Université Lumière Lyon2
Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

Pierre-Emmanuel JOUVE
le 10 décembre 2003

Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

Jean-Paul Rasson, Rapporteur	Professeur, Facultés Universitaires N.D. de la Paix, Namur
Gilles Venturini, Rapporteur	Professeur, Université de Tours
Mohand-Saïd Hacid, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Michel Lamure, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Gilbert Ritschard, Examinateur	Professeur, Université de Genève
Nicolas Nicoloyannis, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

1 Introduction, Préambule

"Le meilleur (...) ce n'est pas le mal réel qu'on se donne pour accoler le mot au mot, pour entasser brique sur brique ; ce sont les préliminaires, le travail à la bêche que l'on fait en silence en toutes circonstances, que ce soit dans le rêve ou à l'état de veille. Bref, la période de gestation. Personne n'a jamais réussi à jeter sur le papier ce qu'il avait primitivement l'intention de dire (...)"

- Henry Miller -
Sexus, (édition Buchet-Chastel) (1949)

Le traitement automatique de l'information fait appel à différentes ressources techniques, technologiques et théoriques issues de domaines variés tels que l'informatique, l'intelligence artificielle, la statistique, la théorie des probabilités, l'analyse de données, l'optimisation... La fin des années 80 et le début des années 90 ont vu un faisceau de situations favorables[1] à l'émergence d'un domaine de recherche transdisciplinaire s'attachant spécifiquement au traitement de vastes volumes d'information et à leur valorisation sous forme de connaissances : le "Knowledge Discovery in Databases" [PSCF+89] (expression que la communauté scientifique francophone a traduit plus tard par Extraction de Connaissances à partir de Données (ECD)).

Une quinzaine d'années plus tard, la communauté mondiale en ECD s'est élargie, structurée et a très largement essaimé dans le monde industriel. L'interaction entre la recherche et l'industrie explique certainement d'ailleurs, la rapidité de cette croissance, tout comme l'enthousiasme et l'activité de cette jeune communauté ont également contribué à cette emergence.

Plus encore, nous pensons que l'agitation, l'ébullition autour du phénomène ECD provient de l'adéquation des problématiques industrielles et académiques : ainsi les promesses de l'ECD en terme de valorisation de l'information ne pouvaient laisser insensibles les acteurs industriels au moment où l'information apparaît comme un élément stratégique déterminant.

1. sur le plan technologique (avancées technologiques en informatique : accroissement des capacités de stockage et de calculs), économique (passage de l'économie de l'ère post-industrielle à l'ère informationnelle)... et peut être épistémologique : le traitement massif et informatisé de l'information souffre alors un peu moins de critiques touchant à la rigueur de cette approche

On peut également chercher des explications non conjoncturelles à ce succès comme la transdisciplinarité d'un domaine qui se fait fort de mettre à profit chacune de ses composantes ainsi que les synergies pouvant exister entre elles. Surtout, l'ECD, comme le rappelle sa définition francophone[2], est un processus anthropocentré tirant profit d'une interaction entre l'homme et la machine. Placer l'humain au centre du processus relève tout d'abord d'un intérêt pragmatique : les limites actuelles des systèmes automatiques peuvent être repoussées par utilisation de l'intelligence et de l'expertise humaine. On peut également se hasarder à avancer un intérêt "psychologique" : un utilisateur intégré au processus de traitement de l'information, non dépossédé de ses capacités d'analyse tend à mieux accepter, comprendre le processus d'ECD.

La tendance actuelle de la communauté ECD à se pencher sur ses échecs passés (afin de les féconder et de préparer les succès futurs...?) [LMF02] souligne d'ailleurs la grande nécessité d'intégrer l'homme au sein des systèmes et processus d'ECD.

Les travaux présentés dans cette thèse s'inscrivent au sein de la problématique ECD.

Aussi, l'intégration de l'expertise et des connaissances humaines, la définition de méthodes "permettant un échange", une interaction entre homme et machine ainsi que la prise en compte des contraintes d'utilisabilité pour les méthodes développées constituent des exigences fondamentales sous-jacentes aux travaux que nous présentons ici.

Préalablement évoquée, la relation forte entre recherche et industrie permet et nécessite la prise en compte des besoins issus de la pratique (tels que la limitation des coûts calculatoires, la limitation des coûts de stockage, la conformité à des exigences issues de la distribution de l'information, etc...). Proposer des solutions ne dérogeant pas non plus à ces nécessités constitue la deuxième exigence que nous imposons à nos travaux.

