EXPLORATION DES BASES DE DONNÉES INDUSTRIELLES À L'AIDE DU DATA MINING – PERSPECTIVES

Bruno Agard (1), Andrew Kusiak (2)
(1) Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal,
C.P. 6079, succ. Centre-ville, Montréal (Québec), H3C3A7, Canada
bruno.agard@polymtl.ca
(2) Intelligent Systems Laboratory, 2139 Seamans Center,
Department of Mechanical and Industrial Engineering, The University of Iowa, IA 52242 - 1527, USA
andrew-kusiak@uiowa.edu

Résumé:

Les industriels d'aujourd'hui sont de plus en plus équipés de systèmes d'acquisition et de décision numériques. Ces systèmes, qu'ils agissent au niveau de la conception des produits, de la gestion ou du suivi de la production génèrent des Giga Octets de données chaque jour. Le volume de données est tel qu'il est désormais impossible à un décideur d'avoir une vue d'ensemble sur ces données. Il est alors nécessaire d'équiper les décideurs d'outils performants leur permettant d'extraire une information pertinente de ces bases de données. Les techniques de data mining sont dédiées à la découverte de modèles non triviaux, implicites, non connus, potentiellement utiles et compréhensibles à partir d'un grand ensemble de données. Nous décrirons tout d'abord en quoi consistent ces différentes techniques. Nous verrons un processus d'extraction d'information adapté à l'ingénierie. Des résultats d'application du data mining sur des problèmes industriels ainsi que des perspectives d'utilisation de ces techniques dans divers champs de l'ingénierie seront montrés.

Mots clés : données industrielles, data mining, perspectives

1 Introduction

Les industriels d'aujourd'hui sont de plus en plus équipés de systèmes d'acquisition et de décision numériques. Il s'agit d'outils de conception des produits, de systèmes de planification de la production, d'outils de fabrication, de gestion de la qualité et de gestion du stock. Ces systèmes, qu'ils agissent au niveau de la conception des produits, de la gestion ou du suivi de la production génèrent des Giga Octets de données chaque jour (Figure 1).

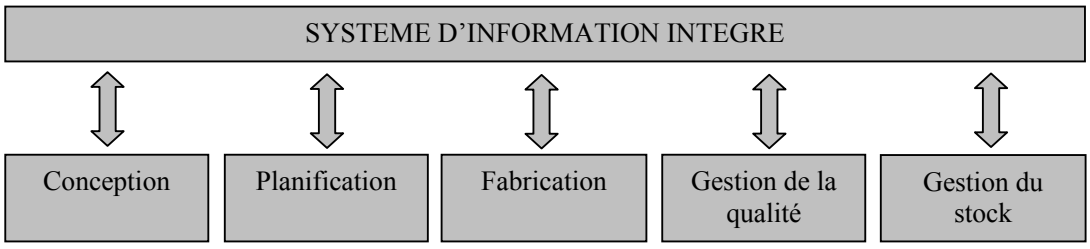

Figure 1: Origines multiples des données dans un système d'information intégré [1].

Le système d'information est alimenté par une quantité sans cesse croissante de données de différentes natures. Il est estimé que le volume d'information a augmenté de 30% par an chaque année entre 1999 et 2002 [2]. Cette masse de données est facilement stockée grâce à l'évolution exponentielle des capacités des systèmes de stockage. Cependant le problème de traitement de ces données est souvent retardé. Les données produites sont le plus souvent simplement archivées dans des buts de traçabilité et de gestion de la qualité. Cet archivage provient tout à la fois de contraintes internes de réutilisabilité ou de contraintes externes de suivi de normes.

Edmunds et Morris [3] déclarent que les systèmes qui ont été jusqu'à maintenant inventés produisent, manipulent et disséminent l'information bien plus vite qu'ils ne sont capables de la traiter. Sur ce même sujet, Shenk [4] parle de brouillard généré par les données. La trop grande quantité d'information cache l'information essentielle tandis que la précision des informations disponibles n'est pas toujours déterminée. Il peut y avoir un écart important entre la réalité et l'information disponible.

Le volume de données est tel qu'il est désormais impossible d'avoir une vue complète sur les informations contenues dans ces données. Les décideurs ont néanmoins besoin d'avoir une vision globale des informations sur les entreprises qu'ils gèrent. Il est alors nécessaire d'équiper les décideurs d'outils performants leur permettant d'extraire une information pertinente de ces masses de données.

