Complexité de l'extraction des connaissances de données : une vision systémique

Walid Ben Ahmed*,***, Mounib Mekhilef*
Michel Bigand**, Yves Page***

*LGI – Laboratoire de Génie Industriel, Ecole Centrale Paris, Grande voie des Vignes 92295 Châtenay-Malabry cedex, France
{walid, mekhilef@lgi.ecp.fr}
**Équipe de Recherche en Génie Industriel, Ecole Centrale de Lille, 59651 Villeneuve d'Ascq, France
michel.bigand@ec-lille.fr
***LAB (PSA-Renault), Laboratoire d'Accidentologie, de Biomécanique et d'études du comportement humain, 132, rue des Suisses-92000 Nanterre
yves.page@lab-france.com

Résumé. Les praticiens et les chercheurs dans le domaine d'Extraction de Connaissances de Données (ECD) sont souvent confrontés à des difficultés qui sont relatives à la nature des données, à l'implication de l'opérateur humain et aux aspects algorithmiques. Aujourd'hui, s'il y a un consensus sur la « complexité » du processus d'ECD, ce n'est pas le cas pour la définition et la caractérisation de cette complexité. Définir la complexité de l'ECD, la caractériser et connaître ses sources sont des questions qui animent aujourd'hui la communauté de fouille de données. Dans cet article, pour répondre à ces questions, nous menons une réflexion sur la notion de complexité en ECD en utilisant l'approche systémique, une approche de modélisation de systèmes complexes.

1 Introduction

Aujourd'hui avec l'informatisation des saisies de données (utilisation des codes à barres, informatisation des transactions, etc.) et la puissance des systèmes de collecte de ces données (satellites, ordinateurs, etc.), des grandes Bases de Données (BD) sont construites et ne cessent de s'agrandir. L'exploitation de ces millions de données en management, en administration, en médecine, en géologie, en biologie et dans beaucoup d'autres domaines fait appel à des techniques d'Extraction de Connaissances de Données.

Le processus d'Extraction de Connaissances de Données (ECD) est défini comme : « *un processus d'identification de modèles (ou paradigmes) valables, nouveaux, potentiellement utiles et compréhensibles dans les données*» (Fayyad, Piatetsky-Shapiro et al. 1996). C'est un processus interactif et itératif, impliquant de nombreuses étapes avec des décisions prises par l'utilisateur (Brachman and Anand 1996). Les praticiens et les chercheurs dans le domaine d'ECD sont souvent confrontés à des difficultés qui sont relatives aux trois phases principales de ce processus (i.e. la *préparation des données*, la *phase de data mining* et *l'interprétation des résultats*). Cependant, s'il y a un consensus sur la « *complexité* » du processus d'ECD, ce n'est pas le cas pour la définition et la caractérisation de cette complexité. Plusieurs facteurs sont généralement considérés comme causes de complexité du

processus d'ECD. Nous citons, par exemple, la quantité de données, la qualité des données (erronées, manquantes, bruyantes, etc.), l'implication de l'opérateur humain, la nature des connaissances à extraire, la complexité des algorithmes utilisés, la multi-disciplinarité du domaine et l'implication de plusieurs points de vue.

Définir la complexité de l'ECD, la caractériser et connaître ses sources sont des questions qui animent aujourd'hui la communauté de fouille de données. Dans cet article, nous menons une réflexion sur la notion de « *complexité* » pour apporter des éléments de réponses à ces questions. Nous utilisons pour cela *l'approche systémique* (à ne pas confondre avec l'approche systématique), qui est une approche de modélisation de systèmes complexes. Dans la première section de cet article, nous présentons un aperçu historique sur les origines de l'approche systémique. Dans la deuxième section, nous présentons ses principes de base. Dans la troisième section, nous utilisons cette approche pour caractériser la complexité du processus d'ECD

2 L'approche systémique : les origines épistémologiques

Aujourd'hui, les termes de la *cybernétique*, *cybernétique du second ordre* (Von Foerster 1995), la *théorie du système général* (Bertalanffy 1969), la *systémique* et la *systémographie* (Le Moigne 1999; Morin and Le Moigne 1999) sont utilisés pour désigner, à peu près, la même approche.

