European Journal of Natural and Social Sciences-Novus, 01(10), 01000140EJNSS
e-ISSN (2785-5236)
https://ejnss.novuspublishers.org/

Research article

Divergence possible des processus de Data mining et Knowledge Discovery in Databases

*** [1] Dr. YENDE RAPHAEL Grevisse, [2] KASEKA KATADI Viviane**

[1]Ingénieur Docteur ès sciences en Systèmes Informatiques de l'Université du Commonwealth et Enseignant-chercheur (Professeur des Universités) attaché à l'Université de Bas-Uélé (RDC)

[2]Ingénieure Licenciée en Conception des systèmes informatiques attachée à la Faculté de l'Informatique de l'Université Notre Dame du Kasayi (RDC).

Received: 4.11.2022
Accepted: 28.1.2023
Final Version: 21.2.2023

*Corresponding Author:
grevisse29@gmail.com

ABSTRACT
Confusion remains to this day between the terms « Data Mining » (DM), translated into French as « Fouille de Données » (FD), and Knowledge Discovery in Databases (KDD), translated into French as « Extraction de Connaissance à partir de Données » (ECD). For many researchers and practitioners, the term DM is used as a synonym for ECD in addition to being used to describe one of the steps in the ECD process. It is for this reason that this article has set itself the main objective of giving the main theoretical meanings between the two concepts (FD and KDD) and trying to explain the difference that would exist in terms of process, storage and data access.

Keywords: Data mining, Knowledge Discovery, Segmentation, Classification, Method, Process, Databases, Divergence, Possible

Introduction

Les systèmes d'information des entreprises actuelles sont de plus en plus « submergés » par des données de tous types: structurées (bases de données, entrepôts de données), semi-structurées (documents XML, fichiers log) et non structurées (textes et multimédia). Ceci a créé de nouveaux défis pour les entreprises et pour la communauté scientifique, parmi lesquels comment classifier, comprendre et analyser de telles masses de données afin d'en extraire des nouvelles connaissances.

En outre traditionnellement, l'exploration de données et la découverte des connaissances étaient effectuées manuellement. Au fil du temps, la quantité de données dans de nombreux systèmes est devenue supérieure à la taille du téraoctet et ne pouvait plus être gérée manuellement. De plus, pour le succès de toute entreprise, la découverte de modèles sous-jacents dans les données est considérée comme essentielle.

La visualisation des résultats dans un processus de Data Mining est une étape très importante. En effet c'est souvent sur cette étape que repose l'acceptation ou le refus par l'utilisateur final (le décideur) de l'outil d'aide à la décision visé. Le

choix d'un cycle de conception de système interactif adapté s'avère déterminant pour son acceptation par les décideurs. Il existe à ce sujet plusieurs modèles classiques provenant du Génie Logiciel, tels les modèles en V, en spirale, en cascade ou par incrémentation. Il en existe aussi d'autres qui sont enrichis sous l'angle des IHM tels les modèles en étoile de Hartson et Hix, le modèle nabla proposé par Kolski ou le modèle en U proposé par Abed et Millot [8].

Par contre, plusieurs outils logiciels ont été développés pour découvrir des données cachées et faire des hypothèses, qui faisaient partie de l'intelligence artificielle, c'est ainsi que le processus KDD a atteint son apogée au cours des 10 dernières années. Il héberge désormais de nombreuses approches différentes de la découverte, notamment l'apprentissage inductif, la sélection de données, le nettoyage de données, les statistiques bayésiennes, l'optimisation des requêtes sémantiques, l'acquisition de connaissances pour les systèmes experts et la théorie de l'information avec comme but ultime d'extraire des connaissances de haut niveau à partir de données de bas niveau. Le processus KDD comprend des activités multidisciplinaires.

Cela comprend le stockage et l'accès aux données, la mise à l'échelle des algorithmes en ensembles de données massifs et l'interprétation des résultats, le processus de nettoyage et d'accès aux données inclus dans l'entrepôt de données ou les bases de données facilite le processus KDD, l'intelligence artificielle soutient également le KDD en découvrant des lois empiriques à partir d'expérimentation et d'observations, les modèles reconnus dans les données doivent être valides sur de nouvelles données et posséder un certain degré de certitude (ces modèles sont considérés comme de nouvelles connaissances) et l'exploration de données, également connue sous le nom de découverte de connaissances dans les bases de données, fait référence à l'extraction non triviale d'informations implicites, auparavant inconnues et potentiellement utiles à partir de données stockées dans des bases de données. Bref, l'article présent nous donne la convergence entre le DM et KDD par rapport aux possibilités qu'ils nous offrent dans la gestion de données et l'extraction des connaissances. *Notre article est composé de trois parties. Dans une première partie nous présentons les notions du processus de Data Mining. Dans la deuxième nous présentons le processus du KDD. La troisième partie vise à faire diverger les deux processus.*

APPROCHE THEORIQUE

Processus du Data Mining

Le *data mining* est considéré comme l'extraction d'informations intéressantes (non triviales, implicites, préalablement inconnues et potentiellement utiles) à partir de grandes bases de données. Il permet d'analyser les données pour trouver des patrons cachés en utilisant des moyens automatiques. C'est un processus non élémentaire de recherche de relations, corrélations, dépendances, associations, modèles, structures, tendances, classes (clusters), segments, lesquelles sont obtenues de grandes quantités de données (généralement stockées sur des bases de données (relationnelles ou non)). Cette recherche est effectuée à l'aide des méthodes mathématiques, statistiques ou algorithmiques. *Data Mining* se considère comme un processus le plus automatique possible, qui part de données élémentaires disponibles dans un *Data Warehouse* à la décision. Le Data Mining est une partie du processus KDD.

Utilisations actuelles de *data mining:* Les domaines d'application du **data mining** sont très nombreux à savoir la prévision de la part de marché, la segmentation de la clientèle, l'implantation de point de vente, l'analyse des portefeuille clientèles, analyse de données scientifiques, le contrôle de données financières en temps réel, la prévision des charges et de demande, le dépouillement de données d'étude, la détection de comportements frauduleux , l'analyse de tendances démographiques, le diagnostic et la maintenance préventive, la modélisation de toxicité, etc [7].

Le profil des participants comporte les décideurs et analystes, désireux d'exploiter leurs données afin de s'assurer une grande réactivité, une meilleure compétitivité et pour améliorer la qualité de leurs prestations dans les environnements évoluant rapidement et fortement concurrentiels : directeurs commerciaux, responsables marketing, responsables de la technologie, responsables d'analyse de données techniques, scientifiques ou commerciales ainsi que les ingénieurs et techniciens devant mettre en œuvre les techniques de data mining au sein des applications qu'ils conçoivent.

