. Un example d'arbre de décision est présenté à la figure 3.

L'induction de règles s'apparente étroitement aux arbres de décision. Pour une meilleure compréhension, interprétation et application des résultats, les arbres de décisions sont convertis en ensemble de règles compréhensibles pour un humain. L'induction de règles est une technique courante pour la construction de bases de connaissances pour les systèmes experts (IMIELINSKI et MANNILA, 1996). Une règle type, générée par induction, consiste en deux parties: la partie gauche, appelée la condition, et la partie droite, appelée l'action. Chacune des parties d'une règle peut être composée de plusieurs composantes. L'example suivant est typique d'une règle pouvant être utilisée dans un système expert.

```
Si:    Température d'échappement des gaz > 600 C et
       temps de démarage < 20 secondes
Alors: Le problème "Hot-fast" existe et les volets
       d'entrée d'air devraient être ajustés
```

La visualisation des données a progressivement évolué de techniques pour la visualisation de données expérimentales à des représentations plus abstraites des données. Il y a deux raisons principales pour cette évolution. Premièrement, plusieurs paramètres quantitatifs importants ne sont pas mesurables directement. Deuxièmement, la représentation appropriée de données hautement multi-variées contenues dans des champs de vecteurs de paramètres, tout en évitant l'engorgement visuel, nécessite une simplification de la représentation visuelle par l'extraction et la fonte des paramètres caractéristiques. Pour une tâche d'analyse courante, on peut diviser le processus d'engendrement de connaissances en trois étapes: *le prétraitement des données, l'analyse, et le post-traitement.* La visualisation des données peut être utile à chacune des étapes. Lors du prétraitement, la visualisation peut être bénifique pour la compréhension de la nature des données, ce qui peut permettre le choix des outils d'analyse les plus appropriés et de décider de la meilleure stratégie de leur utilisation. À l'étape d'analyse, la visualisation est utilisée pour présenter les résultats. Les surfaces tri-dimentionnelles avec facade de projection sont souvent utilisées pour présenter les résultats d'induction par arbre de décision. De façon similaire, lors du post-traitement, la visualisation est utilisée pour récapituler les résultats.

e. **Le choix de la méthode d'analyse:** cette tâche est crucialle parce que les résultats générés en dépendent directement et la méthode choisie dépend aussi du but poursuivi par l'analyse. La classification et le regroupement sont deux examples de buts d'analyse.

f. **Le choix de l'outil d'analyse:** ceci consiste à choisir un ou plusieurs outils, méthodes, ou algorithmes à utiliser pour l'analyse. Cette tâche dépend de la nature des données, qui a été préalablement identifiée durant le prétraitement, et du but de l'analyse. Des études ont été effectuées pour automatiser ce processus et des boîtes à outils logiciels ont été développées permettant un choix parmi plusieurs outils d'analyse.

g. **Forage de données**: Cette étape est la principale de tout le processus et elle consiste à chercher des formes significatives parmis un ou plusieurs ensembles de données choisis pour l'analyse. Les arbres de décision, les règles de classification, ou le regroupement d'évènements sont des exemples de résultats.

h. **Interprétation des résultats:** la plupart des cas, cette étape requiert la suppression des informations non pertinentes obtenues de l'étape de forage de données. Cette tâche fait appel à la participation d'experts du domaine et peut nécessiter de revenir sur des étapes précédentes pour une analyse plus pointue. Les itérations peuvent impliquer un changement des ensembles de données principaux originalement sélectionnés pour l'analyse.

i. **Intégration des résultats:** ceci consiste à incorporer les nouvelles connaissances dans un mécanisme qui sera utiliser par plusieurs membres de l'entreprise. Ceci donne lieu à la documentation des résultats, préférablement sous forme électronique, ou à l'incorporation à une base de connaissances. Si cette connaissance est utilisé pour construire une base de connaissances, elle est combinée avec d'autre connaissances obtenues de personnes ou de documents, et souvent de résultats empiriques obtenus précédemment.

4 Les méthodes d'engendrement de connaissances

Il y a quatre méthodes principales pour l'engendrement de connaissances: *les réseaux neuroniques, les arbres de décisions, l'induction de règles* et *la visualisation des données*. Quelques outils commerciaux disponibles sont basés sur une combinaison de ces méthodes.