L'*approche cybernétique* ne se focalise pas sur la composition matérielle d'un système, mais insiste sur les interactions entre les composants. L'*observateur* et l'*objet observé* ne sont plus séparés et le *résultat de l'observation* dépend de leur interaction. Il y a eu reconnaissance du fait que toutes nos connaissances sur les systèmes sont basées sur des représentations simplifiées (i.e. des modèles). En cybernétique, Un système n'est plus considéré comme une entité passive qu'on peut observer et manipuler, mais comme un agent qui interagit avec son environnement et avec un autre agent qu'est l'observateur. Ce dernier est lui aussi perçu comme un système cybernétique (i.e. complexe, voir la définition dans le paragraphe 3) qui construit un modèle d'un autre système cybernétique. Il s'agit donc d'appliquer la cybernétique à elle-même ou ce que le fondateur de cette théorie, Heinz von Foerster, appelle la *cybernétique du second ordre* dans son livre « *Cybernetics of Cybernetics* » (Von Foerster 1995).

La systémique (ou la cybernétique) est issue de l'épistémologie constructiviste. Cette épistémologie *reconnaît le caractère relatif de la connaissance et sa dépendance de la construction du sens par les individus en se basant sur leurs expériences et leurs interactions avec leur environnement (contexte)*. Un *système* dans une perspective constructiviste est défini comme une représentation de la réalité perçue par un certain nombre d'individus dans un contexte donné. Un *modèle* est donc une représentation de la réalité et il n'est valide que dans un contexte donné. Les *problèmes* ne sont pas indépendants des perceptions des individus qui les traitent. Il existe alors plusieurs *solutions* et la solution optimale est celle qui est acceptable pour la majorité. Les *méthodes de recherche* ainsi que leurs résultats reflètent donc la perception et l'interprétation des chercheurs. Les *sources d'information* doivent être diversifiées pour couvrir la diversité des opinions. En ce qui concerne les *données*, on accorde plus d'importance au processus de leur collecte qu'aux données elles-mêmes (une synthèse de ces principes peut être trouvée dans (Richard 2003)).

3 L'approche systémique : les principes de base

La systémique distingue les *systèmes complexes* et les *systèmes compliqués*, « *la complexité n'est pas la complication* », nous dit Edgar Morin (Le Moigne 1999; Morin and Le Moigne 1999).

Un **système compliqué** est un système qui est caractérisé par un comportement qui peut être prévu par l'analyse des interactions entre ses composantes, il est déterministe (e.g. un ordinateur). Les approches analytiques sont adaptées à la modélisation de ce type de système. Un **système complexe** ou **système cybernétique** est un système non-déterministe dont le comportement ne peut pas être prévisible par analyse disjonctive de ses différents éléments. Selon Miller (Miller 1995), c'est un système *vivant*, *évolutif* et *ouvert à son environnement* avec lequel il est en *interaction continue*. C'est donc un système qui *fonctionne* et se *transforme* en même temps. Son comportement peut être décrit en terme de *feedbacks* et *boucles récursives* (Morin and Le Moigne 1999). Les éléments d'un système complexe sont en interaction réciproque. L'action d'un élément sur un autre entraîne en retour une réponse du second élément vers le premier. On dit alors que ces deux éléments sont reliés par une *boucle de feedback* (ou *boucle de rétroaction*). Les interactions entre les éléments d'un système sont régies selon le principe du *holisme* : « *le tout est supérieur à la somme des parties* ». Les interactions entre les éléments d'un système donnent à l'ensemble des propriétés que ne possèdent pas les éléments pris séparément.