Processus du *data mining:* Littéralement, Data Mining signifie « fouille des données » ; le *Data Mining* désigne l'ensemble des techniques informatiques, outils et applications, permettant de découvrir automatiquement des connaissances nouvelles au sein des grandes bases de données [10]. En fonction du type de données stockées et des informations recherchées, le *Data Mining* utilise des outils issus de nombreuses spécialités différentes: statistiques, algorithmes génétiques, réseaux neuronaux, etc. C'est un champ de recherche multidisciplinaire qui s'est progressivement constitué au cours des dix dernières années.

Par conséquent, on peut regrouper les objectifs des méthodes de *Data Mining* en cinq grandes fonctions: Classification, estimation, segmentation, prédiction, explication. Le choix de la méthode dépendra de la nature du problème posé et du type de données dont on dispose. On peut résumer le processus du *Data Mining* à la mise en œuvre dans les systèmes

d'informations de la succession des tâches suivantes : (1) identifier les données d'intervention, (2) utiliser les techniques du *Data Mining* pour transformer les données en informations utiles, (3) transformer les informations en actions concretes, (4) évaluer les resultants [4]. D'une manière plus précise, le *Data Mining* peut être redéfini par cette suite d'opérations de transformation et d'analyse des données ***(4)***:

- Nettoyage des données: selon une phase d'élimination du bruit et des données inutiles, on filtre, trie, homogénéise, nettoie. En effet, les données peuvent être incomplètes, contradictoires, ou contenir des erreurs humaines ou informatiques.
- Intégration des données: elle consiste en une phase d'association de multiples sources des données sous une forme unique (consolidation), généralement dans le cadre de l'architecture d'un entrepôt de données (*data warehouse*).
- Sélection des données: les données ayant un rapport avec l'analyse demandée sont retrouvées dans la base.
- Transformation des données: les données sont regroupées, normalisées, et transformées dans un format qui les prépare au *Mining*. C'est une sorte de prétraitement avant la fouille des données.
- Data Mining ou fouille de données: c'est un processus essentiel consistant à appliquer des méthodes « intelligentes » pour extraire des éléments remarquables, des patterns. Il s'agit de configurations de données dont la structure est inhabituelle, qui présentent des corrélations imprévues, des écarts statistiques, ou tout ce qui sort de l'ordinaire.
- Evaluation et interprétation des Patterns: on identifie les patterns intéressants, ceux qui représentent de l'information. L'intérêt des patterns est évalué par les outils de Data mining en utilisant des règles objectives basées sur la structure des patterns et les statistiques qui les sous-tendent, ainsi que des règles subjectives basées sur les croyances des utilisateurs (ce savoir est stocké dans une base de connaissances).
- Présentation de la connaissance: des techniques de visualisation et de représentation doivent être utilisées pour présenter clairement à l'utilisateur le savoir extrait des données: tables, arbres, règles, graphiques, courbes, matrices, cubes, etc.

Voici d'une façon synthétique comment se présente le processus du data mining

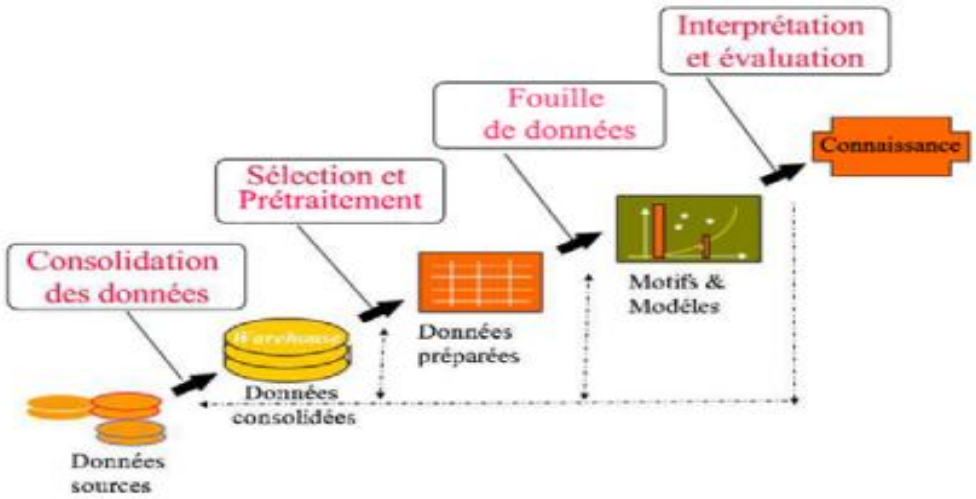

Figure 1: Processus du Data mining

Logiciels de data mining*:* Les principaux logiciels sont présentés ci-dessous selon leur prix et le niveau de compétence requis.

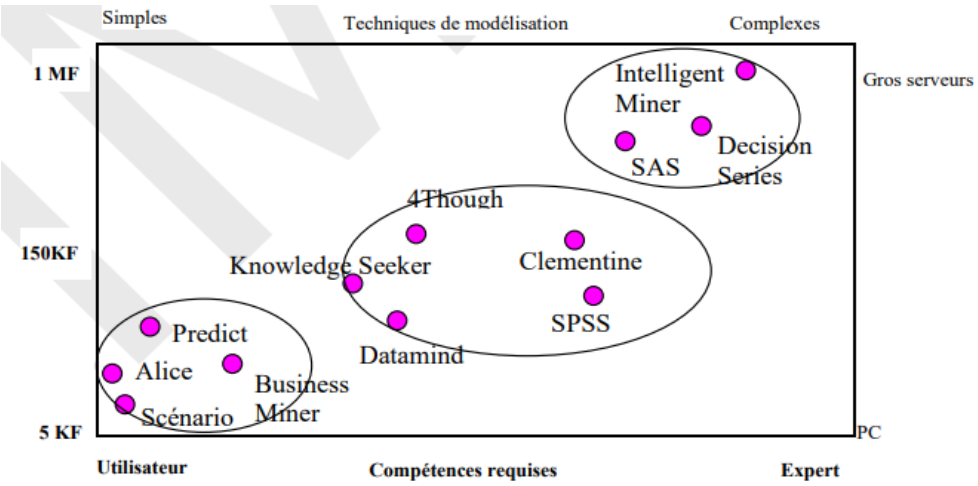

Figure 2: Principaux logiciels du data mining

On constate que se dégagent dans cette figure différents groupes de logiciels qui peuvent être caractérisés comme suit:

- Des logiciels simples demandant peu de compétences particulières de la part des utilisateurs. On retrouve dans cette catégorie des produits tels que ***Business Miner, Alice,*** … La plupart de ces produits sont dédiés à une seule méthode. Ces outils souvent utilisés sur PC et ont la particularité d'être fortement conviviaux et faciles d'apprentissage.
- Une deuxième catégorie de produit qui nécessite certaines connaissances mais restent utilisables par un utilisateur non-averti et offrent des fonctionnalités permettant aux experts du domaine de préciser certains paramètres. On retrouve ici des produits tels que *Datamind, Clémentine*. Les produits de cette gamme fonctionnent essentiellement sur PC mais peuvent également être utilisés en mode client-serveur.
- Enfin, une troisième catégorie regroupe les outils demandant le plus d'expertise dans le domaine, même si des efforts et des progrès en termes de convivialité et facilité d'apprentissage ont été faits par les éditeurs. On retrouve ici des produits tels *que SAS Enterprise Miner, Intelligent Miner (IBM)* Ces outils permettent le traitement de forts volumes de données et offrent une gamme complète de méthodes. Pour le traitement de données à forte volumétrie, il est évident que leur mise en œuvre s'effectue essentiellement sur les gros serveurs.

Processus KDD

L'extraction de connaissances à partir de données désigne un ensemble de méthodes issues des statistiques, de l'intelligence artificielle et de la reconnaissance de formes dont le but est de valider des éléments de connaissances à partir du traitement des données [5].

L'extraction de connaissances et *data mining* consiste à donner un sens aux grandes quantités de données, d'un certain domaine, capturées et stockées massivement par les entreprises d'aujourd'hui. En effet, la vraie valeur n'est pas dans l'acquisition et le stockage des données, mais plutôt dans notre capacité d'en extraire des rapports utiles et de trouver des tendances et des corrélations intéressantes pour appuyer les décisions faites par les décideurs d'entreprises et par les scientifiques. Cette extraction fait appel à une panoplie de techniques, méthodes, algorithmes et outils d'origines statistiques, intelligence artificielle, bases de données, etc. Cependant, avant de tenter d'extraire des connaissances utiles à partir de données, il est important d'avoir une procédure bien claire et d'en comprendre l'approche globale.

En effet, simplement savoir des algorithmes d'analyse de données et les appliquer sur des données en main n'est pas suffisant pour la bonne conduite d'un projet de data mining. Certes, une application aveugle des méthodes de data mining sur les données en main peut mener à la découverte de connaissances incompréhensibles voire même inutiles pour l'utilisateur final [6].

C'est principalement pour cette raison que l'activité de l'extraction de connaissances et *data mining* a été rapidement organisée sous forme d'un processus appelé processus d'Extraction de Connaissances à partir de Données (ECD).

Ce processus se présente comme un processus complexe, non trivial, composé de plusieurs étapes itératives, et nécessitant une interactivité permanente de la part de l'utilisateur expert. Le processus constitue une feuille de route à suivre par les praticiens lors de la planification et la réalisation des projets d'extraction de connaissances à partir de données. Dans ce contexte, l'ECD émerge comme un domaine à part entière, sans remettre en cause ses origines, qui intègre de nouvelles problématiques [16]. Et on peut même annoncer, sans craindre de critiques, qu'il se développe petit à petit en une ingénierie d'Extraction et Gestion de Connaissances (EGC) [16] qui dispose actuellement de ses propres modèles, méthodologies et langages.

L'objectif du présent sous point est de survoler les différents aspects du processus d'ECD, de la terminologie du domaine aux modèles et méthodologies du processus, passant par les étapes, les tâches et les outils de l'ECD. Voici comment se présente le schéma du processus KDD:

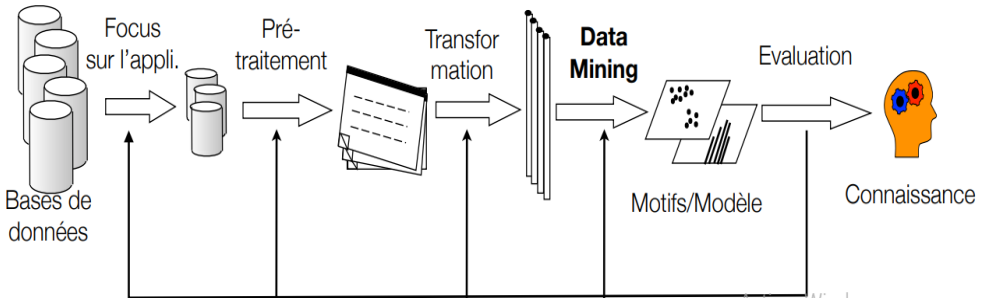

Figure 3: Processus du KDD

Définition

L'Extraction de Connaissances à partir de Données (ECD) est le domaine de recherche où on étudie les techniques, les outils et les méthodologies visant l'extraction de connaissances nouvelles et potentiellement utiles à partir de grandes masses de données. On retient ici la définition donnée par Fayyad [15] « *le processus d'Extraction de Connaissances à partir de Données est un processus non trivial qui permet d'identifier, dans des données, des patterns ultimement compréhensibles, valides, nouveaux et potentiellement utiles* ». Ce processus se présente comme un processus itératif et interactif effectué sur plusieurs étapes interrompues continuellement par des prises de décision par l'utilisateur expert.
L'ECD est un processus itératif et interactif qui fait appel à un ensemble de techniques et outils issus de différents domaines tels que: les bases de données, la statistique, l'intelligence artificielle, l'apprentissage automatique, la reconnaissance de formes, l'analyse de données, et les techniques de visualisation.