Les techniques de data mining ("fouille de données") [5] sont dédiées à l'extraction d'information à partir d'une masse de données. Ils seront les outils considérés dans cette communication. Nous décrirons tout d'abord en quoi consistent ces différentes techniques (Section 2). Nous verrons un processus d'extraction d'information adapté à l'ingénierie (Section 3). La section 4 présentera des perspectives d'utilisation de ces techniques dans divers champs de l'ingénierie. La section 5 conclura ces propos.

2 Le data mining

2.1 Définition

D'après Anand et Buchner [6] le data mining offre des algorithmes et des outils pour la découverte de modèles non triviaux, implicites, non connus, potentiellement utiles et compréhensibles à partir d'une grande masse de données.

Le data mining n'est pas un nouvel outil magique qui résoudrait tous les problèmes d'extraction de l'information qui seraient soudainement apparus. Le data mining s'appuie à la fois sur les techniques statistiques, les réseaux de neurones, les techniques de visualisation… ainsi que sur des techniques spécialement développées pour parcourir d'immenses bases de données à la recherche de patrons fréquents. Le data mining se base sur des données désagrégées plutôt que sur des caractéristiques de population.

Le section suivante introduit les techniques de data mining.

2.2 Un ensemble de techniques complémentaires

Le data mining est en fait un ensemble de techniques complémentaires dédiées à différentes tâches.

D'après Westphal et Blaxton [7], les tâches du data mining se partagent entre la classification, l'estimation, la segmentation et la description (Figure 2).

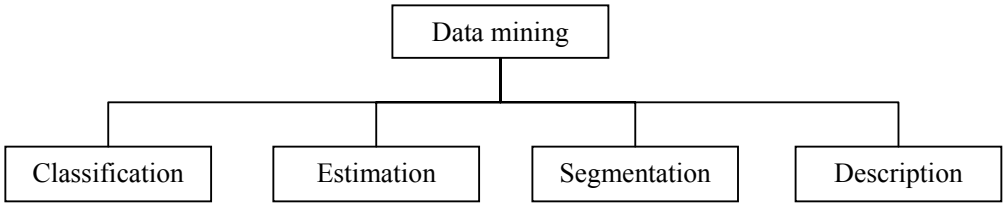

Figure 2: Les tâches du data mining.

Apté [8] propose une classification des algorithmes du data mining en trois catégories : prédiction, segmentation et recherche de patrons fréquents (Figure 3).

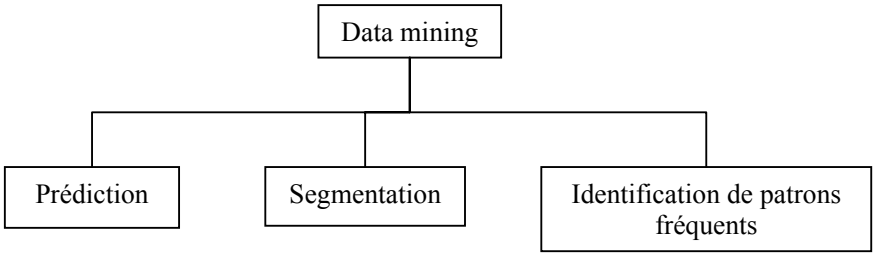

Figure 3: Les algorithmes du data mining.

Ces représentations se recoupent. Ceci permet d'arriver à la description suivante :

Les **modèles prédictifs** ont pour but de déterminer une fonction (ou un modèle) qui associe des entrées et des sorties. Les réseaux de neurone, les régressions, les arbres de décision et les règles de décision sont les méthodes les plus utilisées. Ces méthodes sont des méthodes supervisées en ce sens qu'il faut préciser quelles sont les entrées et quelle est la sortie à prédire. En fonction de la nature de la variable de sortie, deux sous catégories d'outils coexistent :

- **Si la variable de sortie est de type discret**, la **classification** aura pour rôle de construire le modèle qui permettra de classifier correctement les enregistrements, c'est-à-dire d'assigner des catégories prédéfinies aux données.

- **Si la variable de sortie est de type continu**, l'**estimation** consistera à compléter une valeur manquante dans un champ particulier en fonction des autres champs de l'enregistrement. Les outils statistiques usuels de régression sont les plus employés. Les réseaux de neurones sont aussi souvent employés dans cette activité.