En se basant sur les fondements du constructivisme, l'approche systémique insiste sur *l'inséparabilité* entre *l'observateur* (e.g. analyste), *l'objet observé* (e.g. données) et le *contexte de l'observation* (e.g. objectif de l'analyse). Une tâche telle que *l'extraction des connaissances de données*, dépend donc non seulement des données, mais aussi de la personne qui l'effectue et du contexte (objectif, environnement, etc.). Puisqu'un système complexe est un système qui *existe*, qui *fonctionne* et se *transform*e en même temps et qui a un *objectif* (ou *téléologie*), la systémique propose de conjoindre quatre points de vue génériques pour l'analyser et appréhender sa complexité. Nous désignons par *point de vue* « *une position conceptuelle par rapport à un objet, cette position servant à en donner une description particulière* ». Les points de vue systémiques sont (Le Moigne 1999) :

- *Le point de vue ontologique* (le « quoi ») : ce qu'est le système. Le terme *ontologie* est issu du domaine de la philosophie où il signifie « explication systématique de l'existence ». L'axe ontologique représente pour nous les composants du système,
- *Le point de vue fonctionnel* (le « faire ») : ce que fait le système,
- *Le point de vue transformationnel* ou *génétique* (le « devenir ») : comment le système évolue, quels sont les états générés. Ce point de vue décrit l'aspect dynamique, évolutionnel et génétique (dans le sens de la genèse et non celui de l'hérédité) du fonctionnement du système,
- *Le point de vue téléologique* ou *motivationnel* (le « pourquoi ») : quels sont l'objectif et la motivation du système. La téléologie signifie en philosophie « l'étude de la finalité ».

4 L'ECD : un système complexe

Dans cette section, nous commençons par proposer une architecture multi-points de vue pour analyser la complexité du processus d'ECD. Pour cela, nous nous basons sur l'approche systémique. Ensuite, nous appliquons cette architecture à la caractérisation de cette complexité.

4.1 Proposition d'une architecture multi-points de vue pour analyser la complexité de l'ECD

Comme le suggère la systémique, nous assimilons le processus d'ECD à un système complexe (ou système cybernétique). Pour analyser la complexité de ce système, nous proposons une architecture constituée de points de vue génériques et ayant deux niveaux d'abstraction : le premier est composé des trois points de vue « *objet observé* », « *observateur* » et « *contexte de l'observation* ». Le deuxième niveau d'abstraction est composé des quatre points de vue *ontologique*, *fonctionnel*, *transformationnel* et *téléologique*. Chacun des trois points de vue du premier niveau est analysé selon les quatre points de vue du deuxième niveau. La Fig. 1 donne une illustration en UML[1] de cette architecture :

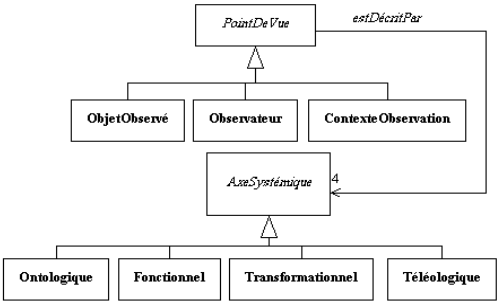

FIG. 1 - *Architecture systémique pour l'analyse de la complexité du processus ECD*

– **Le point de vue « objet observé »**. Dans un processus d'ECD, ce point de vue concerne les données à traiter. Ci-dessous, nous analysons ce point de vue selon les quatre aspects systémiques :

  – *Aspect ontologique* : cet aspect est relatif à la nature des données à traiter. Ces données peuvent être multi-domaines, multi-formes (textuelles, images, vidéo etc.), multi-niveaux de granularité. Elles peuvent être incomplètes, bruitées, aberrantes et incohérentes. Cela a un lien avec le processus de leur collecte que nous abordons dans le point suivant,

  – *Aspect fonctionnel* : cet aspect est relatif à la façon dont les données ont été collectées. En d'autres termes, il concerne le processus de la collecte des données. Ces dernières peuvent être issues de mesures automatiques effectuées par des machines (e.g. température). Elles peuvent être aussi issues d'entretiens effectués par des opérateurs humains. Les opérateurs peuvent être issus de disciplines différentes et dans ce cas là, on aura des données multi-sources. Ils peuvent avoir des expériences différentes et/ou des points de vue différents même s'ils sont de la même discipline,