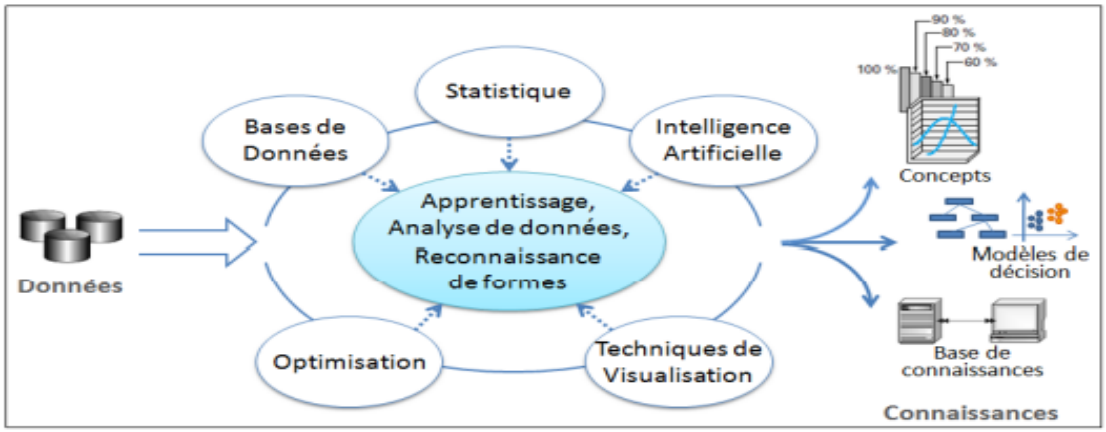

Figure 4: Techniques et domaines en relation avec le processus d'ECD

Le processus d'ECD vise à transformer des données (volumineuses, multiformes, stockées sous différents formats sur des supports pouvant être distribués) en connaissances. Ces connaissances peuvent s'exprimer sous forme de concepts généraux qui enrichissent le champ sémantique de l'usager par rapport à une question qui le préoccupe. Elles peuvent prendre la forme d'un rapport ou d'un graphique. Elles peuvent s'exprimer comme un modèle mathématique ou logique pour la prise de décision. Les connaissances extraites doivent être les plus intelligibles possibles pour l'utilisateur. Elles doivent être validées, mises en forme et agencies [18]. Le processus d'ECD s'effectue sur plusieurs étapes interrompues continuellement par des prises de décision par l'utilisateur expert [17]. Il nécessite sommairement la préparation des données, la recherche de patterns et l'évaluation des connaissances extraites et leur raffinement, toutes répétées dans plusieurs iterations.

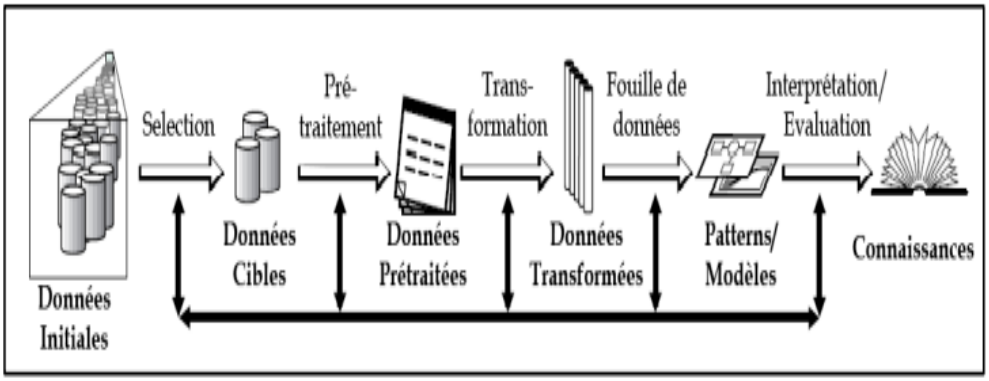

Figure 5 : Le processus de l'ECD

Etapes du processus de KDD: Neuf étapes ont été proposées initialement pour le processus d'ECD mettant en valeur le caractère itératif du processus ainsi que le rôle central de l'utilisateur expert.
Ces étapes se résument comme suit [3].

1. **Compréhension du domaine d'application**: consiste à développer une compréhension du domaine d'application et des connaissances pertinentes préalables. Cette étape prépare l'analyste pour comprendre et définir les objectifs opérationnels du processus d'ECD du point de vue des utilisateurs immédiats de ses résultats.
2. **Création d'un jeu de données cibles**: l'analyste doit sélectionner les données à utilisées (pour la fouille, la validation et le teste) et les attributs pertinents pour la tâche de fouille de données.
3. **Nettoyage des données et prétraitement**: cette étape vise la préparation d'un jeu de données « propre » et bien structuré. Elle comprend des opérations de base telles que: l'élimination des données bruyantes et/ou des valeurs aberrantes (si jugée convenable), recueil des informations nécessaires pour modéliser et tenir compte du bruit, choix des stratégies de traitement des valeurs manquantes, ainsi que de décider des questions sur la base de données à utiliser (les types de données, le schéma à utiliser et le mapping des valeurs manquantes ou inconnues).
4. **Réduction et projection des données**: il s'agit de trouver des attributs utiles pour représenter les données en fonction de l'objectif de la tâche d'extraction, et d'utiliser des méthodes de réduction de dimensionnalité ou de transformation afin de réduire le nombre effectif de variables d'étude et dégager de nouvelles variables plus pertinentes. Cette étape est très importante pour la réussite du projet d'ECD et doit être adaptée en fonction de la base de données et des objectifs opérationnels du projet.
5. **Choix de la tâche de fouille (*data mining task)***: faire correspondre les objectifs opérationnels du processus d'ECD (étape 1) à une tâche particulière de fouille de données, comme la classification, la régression, le clustering, ou la description et synthèse de données.
6. **Choix des algorithmes de fouille de données appropriés:** consiste à sélectionner les méthodes à utiliser pour la chercher de patterns dans les données, décider quels sont les modèles et paramètres appropriés (les modèles pour les données qualitatives diffèrent de ceux conformes aux vecteurs de valeurs réelles), et conclure par le choix d'une méthode particulière de fouille de données en accord avec le critère global du processus d'ECD (par exemple l'utilisateur final peut être intéressé plus par la compréhension du modèle que par ses capacités de prédiction).
7. **Fouille de données (data mining):** il s'agit d'exécuter la ou les méthodes choisies (étape 6) avec leurs paramètres afin d'extraire des patterns d'intérêt sous une forme de représentation particulière. Par exemple des règles ou arbres de classification, des modèles de régression, des clusters, et autres. Parfois il sera nécessaire d'appliquer la méthode de fouille plusieurs fois pour obtenir le résultat escompté.
8. **Interprétation des patterns extraits**: cette étape comprend l'évaluation et l'interprétation des modèles découverts dans les données. Il peut être nécessaire de retourner à l'une des étapes 1 à 7 pour des itérations éventuelles. Cette étape donne l'occasion de revenir sur les étapes précédentes, mais aussi d'avoir une représentation visuelle des patterns, supprimer les patterns redondants ou non représentatifs et de transformer le résultat en informations compréhensibles par l'utilisateur final.
9. **Consolidation des connaissances extraites:** en utilisant directement ces connaissances, en les incorporant dans d'autres systèmes pour des actions ultérieures, ou simplement en les documentant et les rapportant aux utilisateurs concernés. Ceci inclue également la détection et la résolution de tout conflit potentiel avec d'autres connaissances déjà confirmées ou extraites. Les étapes présentées ci-dessus ne sont pas séquentielles. En effet, lors d'un processus d'ECD, l'analyste est amené à faire des itérations et des révisions entre étapes. De plus, la plupart des travaux de définitions et

de modélisation du processus d'ECD se croisent pour confirmer que ce processus est composé de trois grandes étapes ***(9)***: *prétraitement des données, fouille de donnée et post-traitement des patterns identifiés.*