Les réseaux neuroniques sont essentiellement des collections de noeuds de traitement avec des entrés et sorties. Entre les niveaux visibles d'entrée et de sortie, il peut y avoir plusieurs niveaux cachés de traitement (CHENG B. et TITTERINGTON D.M. 1994). Le réseau est capable d'apprentissage; pour l'apprentissage on utilise un ensemble de données pour lesquelles les résultats (sorties) sont connus. Chaque cas de l'ensemble d'apprentissage est comparé avec le résultat connu; si le résultat du réseau est différent, une correction est calculée et appliquée aux noeuds de traitement du réseau. Ces étapes sont répétées jusqu'à ce que la condition d'arrêt d'apprentissage soit satisfaite. Les réseaux sont un processus opaque, i.e. le modèle obtenu ne se prête pas à une interprétation claire des résultats. Les réseaux neuroniques sont souvent utilisés pour la reconnaissance des formes comme l'interprétation d'électrocardiogrammes ou la reconnaissance de l'écriture (figure 2).

L'objectif de l'engendrement de connaissances est d'aider les organisations à trouver des formes et interdépendances entre paramètres qui sont dans leur données. Des algorithmes spécifiques sont utilisés pour l'extraction de résultats potentiellement utiles. En général, l'extraction de données et l'engendrement de connaissances sont différents de OLAP (on-line analytical processing). Ce sont des approches très différentes mais elles se complètent l'une l'autre.

Les requêtes d'informations et les outils traditionnels sont utilisés pour décrire et extraire l'information d'une base de données. OLAP est utilisé pour l'analyse des données en identifiant pourquoi certaines choses sont vraies. L'utilisateur forme une hypothèse à propos d'une relation et vérifie cette hypothèse à l'aide d'une série de requêtes sur les données. Par exemple, un analyste pourrait faire l'hypothèse que les personnes avec peu de revenus et ayant beaucoup de dettes constituent un risque élevé de crédit, cette hypothèse serait confirmer ou infirmer à l'aide de requêtes d'information utilisant une base d'information particulière.

3 Le processus d'engendrement de connaissances

L'engendrement de connaissances est un processus typiquement itératif qui implique la prise en considération de plusieurs critères, contraintes, et composantes. Ce processus nécessite l'implication directe d'experts du domaine qui peuvent fournir de la rétroaction à toutes les étapes du processus. La figure 1 montre les étapes préliminaires du processus. Ce diagramme n'est pas restreint à un domaine d'application particulier. Pour être efficace, le processus se doit d'être à la fois interactif et itératif. Il met en cause plusieurs étapes avec plusieurs décisions à être prises à chaque étape.

Les étapes de base de l'engendrement de connaissances sont:

a. **Compréhension du domaine d'application:** ceci implique l'acquisition et l'intégration de toutes les connaissances disponibles, de sources variées, et l'identification des buts pour l'engendrement de nouvelles connaissances.

b. **Acquérir et organiser les données nécessaires:** cette étape donne lieu à l'extraction des données de toutes les sources de données, à la sélection des données appropriées et à leur organisation pour fins d'analyse.

c. **Prétraitement des données:** ceci comprend toutes les actions requises avant de débuter le processus d'analyse. Il se matérialise par de simples tâches effectuées pour améliorer la qualité des données comme s'occuper des données manquantes ou non pertinentes, et par des tâches complexes comme la fusion de données où certaines données de sources différentes sont combinées pour n'en former qu'une. Il y a trois raisons principales pour effectuer le prétraitement des données: résoudre les problèmes reliés aux données, comprendre la nature des données et générer des connaissances utiles à partir des mêmes données. Dans la plupart des cas, la présence d'imperfections dans les données n'est pas détectée avant le début le l'analyse.

d. **Sélection et simplification des données:** ceci consiste à choisir les sous-ensembles voulus de l'ensemble des données à partir des paramètres utiles. Cette tâche est reliée au but fixé pour l'analyse afin d'engendrer de la connaissance utile.

nécessaire d'accumuler les données, de les filtrer, et de les combiner manuellement. Maintenant, dans plusieurs cas, ces tâches ont déjà été accomplies et les données sont prêtes à être analysées directement à partir des bases de données (CHARLES, 1996).

Dans cette article, nous définissons premièrement ce qu'est la génération de nouvelles informations et expliquons le processus d'engendrement de connaissances. Aux sections trois et quatre, nous fournissons un survol des techniques reliées au processus et à la section cinq nous discutons de l'application des techniques à partir d'une application réelle. Nous concluons à la section six en décrivant des défis qui devraient occuper les chercheurs pour quelques années à venir.

2 Génération d'informations nouvelles

Il y a cinq types usuels d'information qui peuvent être générés par engendrement de connaissances. Ces types sont: (i) *les associations*, (ii) *les séquences*, (iii) *les classes*, (iv) *les groupes*, et (v) *les prévisions*.