  – *Aspect transformationnel et génétique* : cet aspect est relatif à l'évolutivité des données. Le fait qu'elles ne soient pas collectées au même moment peut avoir une influence sur la façon de les traiter et sur la nature des connaissances qu'on peut y trouver,

  – *Aspect téléologique* : cet aspect est relatif aux objectifs pour lesquels les données ont été collectées. Sachant que cet objectif peut évoluer au fil du temps, l'objectif de l'étude en

[1] UML : Unified Modeling Language (Booch et al. 1999).

question peut ne pas être prévu pour ces données. Cela affecte les méthodes à utiliser ainsi que les résultats à espérer,

- **Le point de vue « observateur ».** Dans un processus d'ECD, ce point de vue concerne l'opérateur (ou les opérateurs) humain(s) chargé(s) de la tâche d'ECD ainsi que la tâche d'ECD elle-même :
  - *Aspect ontologique* : cet aspect est relatif à l'expérience de l'opérateur (statisticiens, expert du domaine, etc.), son domaine et sa discipline. Il est relatif aussi aux ressources (humaines et matérielles) utilisées pour résoudre la tâche,
  - *Aspect fonctionnel* : cet aspect est relatif aux différentes fonctions effectuées durant le processus d'ECD. Elles peuvent être regroupées en trois catégories principales : la préparation de données, la phase de data mining et l'interprétation des résultats. La dépendance des différentes tâches a une influence certaine sur le déroulement du processus ECD ainsi que sur les résultats. En effet, les résultats d'une tâche telle que la tâche du data mining peuvent amener à revoir ses inputs qui sont les outputs d'une autre tâche. Il s'agit donc de revoir la tâche précédente, qui est la préparation des données dans ce cas là. Les outputs d'une tâche conditionnent aussi le déroulement de la tâche suivante. C'est le cas par exemple de la relation entre les tâches préparation de données/data mining ou data mining/interprétation des résultats,
  - *Aspect transformationnel et génétique* : cet aspect est relatif à la transformation des données tout au long du processus d'ECD. Quant à l'aspect génétique, il est relatif aux résultats générés durant le processus. Ces deux aspects ont une influence sur la perception de l'observateur qui peut elle-même évoluer et donc faire évoluer la façon dont les données sont traitées ou même collectées,
  - *Aspect téléologique* : cet aspect concerne l'objectif de l'étude. La sélection des données pour l'étude ainsi que les méthodes d'analyse utilisées dépendent étroitement de cet objectif. Ce dernier peut même évoluer au cours de l'analyse, influencé par exemple, par des résultats intermédiaires,
- **Le point de vue « contexte de l'observation »**. Dans un processus d'ECD, ce point de vue concerne l'environnement dans lequel s'effectue la tâche d'ECD.
  - *Aspect ontologique* : cet aspect est relatif à l'environnement physique dans lequel s'effectue le processus d'ECD. Il inclut, entre autres, l'utilisateur final des connaissances à extraire (un apprenti, un expert, un groupe d'experts, une organisation),
  - *Aspect fonctionnel* : cet aspect est relatif à la question suivante : « comment ces connaissances seront-elles utilisées ? ». Faut-il les représenter ou traduire sous une forme bien déterminée en vue d'une exploitation précise ? Toutes ces questions induisent des contraintes au niveau du processus d'ECD,
  - *Aspect transformationnel* : cet aspect est relatif à l'évolutivité de l'environnement auquel les connaissances sont destinées. Des connaissances extraites de données et qui sont exploitables aujourd'hui, ne le seront pas forcément si l'environnement de leur exploitation change,
  - *Aspect téléologique* : cet aspect est relatif à l'objectif global pour lequel les connaissances ont été extraites. Généralement cet objectif est décliné en sous-objectifs au niveau du processus ECD. On peut très bien imaginer qu'un objectif au niveau du contexte n'est pas complètement atteint par le processus ECD. Par exemple, construire des connaissances en accidentologie pour améliorer la sécurité routière ne peut pas être atteint uniquement par l'application de l'ECD sur une base de données d'accidents.