La littérature sur les techniques, outils, et algorithmes d'ECD est abondante. On a même tendance actuellement à publier les outils de *data mining* sous forme de Services Web et Services *Computing* [14]. Cependant, durant un projet d'extraction de connaissances, l'utilisateur/analyste n'est généralement pas un expert d'ECD, mais une personne chargée de donner un sens aux données à l'aide des techniques de fouille disponibles. Et puisque le processus d'ECD est, par nature, interactif et itératif complexe, un des défis du domaine est d'offrir un environnement « intelligent » où les utilisateurs d'ECD seront assistés dans le choix, le paramétrage et la composition de méthodes et d'outils appropriés pour atteindre leurs objectifs. Certes, une application aveugle des méthodes de fouille de données sur les données en main peut mener à la découverte de connaissances incompréhensibles, voire inutiles pour l'utilisateur final.

Complexité du processus d'ECD

Le processus d'ECD, de sa nature itérative et interactive, est largement accepté comme un processus complexe pour plusieurs raisons ci-dessous ***(2):***

- **La nature des données manipulées**: la plupart des algorithmes et techniques de fouille de données sont conçus pour manipuler des données à forme tabulaire simple et à types atomiques (entiers, réel, nominal). Or ces dernières années, notamment dans le cas du Web, les données à analyser sont hétérogènes (textes, hypertextes, images, audio, et vidéo), incohérentes, évolutives, incomplètes et distribuées. Et la situation est de plus en plus complexe avec l'avènement de nouvelles technologies telles que le Cloud Computing, les systèmes de fichiers distribués et Big Data. Des données avec telles caractéristiques nécessitent un certain niveau d'expertise de la part de l'utilisateur d'ECD pour les préparer et les adapter aux conditions d'exécution des algorithmes de fouille disponibles.
- **La nature des tâches effectuées durant les différentes phases du processus d'ECD:** si ces tâches se basent sur l'expertise dans leur application ou l'interprétation de leurs résultats, le processus d'ECD devient plus complexe. Certains algorithmes exigent une certaine forme de données, ce qui nécessite de créer des objets intermédiaires rendant compliqué le processus.
- **L'interdépendance des étapes d'ECD:** plus les étapes sont dépendantes, plus les feedbacks et la récursivité sont intenses et plus l'implication de l'utilisateur est accrue.
- **L'objectif de l'étude et la nature des connaissances recherchées**: pour une analyse de données et l'extraction de connaissances, il faut trouver un support adéquat pour ces connaissances qui réduit au maximum la perte d'informations et qui permet leur exploitation.
- **La multitude des méthodes et des algorithmes pour réaliser une même tâche d'ECD**, le grand nombre de paramètres de ces algorithmes, et les préconditions et post conditions de leur exécution (nature des attributs, traitement des valeurs manquantes et des valeurs aberrantes). Par exemple, pour un problème de classification, plusieurs méthodes de classification sont disponibles (Arbres de Décision, Classification Bayesienne, Réseaux de Neurones, Régression Logistique).

Et pour chaque méthode, plusieurs algorithmes sont possibles et avec un certain nombre de paramètres (pour les arbres de décision: CART, ID3, C4.5.) Ainsi, la mise en œuvre efficace d'un projet de fouille de données nécessite des connaissances pointues et des décisions appropriées sur un bon nombre de techniques spécialisées (la préparation des données, la transformation des attributs, le choix d'algorithmes et de paramètres, les méthodes d'évaluations des résultats, etc.). Ce qui fait qu'un utilisateur novice de l'ECD a besoin d'une assistance pour bien réussir ces objectifs.

Méthodologie

Méthodes utilisées

Plusieurs techniques de *data mining existent en voici quelques-unes* ***(7):***

Les arbres de decision: Les arbres de décision sont particulièrement appréciés et utilisés dans les domaines de la prévision et la segmentation. Ils sont mis en œuvre pour analyser les relations entre une variable Y (variable à expliquer ou dépendant variable) et un ensemble de p variables X j (variables explicatives ou indépendant variable) dans le but d'élaborer des règles.

Les règles sont alors utilisées pour prévoir quelle valeur sera prise par la variable à expliquer en fonction des variables explicatives. Un arbre est une représentation sous une forme particulière de règles de décision comprenant des opérateurs logiques " ET ", " OU ".

Les règles d'association: Les règles d'associations visent à construire un modèle constitué d'un ensemble de règles conditionnelles de type « Si…. Sinon… Alors ». Elles sont élaborées à partir d'un fichier de données. La recherche des associations peut s'opérer sur l'ensemble des données en testant toutes les conclusions possibles ou sur une donnée cible où la conclusion est fixée par l'utilisateur. Les principales utilisations de cette méthode concernent essentiellement le diagnostic de crédit ainsi que l'analyse des tickets de caisse de magasin, ou encore le fonctionnement des cartes de crédit ou de fidélité. Néanmoins, cette technique est également applicable dans l'industrie pour l'analyse des pannes. Plus généralement, ces techniques s'appliquent pour les problèmes où l'apparition d'un événement est conditionnée par des événements passés. Cette méthode consiste à évaluer les affinités existantes entre les variables. Dans le cas de la vente d'articles, on cherche à mettre en évidence les produits vendus simultanément par identification des liaisons existantes.

Les réseaux de neurons: Les réseaux de neurones représentent l'une des techniques de *data mining* la plus utilisée mais qui en même temps est la plus mystifiée. Les statisticiens hésitent à la mettre en œuvre car ils ont l'impression d'une "boîte noire": il est difficile de savoir comment les résultats sont produits, ce qui rend les explications délicates, même si les résultats sont corrects. Cette technique est une transposition simplifiée des neurones du cerveau humain. Dans leur variante la plus courante, les réseaux de neurones apprennent sur une population d'origine puis sont capables d'exprimer des résultats sur des données inconnues.