Plus précisément, il s'agit dans ce document de présenter un ensemble de contributions pour l'apprentissage non supervisé. De manière plus détaillée, nous tentons d'apporter ici un ensemble de nouvelles solutions pour l'intégration de l'apprentissage non supervisé au sein d'un processus ECD[3].

2. L'ECD est définie par la communauté scientifique francophone comme le processus itératif, anthropocentré (interactif), non trivial d'identification de connaissances valides, nouvelles, potentiellement utiles et intelligibles au sein d'un ensemble de données.

3. Le processus ECD est schématisé sur la figure 1.1. Ce processus comprend classiquement les étapes suivantes :

   1. comprehension du domaine étudié
   2. création et nettoyage de l'ensemble des données à traiter (Sélection, Construction de Variables....)

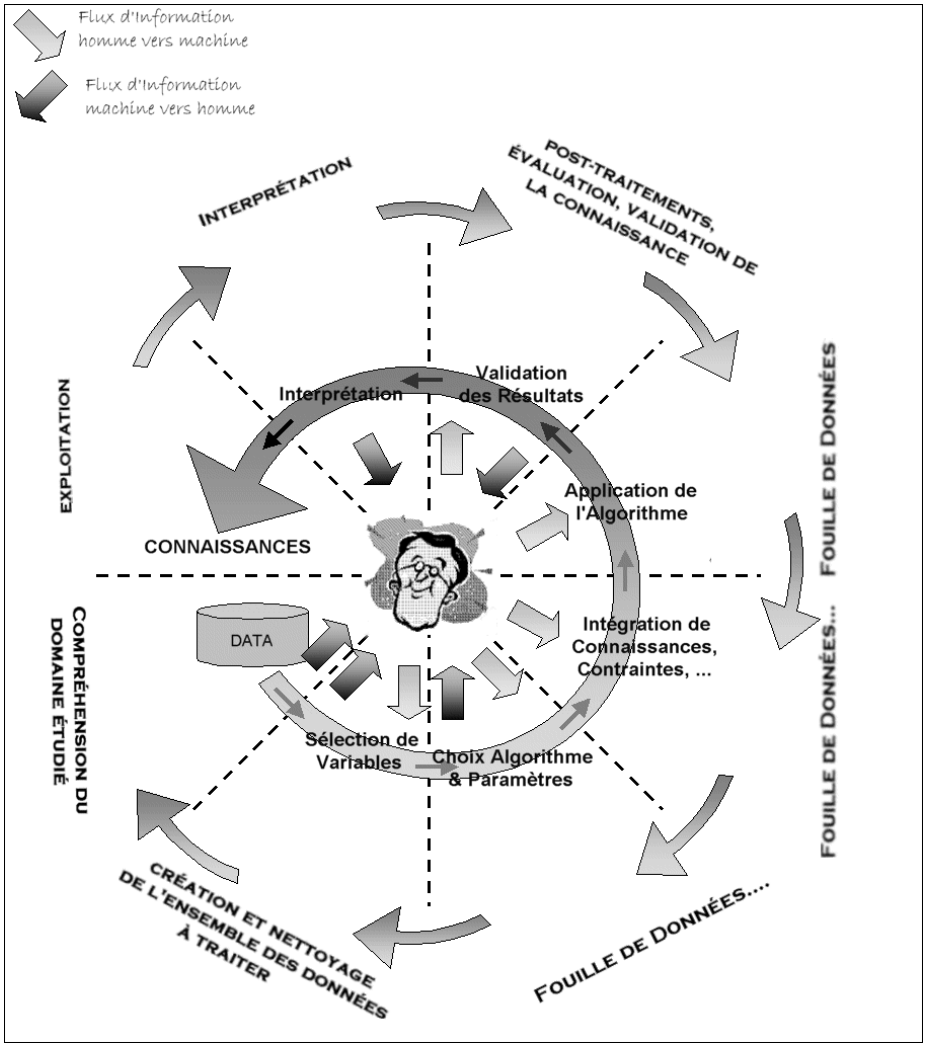

FIG. 1.1 –: *Eléments du Processus ECD*

Ces solutions concernent les différentes étapes clés de ce processus, à savoir, la sélection de variables (chapitre 5), l'application de la méthode d'apprentissage non supervisé 3, la validation / l'estimation de la qualité d'un modèle d'apprentissage non supervisé 4. Un chapitre est également consacré à l'agrégation de modèles d'apprentissage non supervisé et aux différents intérêts que

3. extraction des régularités cachées dans les données et formulation des connaissances mises à jour sous forme de modèles ou de règles (cette étape dans le processus global d'ECD est habituellement désignée sous le nom de fouille des données)
4. post-traitements, évaluation, validation de la connaissance découverte
5. interprétation des résultats
6. exploitation des résultats

cela revêt dans le cadre de la prise en compte des deux niveaux d'exigence évoqués plus tôt.