L'architecture que nous venons de présenter sert à présenter les différents points de vue dont il faut tenir compte quand il s'agit d'analyser la complexité de l'ECD. Cependant, nous pensons que cette complexité ne peut pas être analysée d'une façon disjonctive au niveau de ces différents points de vue. Nous ne pouvons pas parler, par exemple, de la complexité de données indépendamment du processus de leur traitement ainsi que des objectifs du traitement. Des données peuvent paraître complexes pour un objectif donné sans l'être pour un autre objectif. Nous ne pouvons pas non plus parler de la complexité d'un processus de traitement indépendamment des données et des objectifs de l'étude, etc. Nous proposons alors, dans la section suivante, la caractérisation de la complexité du processus d'ECD à travers des concepts qui font la conjonction de ces différents points de vue.

4.2 Caractérisation systémique de la complexité de l'ECD

Aucun des aspects systémiques des trois points de vue (i.e. données, tâches effectuées et contexte de l'étude) ne peut être considéré comme la source de la complexité du processus d'ECD s'il est considéré isolément des autres. Mais, c'est la conjonction des différents aspects au niveau de ces trois points de vue qui rend complexe ce processus. Il s'agit précisément des concepts de comportement circulaire tels que l'auto-application, le comportement projectif, les boucles de rétroaction. Dans cette section, nous définissons ces concepts qui sont des caractéristiques de la complexité du processus d'ECD.

4.2.1 Le concept d'auto-application

Le concept *d'auto-application* (*self-application*) est le plus général parmi les concepts de circularité. Son expression mathématique est donnée à travers l'équation suivante : $y = f(y)$. La forme discrète de circularité est exprimée à travers l'équation $y_{t+1} = f(y_t)$. La forme plus générale de ce principe est exprimée à travers la formule suivante : $y = kf(y)$[2]. Pour expliquer ce principe, nous prenons l'exemple suivant : *y = un écran TV* et *f = une caméra pointée sur cet écran et en même temps transmettant l'image sur lui*. L'image dans cette situation est cause et effet en même temps, donc $y = f(y)$.

Revenons à notre processus d'ECD que nous assimilons à un système composé par « *les données* », « *l'opérateur humain* » et « *le contexte de l'étude* ». Nous définissons la notion « *d'état* » de ce système de la manière suivante : à chaque instant *t*, ce système est défini par *{le résultat de la transformation des données, la perception de l'opérateur, l'objectif de l'étude}*. En définissant ainsi notre système, nous identifions le même *phénomène d'auto-application*. En effet, tout au long de ce processus, l'opérateur applique des tâches (e.g. nettoyer les données, appliquer une technique de data mining, etc.) ce qui génère des transformations et des résultats. Ces derniers ont une influence sur la perception de l'opérateur lui-même car en fonction de ces résultats l'opérateur essaye d'affiner son analyse en définissant de nouvelles tâches (e.g. réutiliser la même technique en changeant les paramètres, utiliser une autre technique, etc.). Il s'agit donc d'un *processus itératif* durant lequel les tâches appliquées dépendent des résultats intermédiaires qu'elles génèrent. Vis-à-vis d'un observateur externe, les tâches sont causes et effet en même temps. Si on pose *y= « l'ensemble des tâches effectuées »* et *f= « application d'une tâche »*, on peut donc écrire $y = f(y)$, ce qui veut dire que le choix des tâches dépend des résultats intermédiaires.

[2] En algèbre linéaire *k* représente les valeurs propres de *f*.

La même chose peut être observée au niveau des données. En effet, les données influencent le choix des techniques appliquées par l'opérateur[3] ce qui génère de nouvelles données dont la nature dépend des données initiales et des techniques de traitement. Ainsi, les données sont cause et effet en même temps. En posant *y= « l'ensemble des données »* et *f= « tâche de traitement appliquée »*. Vis-à-vis d'un observateur externe, les données sont les résultats d'eux-mêmes, c'est-à-dire *y =f(y)*.