La **segmentation** est un apprentissage non supervisé (on ne définit pas ce qui est "entrées" et ce qui est "sorties") qui vise à identifier des ensembles d'éléments qui partagent certaines similarités. Les algorithmes de segmentation maximisent l'homogénéité à l'intérieur de chaque ensemble et maximisent l'hétérogénéité entre les ensembles. Différentes méthodes sont utilisées pour définir ces groupes : k-médian, algorithmes hiérarchiques, réseaux de neurones…

La **description** (ou **identification de patrons fréquents**) consiste à expliquer les relations existantes dans les données. Les méthodes cherchent à identifier les associations entre des variables. L'analyse des liens et les techniques de visualisations sont couramment utilisées dans ce but. Les techniques de visualisation sont utilisées pour simplifier la compréhension des données à l'aide de représentations graphiques adaptées. Les règles d'associations représentent des combinaisons de variables avec des niveaux prédéfinis de régularité. Par exemple une règle $A \Rightarrow B$ peut être restituée avec deux indicateurs de mesure : le support et la confiance. Le support caractérise le nombre de fois

où A est présent dans l'ensemble des données. La confiance identifie la proportion de fois où B existe quand A est présent. Ces indicateurs permettent de montrer à quel niveau une règle permet de représenter les données considérées. Plus ces indicateurs sont élevés, plus la règle est "forte", elle est alors indiquée à l'utilisateur qui jugera de sa pertinence par rapport au problème posé.

3 Le processus d'extraction d'information

Le processus général d'extraction d'information à l'aide du data mining (Figure 4) est proposé dans Fayyad *et al.* [5].

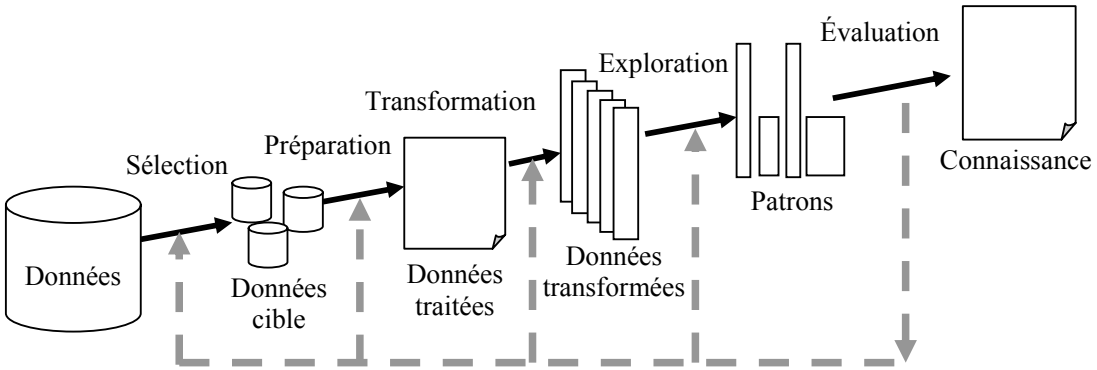

Figure 4 : Processus général du data mining [5].

Ce processus comprend des étapes de définition du problème (définition du domaine, but de l'utilisateur final), de préparation des données (sélection, préparation, transformation), de fouille de données (sélection des outils de data mining appropriés, recherche des patrons) et d'évaluation des résultats pour aboutir aux nouvelles connaissances. Le processus présenté est itératif et plusieurs retours en arrière dans les différentes étapes sont nécessaires pour affiner les résultats.

Ce processus générique est repris par Büchner *et al.* [9] qui le spécialisent au niveau de la résolution des problèmes industriels (Figure 5). Les différentes étapes de résolution sont détaillées ci après.

