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