Le concept d'auto-application peut être généralisé au niveau de tout le système ECD en posant *y= « système ECD »* et *f= « l'ensemble des tâches effectuées »*. Ce système est le résultat de lui-même, donc *y =f(y)*. L'auto-application peut être perçue comme un comportement du système ECD qui essaye de s'adapter à un objectif fixe (celui de l'étude), qui est dans notre cas « *extraire des connaissances des données pour un autre objectif* ». Ce comportement est appelé *un comportement projectif* que nous présentons dans la section suivante.

4.2.2 Le concept de comportement projectif

Une des principales caractéristiques d'un système complexe réside dans le fait qu'il a ses propres objectifs qu'il essaye de réaliser en résistant à toutes les perturbations. Un *comportement projectif* (*goal-directness*) implique une régulation ou un contrôle des perturbations guidé par cet objectif. Ce rôle de contrôle dans un système d'ECD est réalisé essentiellement par l'opérateur humain. En effet, ce dernier choisit ses actions en fonction de l'évolution du système (e.g. résultats intermédiaires), mais aussi en fonction de l'objectif qu'il s'est fixé pour l'étude qu'il effectue. Il essaye d'une façon continue à résister aux perturbations (e.g. bruits au niveau des données) pour atteindre l'objectif de l'étude. La question maintenant est : *pourquoi un comportement projectif est-il source de complexité* ?

- La première réponse est que ce comportement nécessite d'abord de décliner l'objectif de l'étude en sous-objectifs à réaliser au cours du processus. Cela nous amène à un autre concept caractéristique de la complexité : le principe de structure hiérarchique de contrôle (cf. 4.2.3) ;
- La deuxième est qu'il fait appel à des processus de régulation non-linaires ce qui nous amène à un autre concept caractéristique de la complexité : les boucles de rétroactions (cf. 4.2.4).

4.2.3 Le concept de structure hiérarchique de contrôle

Dans un système complexe, les objectifs sont organisés en hiérarchie. Si une boucle de contrôle ne suffit pas pour réduire les effets d'une perturbation, il faut ajouter une autre (cf. Fig. 2). Un exemple typique dans les organisations est la tendance d'augmenter le nombre des niveaux bureaucratiques. Dans le cas du processus ECD, atteindre l'objectif d'une étude nécessite de préparer les données, appliquer une technique de data mining et interpréter les résultats. Chacune de ces trois tâches est ensuite déclinée en sous-objectifs, etc.

La Fig. 2 représente cette structure hiérarchique. Pour atteindre un objectif, l'opérateur effectue une perception pour faire une représentation interne[4] des perturbations. Il effectue ensuite, moyennant un processus de traitement d'information, une confrontation de ces

[3] Le type de données (texte, image, etc.) ainsi que leur nature (qualitative, quantitative) jouent un rôle dans le choix des techniques de traitement.
[4] interne au système ECD.

perceptions avec les objectifs du système ECD suite à quoi il décide des actions. Ces dernières, moyennant un processus dynamique de transformation, modifient l'effet des perturbations. Cette boucle est exécutée jusqu'à l'atteinte d'un résultat satisfaisant pour le régulateur (i.e. l'opérateur dans notre cas). Si cette boucle de contrôle ne suffit pas pour atteindre l'objectif, il faut en ajouter une autre. Cependant, plus le nombre de couches d'hiérarchie de contrôle est important, plus les bruits sur les perceptions et les actions à entreprendre par l'opérateur sont importants. Il est donc préférable de maximiser la capacité de régulation d'une seule couche d'hiérarchie et réduire le nombre de couches.