Ils sont utilisés dans la prédiction et la classification dans le cadre de la découverte de connaissances dirigée. Certaines variantes permettent l'exploration des séries temporelles et des analyses non dirigées (*réseaux de Kohonen*). Le champ d'application est très vaste et l'offre logicielle importante. Le fonctionnement d'un réseau de neurones est inspiré de celui du cerveau humain. Il reçoit des impulsions, qui sont traitées, et en sortie d'autres impulsions sont émises pour activer les muscles. Il existe deux types de réseaux : à apprentissage supervisé où la réponse est connue ; à apprentissage non supervisé où le réseau ne connaît pas le résultat.

Les algorithmes génétiques: Les algorithmes génétiques représentent une technique dont la vocation principale est l'optimisation, mais on peut également les utiliser pour des prédictions ou classifications. Leur champ d'application est très large. C'est une technique récente : les premiers travaux datent de la fin des années 50 où les biologistes et informaticiens ont coopéré pour modéliser les mécanismes génétiques sur ordinateur. Mais c'est surtout au début des années 60 que John Holland développa ses travaux de recherche sur ce thème. Le terme d'algorithme génétique, quant à lui, date de 1967, tandis que l'algorithme date de 1975. Ces concepts étaient mal perçus car, à cette époque, régnait la recherche opérationnelle capable de trouver, par définition, la meilleure solution, tandis que les algorithmes génétiques se basent sur un processus aléatoire dont l'objectif est de trouver une meilleure solution que celle en cours.

Actuellement, peu de produits commerciaux proposent ces algorithmes. En fait ils sont intégrés de manière transparente pour optimiser l'apprentissage des réseaux de neurones. Les algorithmes génétiques trouvent de nombreux domaines d'applications. Ils sont utilisés dans : l'industrie pour optimiser ou contrôler les processus (pression d'un cylindre, température d'un four, etc…) ; les domaines spatiaux (choix des meilleures implantations d'un distributeur automatique de billets de banque) ; le domaine marketing (choix des meilleurs candidats à une offre).

Le raisonnement à base de cas: C'est une technique de découverte de connaissances dirigée utilisée dans un but de classification et de prédiction. Appelée également raisonnement basé sur la mémoire (RBM), ou Case Based Reasoning (CBR) en anglais, cette technique est l'équivalence de l'expérience chez l'homme : en fonction d'elle, tout le monde peut prendre une décision. Lorsqu'un médecin pose un diagnostic et un traitement, il applique sa connaissance du patient et de symptômes similaires au cas présent. Son processus sera l'identification des cas similaires, puis l'application de l'information provenant de ces cas au problème actuel. Le RBC fonctionne sur le même Principe: lorsque l'on présente un nouvel enregistrement, le RBC trouve les voisins les plus proches et positionne ce nouvel élément.

Le RBC s'applique à tous les types de données. Le RBC s'adapte bien aux bases de données relationnelles, qui sont les plus courantes dans le domaine de gestion. Sa mise en œuvre est simple, ce qui en fait un outil apprécié. On peut l'utiliser pour : estimer des éléments manquants ; détecter des fraudes ; déterminer le meilleur traitement d'un malade ; prédire si un client sera intéressé ou non par telle offre ; classifier les réponses en texte libre.

Les réseaux bayésiens: Les réseaux bayésiens sont une méthode probabiliste classique utilisée pour estimer une probabilité d'apparition d'un événement, étant donné la connaissance de certains autres événements. Ils consistent en un modèle graphique qui encode les probabilités entre les variables les plus pertinentes. Cette technique complétée par des statistiques classiques permet de comprendre les relations causales existant entre les variables (mesure d'impact) pour déclencher une action. En outre, un réseau bayésien est un graphe orienté dont les nœuds représentent des variables et les arcs symbolisent les dépendances entre les variables. Chaque nœud ne peut être

relié que par un nœud ou une variable le précédant. La probabilité d'une variable est mesurée par sa fréquence d'apparition. La force des relations entre les variables est mesurée par les probabilités conditionnelles.

Comparaison des méthodes

L'ensemble des méthodes présentées dans le paragraphe précédent répond globalement à la même problématique : dégager la connaissance à partir des données. Les techniques auxquelles nous sommes susceptibles de faire appel sont diverses mais permettent de comprendre, segmenter et prévoir le comportement traduit par une variable. Face à cette diversité de méthodes, le problème de choix de la bonne méthode persiste. En effet, si la nature des données et l'objectif poursuivi nous guide, il n'en reste pas moins qu'il se dégage une impression de redondance et de profusion des méthodes [13]. Pour comprendre cette diversité, il faut savoir qu'à son origine le *data mining* recouvrait les quelques méthodes que nous avons présentées précédemment. Auparavant, les seules méthodes étaient celles dites plus classiques qui nécessitaient une culture statistique pour leur utilisation et surtout leur interprétation. Le *data mining* visait à mettre à la disposition des décideurs ces techniques. On assistait à un phénomène de vulgarisation des techniques.
Aujourd'hui le terme de *data mining* recouvre l'ensemble des techniques permettant de comprendre les données. Les techniques de *data mining* proprement dites telles que les arbres de décision ou les réseaux de neurones apparaissent fréquemment comme une solution miracle, mais qu'en est-il vraiment ? Sont-elles si différentes des techniques classiques ? Quels sont leurs atouts et leurs faiblesses ? Ont- elles des limites ? Les réponses à ces questions nous sont données dans les sous points qui suivent.