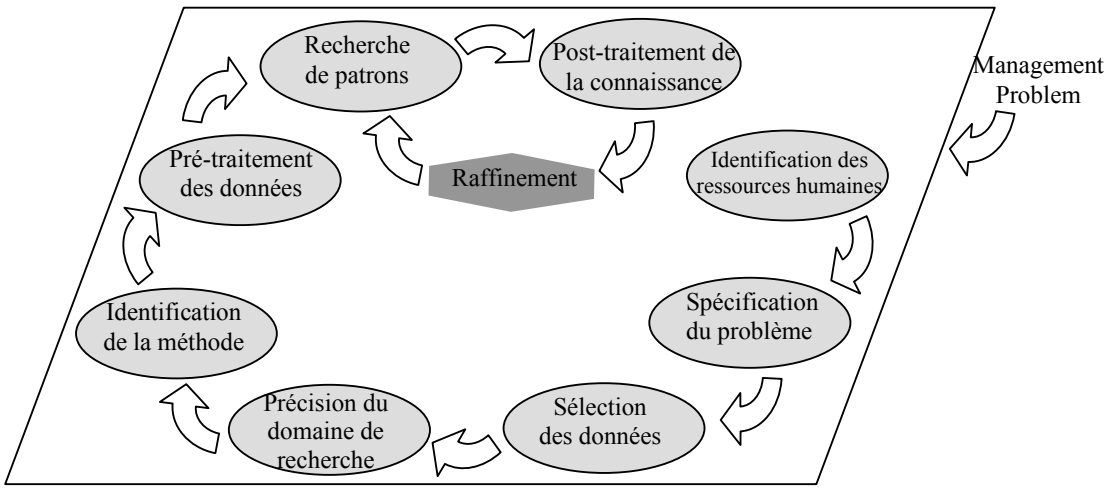

Figure 5: Processus du data mining adapté aux problèmes industriels [9].

Le processus s'amorce par l'identification d'un problème. S'en suit l'identification des ressources humaines adaptées à la résolution de ce problème.

L'**identification des ressources humaines** est nécessaire pour la résolution d'un problème industriel. Il est conseillé de faire appel à un *expert du domaine industriel* considéré. Il aura pour fonction principale d'analyser les résultats et de juger de la pertinence de ceux-ci. Un *expert du système d'information* de l'entreprise sera nécessaire pour trouver et extraire les données indispensables. Enfin un *expert en data mining* sera en charge de l'extraction de l'information. Cet expert devra sélectionner les algorithmes et paramètres les mieux adaptés pour effectuer une analyse précise.

L'étape de **spécification du problème** a pour but d'améliorer la compréhension du problème. Il s'agit d'une analyse et d'une discussion des experts qui permettra si besoin de décomposer le problème en sous problèmes plus faciles à résoudre. Les sous problèmes qui peuvent être résolus à l'aide des techniques de data mining sont identifiées et la meilleure approche de résolution (règles d'association, classification, estimation, segmentation, régression, recherche de patrons fréquents, détection d'écart ou techniques de visualisation – voir Fayyad *et al.* [5]) est discutée.

La **sélection des données** permet de sélectionner et d'analyser l'état des données requises. Il s'agit d'identifier les attributs pertinents, l'accessibilité des données, considérer le traitement des données manquantes et évaluer la distribution et l'hétérogénéité des données. Le but est de disposer de suffisamment de données représentatives pour permettre d'extraire de l'information pertinente et fiable.

Le **domaine de recherche sera précisé**, les connaissances sur le problème permettant de réduire l'espace de recherche sont intégrées.

Identification de la méthodologie : la meilleure méthodologie pour la résolution du problème est sélectionnée.

Le **pré-traitement** des données permet d'adapter le format des données à l'algorithme de data mining utilisé. Cela peut nécessiter de compléter les données manquantes. Il peut être envisagé de réduire le volume de données.

Lors de la **recherche de patrons**, l'algorithme préalablement défini va automatiquement parcourir les données à la recherche du modèle demandé. Dans le cas d'un très grand volume de données qui nécessiterait un temps de calcul relativement élevé, des méthodes de décomposition [10] et de transformation [11] des données sont possibles. Ce processus de recherche est itératif et nécessite plusieurs raffinements.

La **connaissance** extraite va demander à être validée. Cette étape de **post-traitement** va faciliter la compréhension des résultats. Les techniques de visualisation, de formulation des règles et de construction d'arbres de décisions vont présenter les résultats d'une manière interprétable pour l'utilisateur. La nouvelle connaissance devra finalement être validée.

Plusieurs itérations de l'ensemble de ce processus de recherche peuvent être nécessaires pour répondre à un problème.

4 Perspectives d'utilisation en milieu industriel

Comme cela a été présenté en introduction, le volume croissant de données produites et stockées dans les entreprises est le plus souvent archivé dans des entrepôts de données ("data warehouse systems") sans être au préalable exploité.