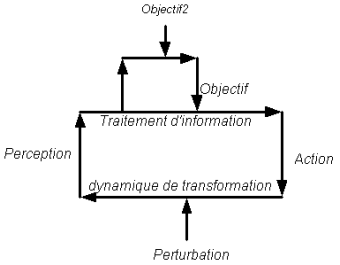

FIG. 2 - *Une structure hiérarchique de contrôle*

L'opérateur est confronté aux questions suivantes : Quel est le nombre de boucles optimal ? Comment traduire un objectif global d'une étude donnée en sous-objectifs en tenant compte des autres composantes du système (i.e. données et contexte) ? Comment gérer le conflit entre les sous-objectifs ? Les réponses à ces questions dépendent des données traitées, du contexte de l'étude et de l'opérateur. Ce dernier est amené sans cesse à effectuer des choix. Mais, dès qu'une décision est prise, des fonctions sont effectuées et des transformations sont induites, l'état du système change et de nouvelles décisions qui peuvent être contradictoires aux premières doivent être prises. Ainsi, une décision dépend finalement d'elle-même ce qui nous renvoie encore au principe d'auto-application. Ce comportement est assuré par ce qu'on appelle les boucles de rétroaction que nous présentons dans la section suivante.

4.2.4 Le concept de boucle de rétroaction

Le principe d'auto-application que nous avons présenté dans la section 4.2.1 et traduit par la formule $y=kf(y)$, peut être utilisé pour analyser la déviation d'un système par rapport à un état donné y_0 (e.g. un état d'équilibre). En d'autres termes la déviation $\Delta y = (y-y_0)$ à un instant $(t+\Delta t)$ dépend de la déviation à l'instant t. Cela est exprimé à travers la formule suivante : $\Delta y(t+\Delta t)= k\Delta y(t)$.

Si on pose « y_0 =*l'état du système quand il atteint son objectif* », le *comportement projectif* (cf. 4.2.2) parait comme la tentative d'atteindre y_0, donc de supprimer la déviation Δy. Ainsi, la déviation de l'état y du système par rapport à la situation objective y_0 dépend d'elle-même, c'est-à-dire : $(y-y_0)_{(t+\Delta t)}= k(y-y_0)_t$. Il s'agit de ce qu'on appelle le principe de rétroaction ou de feedback.

Dans le cadre du processus d'ECD, le principe de feedback exprime bien l'interdépendance entre les différentes tâches du processus, i.e. préparation de données, data mining et interprétation des résultats. En effet, selon l'interprétation des résultats de la fouille, on décide de refaire ou pas l'étape de data mining (e.g. rappliquer la technique de data mining en changeant des paramètres,

appliquer une nouvelle technique, etc.). On peut choisir aussi de revenir à l'étape de préparation des données (e.g. sélectionner d'autres données, refaire le nettoyage, etc.). Il s'agit bien donc, conformément à la définition du processus d'ECD donnée dans (Brachman and Anand 1996), d'un *processus itératif*.

Les feedbacks peuvent être négatifs[5] et tendent vers la stabilisation du système quand une déviation positive (par rapport à y_0) génère une déviation négative. Prenons comme exemple *y=nombre de variables* à considérer dans une étude de fouille de données. Soit y_0 le nombre optimum. Si l'opérateur augmente y au-dessus du seuil y_0, la connaissance extraite diminue[6] ce qui conduit l'opérateur à réduire y. Cela va faciliter l'interprétation et augmenter la connaissance extraite. Si y continue à diminuer et passe au-dessous du seuil y_0, la connaissance extraite diminue aussi ce qui va amener l'opérateur à augmenter y. Ainsi le système ECD, grâce à son aspect itératif, oscille autour de la position y_0 qui correspond en quelque sorte à un état d'équilibre.

Les feedbacks peuvent être aussi positifs[7] et tendent dans ce cas vers l'amplification quand une déviation positive (par rapport à y_0) génère une déviation positive. Prenons l'exemple précédent, mais avec un autre mode de régulation se basant sur la règle suivante : « $\forall$ *l'état de y par rapport à* y_0, *Si (connaissance diminue), Alors (augmenter le nombre de variables y)* ». Dans ce cas, la boucle de rétroaction devient positive. Elle est traduite par : *Nombre de variables augmente* $\Rightarrow$ *Connaissance diminue* $\Rightarrow$ *Nombre de variables augmente* $\Rightarrow$ *Connaissances diminue, etc*.