(i) **Les associations** se produisent quand les faits sont liés par un seul évènement. Par exemple, l'étude du contenu d'un panier de provisions pourait révéler que lorsque des croustilles au maïs sont achetés, 65% du temps le consommateur achète aussi du cola, à moins qu'il y ait une promotion, dans lequel cas le cola est acheté dans 85% des cas. Ces connaissances permettent au gérant d'évaluer l'efficacité de la promotion.

(ii) **Le groupement séquentiel** permet de lier les faits pour une période donnée. Suite à l'achat d'une maison, un nouvelle cuisinière sera achetée dans le mois suivant dans 45% des cas et un nouveau réfrigérateur sera acheté au cours des deux semaines suivantes dans 60% des cas.

(iii) **La classification** est l'activité la plus commune de l'extraction des données. Cette activité reconnait les formes qui décrivent le groupe de données auquel un cas ou une observation appartient. Ceci est effectué par l'examen de cas existants qui ont déjà été classifiés et desquels un ensemble de règles ont été déduites. Par exemple, la perte de clients habituels est un problème qui affecte plusieurs entreprises. La classification peut aider à la découverte d'attributs qui caractérisent le client type qui va probablement faire défection. Ces attributs peuvent constituer un modèle qui pourrait être utilisé pour identifier ces clients et déterminer le type de promotions qui est le plus efficace pour retenir la clientèle cible.

(iv) **Le regroupement** est relié à la classification mais il diffère par le fait qu'aucun groupe n'a encore été défini. En utilisant la technique de regroupement, l'outil d'extraction des données fait resortir les différents groupes de données. Ceci peut être appliqué à des problèmes aussi diversifiés que la détection de défauts pour des produits manufacturés ou la recherche de groupes distinctifs pour les cartes bancaires. Toutes ces applications peuvent être associées à des prédictions comme prédire si un client va renouveler son abonnement ou non.

(v) **Les prévisions** sont obtenus par régression. Cette activité consiste à estimer la valeur future d'une variable continue en se basant sur les formes et tendances des données.

Engendrement de connaissances à partir de données industrielles

A. (Fazel) Famili et François Dubé

Institut de technologie de l'information
Conseil national de recherches Canada
Ottawa, Canada
K1A 0R6
(famili@ai.iit.nrc.ca, francois@ai.iit.nrc.ca)

Résumé

Cette article présente une perspective de l'approche pour l'engendrement de connaissances à partir de bases de données. Tout particulièrement, nous proposons une vue d'ensemble des raisons de l'importance de la découverte de connaissances à partir de données industrielles, qui sont les bénificiaires des ces connaissances nouvellement acquises, et comment ceci peut être accompli. Les étapes requises pour l'engendrement de connaissances sont sommairements décrites et quelques méthodes sont commentées. Tout particulièrement nous expliquons le pré-traitement des données, l'utilisation des techniques d'apprentissage automatique pour l'analyse de données industrielles, et le processus de récapitulation des résultats générés par le processus d'analyse. Finallement, nous fournissons un example d'application des techniques d'engendrement de connaissances appliquées à des données provenant du dommaine de l'industrie du transport aérien. En fin d'article, nous nous permettons d'émettres des remarques stimulantes pour générer de nouvelles idées.

Mots clés

Analyse des données, engendrement de connaissances, apprentissage automatique.

1 Introduction

L'engendrement de connaissances est la base pour la conduites d'enquêtes dans plusieurs champs d'expertise, de la science au génie ou de la gestion au contrôle des processus (BRACHMAN *etal.*1996). Des données sur un sujet particulier sont acquises sous forme d'attributs symboliques ou numériques. La provenance de ces données peut être multiple, i.e. les données peuvent être fournies par un humain ou par des capteurs ayant différents degrés de complexité et de fiabilité. L'analyse de ces données fournit une meilleure compréhension du phénomène étudié. L'objectif principal de l'engendrement de connaissances est donc de produire des connaissances nouvelles et utiles pour la résolution de problèmes et la prise de meilleures décisions (GLYMOUR *etal,* 1996).

L'engendrement de connaissances utiles à partir de bases de données constitue un défi de taille, c'est un processus non trivial et dans certains cas très complexe (FOLEY, 1996). Le processus est défini comme l'extraction implicite d'information potentiellement utile et étant préalablement inconnues à partir de données (FAYYAD *et al*, 1996). La popularité du "data warehousing" et des outils associés pour le suivi, l'acquisition, l'organisation, et l'emmagasinage des données a grandement réduit les barrières à l'engendrement de connaissances. Par le passé, il était souvent