Notons enfin que le choix de l'apprentissage non supervisé s'explique, d'une part, par l'intégration de ces travaux au sein du projet de recherche universitaire BC3[4] dont l'un des objectifs était la mise au point de méthodes de fusion de données (l'apprentissage non supervisé constituant une piste envisagée), et d'autre part car nous pensons que l'apprentissage non supervisé occupe une place particulière et centrale au sein du processus ECD. Ainsi, la présentation de nos travaux sur l'apprentissage non supervisé est complétée par différentes contributions pour l'apprentissage supervisé exploitant justement nos contributions pour l'apprentissage non supervisé.

Ce document s'organise de la manière suivante : le premier chapitre introduit les éléments nécessaires pour la lecture du reste du document ; les chapitres 3 (présentation d'une nouvelle méthode d'apprentissage supervisé) et 4 (présentation d'une nouvelle méthodologie pour l'évaluation de la validité d'un modèle d'apprentissage non supervisé) peuvent être abordés de manière indépendante des autres alors que la lecture du chapitre 5 (présentation de méthodes de sélection de variables pour l'apprentissage supervisé et non supervisé) et du chapitre 6 (présentation de méthodes d'agrégation de modèles d'apprentissage non supervisé) nécessite respectivement la lecture préalable des chapitres 4 et 3. Notons enfin, que contrairement à la majorité des thèses,

4. Le projet BC3 (Projet Base de Connaissances Cœur-Cerveau, $http : //kbbrain.free.fr$) initié par différentes équipes universitaires et hospitalières vise à la création d'une base de connaissances (BDC) sur les pathologies organiques touchant deux organes vitaux : le cœur et le cerveau. Une telle BDC a pour objectif premier de servir de support pour le recueil et la conservation de données médicales, permettant ainsi la capitalisation de l'expérience et des connaissances de chercheurs et cliniciens des domaines concernés. La métaphore d'une encyclopédie médicale numérique ayant pour sujet les 2 organes vitaux que constituent le cœur et le cerveau peut dans un premier temps être adoptée pour décrire cet outil. Les travaux que nous menons dans le cadre de ce projet impliquent toutefois d'étendre cette métaphore à celle d'une encyclopédie stockant non seulement de l'information mais se voulant également "génératrice" de connaissances. Nous proposons en effet d'intégrer des méthodes de traitement de l'information à la BDC de manière à permettre l'exploitation de l'information stockée dans cette base et la découverte de nouvelles connaissances.
   Equipes Universitaires :
   - CREATIS : Centre de Recherche et d'Applications en Traitement de l'Image et du Signal, Université Claude Bernard Lyon 1, Hôpital Cardiologique L.Pradel Service de Radiologie
   - ERIC : Equipe de Recherche en Ingénierie des Connaissances, Université Lumière Lyon2
   - ISC : Institut des Sciences Cognitives, Université Claude Bernard Lyon1
   - LASS : Laboratoire d'Analyse des Systèmes de Santé, Université Claude Bernard Lyon1
   - TIMC : Techniques en Imagerie, Modélisation et Cognition, Faculté de Médecine de Grenoble

   Equipe hospitalière : Centre de Neuropsychologie de l'hôpital de la Pitié-Salpêtrière (Paris)

le début de ce document n'est pas consacré à un état de l'art général car chacun des chapitres le constituant est amorcé par un état de l'art spécifique.

Nous tenons également à signaler que les différentes expérimentations proposées dans cette dissertation ont été possibles grâce à l'utilisation des logiciels libres : Sipina développé au laboratoire ERIC[5], WEKA de l'Université de Waikato en Nouvelle-Zélande[6] et d'un logiciel mis au point au cours de cette thèse.

5. Sipina est disponible au téléchargement à l'adresse *http://eric.univ-lyon2.fr/%7Ericco/sipina.html*

6. WEKA est disponible au téléchargement à l'adresse *http://www.cs.waikato.ac.nz/ml/weka*