Ce sont essentiellement les feedbacks négatifs qui assurent la stabilité et la convergence d'un processus d'ECD. Le rôle de l'opérateur peut donc être perçu comme celui d'un régulateur utilisant ce type de feedbacks.

5 Conclusion & perspectives

Nous avons montré dans cet article que la complexité d'un processus d'ECD doit être analysée selon différents points de vue. En se basant sur une approche de modélisation de systèmes complexes (i.e. la *systémique*) nous avons défini une architecture d'analyse de la complexité de l'ECD. Cette architecture est composée de deux niveaux d'abstraction : le premier est composé des trois points de vue « *objet observé* », « *observateur* » et « *contexte de l'observation* » qui correspondent respectivement aux « *données* », « *l'opérateur humain* » et « *le contexte de l'étude* ». Le deuxième niveau d'abstraction est composé des quatre points de vue *ontologique*, *fonctionnel*, *transformationnel* et *téléologique*. Chacun des trois points de vue du premier niveau est analysé selon les quatre points de vue du deuxième niveau. Les sources de complexité du processus d'ECD ont été analysées selon cette architecture.

Nous avons montré ensuite qu'aucun des aspects systémiques des trois points de vue ne peut être considéré comme la source de la complexité du processus d'ECD s'il est considéré isolément des autres. Mais, c'est la conjonction des différents aspects au niveau de ces trois points de vue qui rend complexe ce processus. Nous avons défini alors des concepts faisant cette conjonction et permettant la caractérisation de la complexité du système ECD. Il s'agit des concepts d'*auto-application*, de *comportement projectif*, de *structure hiérarchique de contrôle* et de *boucles de rétroaction*.

[5] Le feed-back est négatif quand *k* dans la formule *y=kf(y)* est négatif.
[6] Car le bruit et la difficulté d'interprétation augmentent.
[7] Le feed-back est positif quand *k* dans la formule *y=kf(y)* est positif.

Le travail dans ce papier a eu pour objectif de comprendre la complexité d'un système d'ECD en la caractérisant. Mais, cet objectif n'est pas une fin en soit car l'objectif final est de *contrôler cette complexité*. En effet, dans la réalité, le choix des actions dans un processus d'ECD n'est ni opéré complètement à l'aveuglette, ni de façon complètement déterministe. Il dépend d'un processus d'apprentissage issu d'expériences antérieures. Les feedbacks aident le système d'ECD à ajuster sa réponse. De nouvelles questions surgissent alors et méritent une réflexion profonde : Quels sont les différents modes de contrôle de la complexité du système d'ECD ? Quel est le rôle de l'apprentissage ? Quel est le rôle des connaissances du domaine ?

Références

Bertalanffy, L. v. (1969). General system theory: foundations, development, applications. New York, George Braziller.

Booch, G., R. J., et al. (1999). Unified Modelling Language User Guide, Addison Wesley Professional.

Brachman, R. and T. Anand (1996). The Process of Knowledge Discovery in Databases: A Human-Cen-tered Approach. In Advances in Knowledge Discovery and Data Mining, 37-58, eds. U. Fayyad, G. Piatet-sky- Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, Calif.

Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications Of The ACM **39**(11): 27-34.

Le Moigne, J.-L. (1999). La modélisation des systèmes complexes, Dunod.

Miller, J. G. (1995). Living Systems, University Press of Colorado.

Morin, E. and J. L. Le Moigne (1999). L'intelligence de la complexité. Paris, L'Harmattan.

Richard, H. (2003). Approaches to Inquiry - Key Concepts, ICRA Learning Materials - Methods.

Von Foerster, H. (1995). The Cybernetics of Cybernetics (2nd edition). Minneapolis, Future Systems Inc.

Summary

Patricians and researchers in knowledge Discovery in Databases (KDD) have to address many issues related to the data nature, the implication of human operator and algorithmic aspects. Nowadays, there is a consensus that the KDD process is complex. However, there is discordance when questions deal with the definition and characterization of this complexity. To answer these questions, we use the systemic approach, which is a complex system modeling approach.