Différences et similitudes

Comme on a pu le constater à travers les paragraphes précédents, les techniques de *data mining* visent à répondre d'une autre façon, qui se veut plus claire et compréhensive pour un néophyte. On peut établir cette comparaison à travers le tableau suivant :

	Techniques classiques			**Nouvelles Techniques**		
	AFC, ACP, ACM	Analyse discriminante	Régressions	Arbres de décision	Règles d'association	Réseaux de neurones
Objectifs	Analyse descriptive des données, compréhension des relations entre les variables	Etablir une prévision pour une variable qualitative à l'aide des variables explicatives qui différencient le plus les individus	Etablir un modèle sous la forme d'une équation linéaire afin de prévoir une variable à expliquer en fonction de variables explicatives	Etablir des règles représentées sous la formes d'un arbre afin d'effectuer une segmentation des prévisions.	Etablir des règles	Etablir des prévisions et segmentation
Principes	Utilisation des notions de distance, de corrélation, projection des individus et/ou variables sur un plan factoriel. Interprétation des axes en s'appuyant sur	Utilisation des notions de corrélation, de variance. Utilisation d'hypothèses à vérifier sur les variables explicatives	Utilisation des principes de corrélation, d'estimation des paramètres du modèle par des moindres carrés. Tests de	Les principes sont différents selon les algorithmes, mais consistent à déterminer les modalités de variables qualitatives ou quantitatives	Reposent sur les notions de fréquence d'apparition de couples de variables	Estimation d'une fonction non-linéaire complexe Repose sur des algorithmes d'apprentissage.

	un ensemble d'indicateurs fournis par les méthodes		significativité du modèle par la méthode des moindres carrés Tests de significativité du modèle et des variables introduites dans le modèle. Intervalle de confiance des paramètres estimés	permettant de séparer la population initiale en sous-ensembles. On parle également de variables discriminantes		
Résultats	Représentation sous forme d'un plan factoriel défini par deux ou plusieurs axes qui sont des résultats synthétiques des variables initiales. Indicateurs de qualité et de représentation	Représentation sous la forme d'équations du problème. Elle permet d'effectuer une prévision.	Une équation avec de type : y=f(x)+G. Elle permet d'effectuer une prévision ainsi que des simulations.	Un arbre qui parcourt de la racine aux feuilles selon les modalités prises par un individu placé en entrée. La lecture s'effectue en termes de probabilités	Un ensemble de règle	Un modèle non explicité. Des indicateurs de qualité du modèle.

Tableau comparatif des différentes techniques de data mining.

Remarque: CART peut concurrencer les méthodes plus classiques que sont la régression multiple, l'analyse discriminante et la régression logistique pour sa problématique. On constate donc que, globalement, ces familles de technique recouvrent les mêmes problématiques et objectifs mais sont différentes relativement aux concepts utilisés au sein de ces techniques. Néanmoins, les méthodes dites de *data mining* font appel à des notions statistiques tels que la corrélation, les tests du Chi-Deux, le calcul de fréquence,

Les différences essentielles concernent:

- ***les utilisateurs potentiels*:** les techniques classiques nécessitent une bonne connaissance du domaine statistique afin d'interpréter les résultats alors que les résultats fournis par les techniques de data mining ne nécessitent aucune connaissance particulière
- ***le volume de données***: les analyses classiques peuvent être parfois limitées par le volume de données à traiter. Les volumes actuels importants. Les techniques d'analyse classique font fréquemment appel à des calculs matriciels forts consommateurs de mémoire. Les techniques de *data mining* répondent plus facilement à cette problématique de forte volumétrie ;
- ***l'utilisation des résultats**:* les résultats fournis par les méthodes dégageant un ensemble de règles sont utilisables et peuvent être exploités tout comme les modèles estimés à l'aide des méthodes dites classiques ;
- ***la clarté des résultats***: les résultats fournis par les techniques de *data mining* sont explicites et facilement

exploitables, en particulier les arbres de décision, ou totalement obscurs (les réseaux de neurones !) alors que ceux fournis par les techniques classiques sont explicités mais nécessitent une bonne culture statistique.

Ces différences explicitent clairement les forces et faiblesses de chaque méthodologie. Les techniques de datamining sont:

- facilement exploitables ;
- permettent l'analyse d'importants volumes de données ;
- aisées d'accès ; mais elles ne permettent pas ou peu de simulation ; Les techniques classiques :
- nécessitent une connaissance du domaine statistique ;
- fournissent des résultats de bonne qualité ;
- permettent la simulation ;
- sont parfois limitées en volumes de données à traiter.

Néanmoins, les techniques de *data mining* ont également leurs limites et contraintes.

Limites et contraintes

Comme toutes techniques, certaines nécessitent une transformation des données qui peuvent introduire un biais dans les résultats. En effet, en particulier, par exemple, les réseaux de neurones nécessitent que les valeurs fournies en entrée soient comprises entre 0 et 1 ce qui peut biaiser les résultats. De même, on a vu que les arbres de décision se présentaient sous une forme conviviale mais ils se trouvent également limités par le volume de données si une variable dispose d'un trop grand nombre de modalités. De plus, il convient de s'assurer que sa lecture et son usage s'effectuent en termes de probabilities [16].

Un certain nombre de règles sont à respecter, en particulier celles préconisées dans la démarche qui est d'échantillonner la population initiale globale en deux sous-populations qui constitueront les échantillons d'apprentissage et de test contenant respectivement 70% et 30% de la population initiale. Le premier sera utilisé pour déterminer et construire le modèle alors que le second sera plus particulièrement utilisé pour valider le modèle obtenu préalablement à sa mise en exploitation.

Aujourd'hui, on constate un engouement pour ces techniques de la part des industriels, mais leur succès tient plus à la présentation des résultats qu'aux performances réelles de ces analyses. Néanmoins, elles répondent ainsi à un besoin d'analyses statistiques réalisées non plus par l'expert du domaine statistique mais par un néophyte en la matière. Le décideur obtient alors des résultats simples et synthétiques lui permettant d'orienter ses décisions. Face à cet engouement, les éditeurs de logiciel ont développé un ensemble d'outils qui vont faire l'objet d'une présentation dans le paragraphe suivant.

Résultats et Discussion

L'extraction de connaissances à partir des données (ECD) se définit comme « l'acquisition de connaissances nouvelles, intelligibles et potentiellement utiles à partir de faits cachés au sein de grandes quantités de données ». En fait, on cherche surtout à isoler des traits structuraux (*patterns*) qui soient valides, non triviaux, nouveaux, utilisables et si possible compréhensibles ou explicables. Deux dénominations courantes, mais pas tout à fait équivalentes, se rencontrent habituellement dans la littérature anglosaxonne : le *Knowledge Discovery in Databases* (KDD) et le *Data Mining* (DM). La différenciation entre ces deux désignations réside dans le type d'approche utilisée : intelligence artificielle pour le KDD avec utilisation d'heuristiques provenant de l'apprentissage symbolique, statistique pour le DM considéré comme une industrialisation des techniques d'analyse des données.