L'hypothèse de ce travail est que les données archivées contiennent de l'information utile, non exploitée, que nous nous proposons d'identifier pour en montrer les perspectives d'utilisation à plus ou moins long terme. Une bonne utilisation des outils de data mining semble être une clé importante pour l'amélioration de l'utilisation des données stockées en augmentant les connaissances de l'entreprise sur elle-même et sur ses clients [12].

La figure suivante (Figure 6) présente des champs d'application du data mining à partir d'une base de données industrielle.

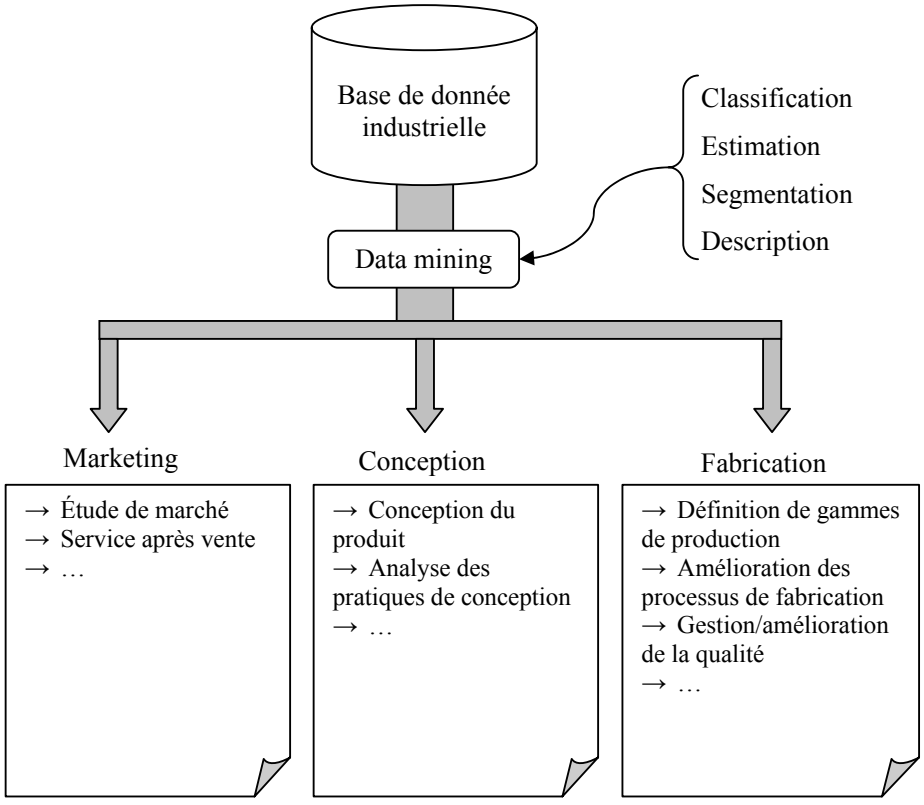

Figure 6 : Champs d'application considérés.

En **étude de marché** : le data mining permet de découvrir des connaissances cachées sur les consommateurs. Berry et Linoff [13] présentent de nombreux exemples et applications du data mining en marketing, vente et suivi du consommateur.

- La segmentation identifie des consommateurs de profils similaires, ces groupes de consommateur permettant de représenter des segments de marchés différents.

- La classification attribuera des produits types, des budgets types… aux différentes catégories de clients.

Couplé aux connaissances actuelles sur les attentes des clients, il sera possible d'identifier des marchés cible en recherchant l'ensemble des clients potentiellement intéressés par tel ou tel produit/service. Il restera ensuite à vérifier si le marché considéré est rentable.

Inversement, il sera intéressant d'identifier les consommateurs qui divergent du profil standard des consommateurs. Les consommateurs divergents peuvent être une catégorie de clients insatisfaits, ceci permettra de faire ressortir des besoins nouveaux, d'où de nombreuses opportunités d'amélioration des produits et services offerts actuellement. Ce point particulier a été traité dans le cas d'un fournisseur d'énergie par Sforma [14].

Le **service après vente** est un point de contact privilégié avec le consommateur. Le service après vente, par un traitement différentié et individuel de chaque client, permettra de découvrir quelles sont les utilisations réelles du produit/service en usage. En fonction des demandes d'entretiens, de pièces de rechange et des re-réglages nécessaires, il sera possible de faire ressortir des classes d'utilisateurs (par segmentation). En identifiant les classes d'utilisateurs et leurs utilisations (par la recherche de relations dans les demandes de ces clients), il apparaîtra des besoins nouveaux (de qualité, de réglages, de fonctions et de services) qui étaient jusqu'alors restés cachés dans les demandes de pièces du service après vente.