Pour certains auteurs les outils de *Data Mining* se résument aux réseaux de neurones et aux arbres de décision autorisant la prédiction d'une variable qualitative (arbres de classification) ou quantitative (arbre de régression) [16]. Néanmoins, les méthodes les plus novatrices concernent la recherche de règles d'associations pouvant conduire à des observations de type « composition du panier d'achat du consommateur », et l'étude des séquences fréquentes permettant d'appréhender le comportement des clients dans le temps [1]. Deux grands types méthodologiques président aux techniques de DM : le mode supervisé qui nécessite la définition d'une variable dépendante (donc certaines hypothèses) et le mode non supervisé où toutes les variables sont considérées sur le même plan (détection des associations, classification, partition, etc.).

Quoi qu'il en soit, KDD et DM ont en commun l'utilisation de mégabases ou *Data Warehouses* (DW) [11], entrepôts de données orientés utilisateurs. Et bien que dans l'esprit, le KDD assimile le cycle complet de traitement des données (le DM faisant plutôt référence à leur analyse statistique), leur finalité reste cependant sur le fond strictement identique, l'objectif étant de fournir une aide décisionnelle aux managers à partir de bases de données souvent volumineuses.
Discipline émergente et multifocale, rassemblant les travaux des chercheurs en statistiques, intelligence artificielle, apprentissage automatique, reconnaissance de formes, bases de données, visualisation des données et linguistique, l'ECD génère des techniques et des outils permettant la révélation de connaissances enfouies dans d'énormes quantités de données hétérogènes et protéiformes.

Conclusion

Pour l'entreprise la révolution induite par les NTIC est loin d'être achevée. Non seulement les flux d'information circulante vont s'accélérer et s'amplifier mais les technologies d'analyse et d'extraction de nouvelles connaissances de ce type de données vont se développer puis se généraliser. De plus l'essor du recours aux techniques de *data mining et* ses outils sans ignorer le KDD permet l'anticipation des choix et orientations stratégiques devenue nécessaire sur un marché concurrentiel.
Parallèlement, les évolutions récentes du Web Mining, notamment en matière de multimédia, devraient permettre une assimilation quasi totale d'un univers informationnel externe. Néanmoins, même si les bases de données, les intranets, les SIAD et les progiciels y afférents sont clairement des outils permettant l'acquisition, la capitalisation puis la diffusion des connaissances, ils ne s'attachent qu'à une partie de la connaissance explicite formalisable et sont généralement focalisés sur le marché servi pouvant devenir alors une rigidité pour l'innovation des produits ou des *Business Models*.
Il était question dans cette recherche scientifique de donner le parallélisme entre les termes *Data mining* et l'*ECD,* termes qui existent depuis quelques décennies. Retenons que l'extraction de données peut faire partie de l'exploration de données lorsque l'objectif est de collecter et d'intégrer des données provenant de différentes sources. L'extraction de connaissances et *data mining* consiste à donner un sens aux grandes quantités de données, d'un certain domaine, capturées et stockées massivement par les entreprises d'aujourd'hui. En effet, la vraie valeur n'est pas dans l'acquisition et le stockage des données, mais plutôt dans notre capacité d'en extraire des rapports utiles et de trouver des tendances et des corrélations intéressantes pour appuyer les décisions faites par les décideurs d'entreprises et par les scientifiques. L'exploration de données, en tant que processus relativement complexe, consiste à découvrir des modèles pour donner un sens aux données et prédire l'avenir. Les deux requièrent des compétences différentes et une expertise, mais la popularité croissante des outils d'extraction de données et des outils d'exploration de données améliore considérablement la productivité et facilite grandement la vie de ses utilisateurs. Le choix de l'une ou l'autre approche dépend de la nature du problème à résoudre.

References

1. AGRAWAL, R., Srikant, R. (1995). Mining Sequential Patterns. In proceedings of the 11th International Conference on Data Engineering (ICDE'95), Tapei, Taiwan, March 1995.
2. BEN A, Safe Next: une approche systémique pour l'extraction de connaissances de données: application à la construction et à l'interpretation de scénarios d'accidents de la route. Thèse de doctorat présentée et soutenue publiquement le 17 janvier 2005, Ecole Centrale de Paris, France.
3. Brachman, R., and Anand, T. (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press.
4. D. HAND, MANNILA, H., P. SMYTH, Principles of Data Mining, MIT Press, Cambridge, 2001.
5. Fadi B., Extraction de connaissances d'adaptation en raisonnement à partir de cas, Thèse de doctorat présentée et soutenue publiquement le 20 novembre 2009, université Henri Poincaré – Nancy 1, département de formation doctorale en informatique, 2009, p.65.
6. FAYYAD, U. M. et al, The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 1996a.
7. HAN J., KAMBER M. Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2000, p.43-45
8. KOLSKI, C., EZZEDINE, H., ABED M. Cycle de développement du logiciel : des cycles classiques aux cycles enrichis sous l'angle des interactions homme machine, Editions Hermes, Paris, 2001

9. KURGAN, L. A., & Musilek, P. A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, Cambridge University Press, USA, Vol. 21(1), March 2006.
10. LEFEBURE R., G. VENTURI, Data Mining: Gestion de la relation client, Editions Eyrolles, Paris, 2001.
11. Michel F., Le Data Warehouse, le Data Mining, Eyrolles, Paris, 1996.
12. MOHAMMED A., MOUNIR B., CHRISTOPHE K., Convergence possible des processus du data mining et de conception-évaluation d'IHM: adaptation du modèle en U, 2005.
13. René L. et Gilles V., Le data mining, Eyrolles, Paris, 1998.
14. TANASA D., Web usage mining: contributions to Intersites Logs Preprocessing and sequential Pattern Extraction with low support, Thèse de Doctorat de l'Université, Nice Sophia-Antipolis, 2005.
15. USAMA F., GREGORY P. S., PADHRAIC S., Knowledge Discovery and Data Mining: Towards a Unifying Framework, 1996b.
16. YENDE RAPHAEL G., KASEKA V. KATADI et al, (2022), Signal performance optimization in the local area network trafic management in the DRC: Models for transmission networks. European Journal of Computer Science and Information Technology, Vol.10, No.5, pp.1-23
17. ZEMMOURI, E., BEHJA, H., & MARZAK, A. (2010). Modèle de gestion des connaissances dans un processus d'ECD. Journées Doctorales en Technologies de l'Information et de la Communication JDTIC'10, Fès, 15-17 Juillet 2010.
18. ZIGHED, D. A., RAKOTOMALALA, R. (2002) Extraction de connaissances à partir de données (ECD). Dans Techniques de l'Ingénieur.