Lors de la **conception des produits**, le data mining peut apporter un ensemble d'informations complémentaires aux outils actuels.

- Ce peut être aussi bien en support à la conception des familles de produits [15]. Le data mining permet de faire ressortir les similarités dans les besoins des clients autour d'une famille de produits, ceci permet de structurer l'architecture de la famille de produit. Il est alors possible de cibler un marché, sélectionner les clients concernés et identifier le(s) produit(s)/service(s) correspondants.

- Le data mining se révèle être un puissant outil de modélisation qui permet de compléter les simulations de comportements complexes. Par exemple, Leu *et al.* [16] ont développé une application réelle pour la conception de tunnels dans des roches de natures différentes, là où les outils classiques étaient insuffisants.

- En aide à la standardisation, pour la conception modulaire le data mining permet de faire ressortir les liens entre les besoins de différents composants/fonctions (règles d'association), ces liens sont potentiellement des éléments à standardiser [17].

Pour l'**analyse des pratiques de conception**, Simoff et Maher [18] analysent la participation de différents intervenants dans un environnement de conception collaborative, ils identifient des comportement et situations plus ou moins favorables à la collaboration. La pratique de conception étant de plus en plus supportée par des outils informatique, aussi bien pour la conception des produits que pour le processus de fabrication, nous disposons de plus en plus d'information sur le déroulement de cette conception. Des avancées dans l'analyse automatique du son et de l'image permettront d'enrichir les possibilités d'analyse automatique des pratiques de conception.

Le data mining est aussi utilisé pour **la définition de gammes de production,** Agard et Kusiak [17] définissent ainsi automatiquement des gammes de production par similarité entre les produits déjà fabriqués et les produits nouveaux à fabriquer. Une segmentation des produits relie les produits similaires, des arbres de décision et des règles de décision extraient un processus de production représentatif pour chaque ensemble de produits, le nouveau produit est ensuite affecté à un groupe connu par classification, et enfin le nouveau produit est affecté par le processus représentatif du groupe. Finalement, il ne reste plus qu'à spécialiser le processus généré automatiquement, qui devient

à son tour une entrée dans la base. Cette application permet d'aller vers une génération automatique des gammes de production.

Pour **améliorer les processus de production**, Gertosio et Dussauchoy [19] appliquent les méthodes de data mining afin de réduire le temps de process et de contrôle lors du réglage électronique de moteurs de camion diesel. Pour l'assemblage, Agard et Kusiak [20] identifient les éléments de sous assemblages à réaliser pour diminuer un temps d'assemblage final au coût minimum. L'idée de base de ces applications repose dans une phase d'apprentissage sur les processus passés afin d'en déduire le comportement des processus à venir.

La **gestion/amélioration de la qualité** est aussi une voie de recherche en exploration. Un exemple d'application a permis de prédire la qualité des produits dans une industrie de semi conducteur [21]. De même, le suivi individuel et automatique des produits par le data mining permettra d'identifier les éléments dont le comportement diverge du groupe auquel ils appartiennent. Ces éléments divergents sont potentiellement hors norme. Remonter automatiquement aux sources de cette divergence au niveau des paramètres du process permettra de les corriger. Ceci permet de remonter automatiquement aux sources ayant provoquées un défaut.

5 Conclusions

Dans le contexte d'une entreprise intégrée numériquement, des quantités très importantes de données sont générées chaque jour. Ces données permettent de décrire tout aussi bien les produits et les processus de l'entreprise. Bien que le volume de données produites augmente sans cesse à un rythme qui s'accélère, ces données sont le plus souvent archivées dans des entrepôts de donnés sans être au préalable exploitées. Nous supposons que ces données contiennent de l'information cachée qu'il peut être pertinent d'extraire et d'utiliser pour l'amélioration des connaissances de l'entreprise sur elle-même, sur ses produits et sur ses clients. Pour cela les techniques de data mining sont proposées au titre de l'extraction d'information dans une grande quantité de données.

Cet article avait pour but de montrer en quoi consiste le data mining, quels en sont les outils et quelle est la méthodologie de résolution d'un problèmes avec le data mining. Un processus de résolution dédié aux problèmes industriels a été plus particulièrement présenté.

Des utilisations actuelles du data mining dans des application industrielles ont été présentées qui concernent en particulier le marketing, la conception et la production. Les perspectives d'utilisations du data mining dans ces différentes secteurs montrent qu'il est envisageable d'automatiser de nombreux processus d'extraction de l'information afin de réutiliser cette information dans des outils d'aide à la décision pouvant eux aussi être appuyés par les techniques de data mining.

Des perspectives particulièrement intéressantes du data mining en milieu industriel sont dans la génération automatique de règles de décision, le pilotage des systèmes de production et la génération automatique de pré gammes de production.

Références

[1] A. KUSIAK, "Computanional Intelligence in Design and Manufacturing", Wiley-Interscience, New York, 2000.

[2] P. LYMAN, H. VARIAN, "How Much Information", Retrieved from http://www.sims.berkeley.edu/how-much-info-2003, on August 2004, Berkeley, 2003

[3] A. EDMUNDS, A. MORRIS, "The problem of information overload in business organisations: a review of the literature", International journal of information management, Vol. 20, No. 1, pp. 17-28, 2000.

[4] D. SHENK, "Data smog: surviving the information glut", Harper, San Francisco, 1997.

[5] U.M. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, R. UTHURUSAMY, "Advances in knowledge discovery and data mining", AAAI Press, The MIT Press, 1996.
[6] S. S. ANAND, A. G. BUCHNER, "Decision Support Using Data Mining", Financial Times Pitman Publishers, London, 1998.
[7] C. WESPHAL, T. BLAXTON, "Data Mining Solutions", John Wiley, New York, 1998.
[8] C. APTÉ, "Data Mining: An Industrial Research Perspective", IEEE Computational Science and Engineering, April-June, pp. 6-9, 1997.
[9] A. G. BUCHNER, S. S. ANAND, J.G. HUGUES, "Data Mining in Manufacturing Environments: Goals, Techniques and Applications", Studies in Informatics and Control, Vol. 6, No. 4, 1997, pp. 319-328.
[10] A. KUSIAK, "Decomposition in Data Mining: An Industrial Case Study", IEEE Transactions on Electronics Packaging Manufacturing, Vol. 23, No. 4, pp. 345-353, October 2000.
[11] A. KUSIAK, "Feature Transformation Methods in Data Mining", IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24, No. 3, 2001, pp. 214-221.
[12] B. AGARD, A. KUSIAK, "Chapter 192: Computer Integrated Manufacturing: A Data Mining Approach", in R.C. Dorf (Ed.), The Engineering Handbook, Second Edition, CRC Press & IEEE Press, Boca Raton, FL, pp.192.1 - 192.11, 2005.
[13] M.J.A. BERRY, and G. LINOFF, "Data Mining Techniques: For Marketing, Sales, and Customer Support", New York: Wiley, 2004.
[14] M. SFORMA, "Data mining in a power company customer database", Electric Power Systems Research, Vol. 55, pp. 201-209, 2000.
[15] B. AGARD, A. KUSIAK, "Data-Mining Based Methodology for the Design of Product Families", International Journal of Production Research, Vol. 42, No. 15, pp. 2955-2969, 2004.
[16] S.S. LEU, C. N. CHEN, S.L. CHANG, "Data mining for tunnel support stability: neural network approach", Automation in Construction, Vol. 10, No. 4, pp. 429-441, 2001.
[17] B. AGARD, A. KUSIAK, "Standardization of Components, Products and Processes with Data Mining", International Conference on Production Research, Santiago, Chile, August 1-4, 2004.
[18] S.J. SIMOFF, M.L. MAHER, "Analysing Participation in collaborative design environments", Design Studies, Vol. 21, No. 2, pp. 119-144, 2000.
[19] C. GERTOSIO, A. DUSSAUCHOY, "Knowledge discovery from industrial databases", Journal of Intelligent Manufacturing, Vol. 15, pp. 29-37, 2004.
[20] B. AGARD, A. KUSIAK, "Data Mining for subassembly selection", ASME Transactions: Journal of Manufacturing Science and Engineering, Vol. 126, No. 3, pp. 627-631, 2004.
[21] A. KUSIAK, "Rough set theory: A data mining tool for semiconductor manufacturing", IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24, No. 1, pp. 44-50, 2001.