État de l'art sur les aspects méthodologiques et processus en Knowledge Discovery in Databases

Cristina OPREAN

Janvier 27, 2010

Encadrant : Lenca Philippe
Equipe : Lab-STICC, pôle CID, équipe DECIDE

Table des matières

Table des figures

Liste des tableaux

Résumé *- Des nombreux progrès sont réalisés régulièrement pour améliorer les méthodes et algorithmes en fouille de données. En revanche, peu d'efforts -ou pas assez- ont été déployés sur les aspects méthodologiques. Ainsi, parmi les différentes méthodologies disponibles, la première version de CRISP-DM 1.0 est sans aucun doute la plus utilisée. C'est également probablement la plus ancienne, et la version 2.0 de CRISP-DM devant notamment tenir compte des évolutions de la fouille de données n'est toujours pas sortie. Pour autant, il y a un besoin méthodologique face à la complexité des études menées et aussi parfois face aux besoins des utilisateurs, qui ne sont pas toujours experts de la fouille de données. Le travail consiste ainsi à réaliser un état de l'art sur les aspects méthodologiques et processus en data mining et à établir une cartographie des qualités et besoins d'une "bonne" méthodologie. Cette cartographie sera mise en regard des processus et des workflows proposés au sein des logiciels de fouille les plus populaires. On mettra alors en évidence les points de convergence et de divergence entre les logiciels et les méthodologies dont la mise en oeuvre devrait pourtant être facilitée au sein des logiciels. On pourra alors proposer si nécessaire une méthodologie, à périmètre réduit, et un démonstrateur à base de workflow supportant cette méthodologie.*

Mots clés : Data Mining, Knowledge Discovery Databases, état de l'art, processus, méthodologie, Data Mining Engineering,Agile Discovery Environment, CRISP-DM.

1 Introduction

Il y a plusieurs définitions pour le processus de Data Mining, parmi lesquelles les suivantes sont les plus utilisées par la communauté scientifique : Data Mining est *un processus* d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans des données [1]. Data Mining est *un processus* complexe permettant l'indentification, au sein des données, des motifs valides, nouveaux, potentiellement intéressants et les plus compréhensibles possibles [2]. Data Mining est *un processus* interactif et itératif d'analyse d'un grand ensemble de données brutes et d'extraction des connaissances exploitables par des utilisateurs-analystes qui jouent un rôle central[3]. Il y a une grande confusion dans la littérature scientifique entre les termes de Knowledge Discovery in Databases (KDD) et Data Mining (DM, fouille de données), quand le terme de DM est utilisé avec le sens de KDD, même si en réalité DM est seulement une étape dans le processus de KDD.

Pour que le processus de Knowledge Discovery in Databases puisse devenir une technologie mature, il y a un grand besoin d'une approche méthodologique. Ainsi, mettre en place des normes va permettre d'avoir des méthodes standardisées et des procédures développées pour le Data Mining de telle sorte que ce processus sera automatisé et plus facile à utiliser par les différents types d'utilisateurs.

Après l'introduction du terme KDD dans les années '90 [4], un grand besoin est apparu afin d'implémenter des algorithmes pour résoudre les problèmes de recherche d'informations, plus précisément des connaissances dans de larges volumes de données. La multitude d'algorithmes développés et de fonctionnalités assurant un support pour toutes les activités qui impliquent le processus de Knowledge Discovery in Databases (visualisation des données, traitement des données, des outils pour faire de la statistique, etc.), ont été intégrées dans des logiciels qui permettent de suivre plus ou moins une approche méthodologique : Clementine [5] , IBM Intelligent Miner [6], Weka [7], Tanagra [8], DBMiner [9], Orange [10], RapidMiner [11]. Le problème est qu'aucun de ces outils ne donne une solution optimale qui pourrait être donnée seulement par une méthode standard idéale (mais qui n'existe pas pour le moment). Une méthodologie standard idéale devrait donner les meilleurs résultats après la première itération. Aucune autre itération n'étant nécessaire pour avoir le résultat final.

La bibliographie est réalisée dans le contexte du stage de recherche « Méthodologies et processus en Data Mining », au sein d'équipe DECIDE, au TELECOM Bretagne. Le but de ce stage est d'indentifier, premièrement, une cartographie des qualités et besoins d'une « bonne »méthodologie, qui sera mise en regard des processus et des workflows proposés au sein des logiciels de fouille de données les plus populaires. Deuxièmement, sera proposé, si nécessaire, une méthodologie, à périmètre réduit, et un démonstrateur à base de workflow supportant cette méthodologie.

Pour mieux comprendre le contexte du projet, l'étude bibliographique est structurée de la manière suivante : Section 2 fait d'abord une classification des processus et des méthodologies en : processus et méthodologies orientées vers la recherche et vers l'industrie, puis, présente une comparaison entre eux. Dans la section 3 des concepts propres d'un domaine mature (Software Engineering), seront proposés pour l'intégration dans le domaine de DM. Également, dans cette section, seront passées en revue de nouvelles tendances (XML, PMML, SOAP, UDDL, etc.) qui constituent des pistes des solutions pour le besoin actuel du Data Mining. Finalement, trouver une méthodologie standard pour le processus de Data Mining reste un problème ouvert, avec une solution qui peut répondre au besoin des utilisateurs d'être guidés pendant la réalisation de certaines tâches mentionnées dans les processus.

2 Une parallèle entre plusieurs méthodologies et processus

Dans la littérature scientifique il y a des confusions entre les termes « **processus**" »et « **méthodologie** ». Afin d'éviter le mélange entre les deux, O. Marban et. al [33] donnent les définitions prises du Software Engineering (SE). **Un processus** est un ensemble de tâches exécutées pour développer un certain élément. Le but d'un modèle de processus est de rendre le processus répétable, mesurable, maintenable. **Une méthodologie** peut être définie comme une instance du processus qui spécifie les tâches, les entrées, les sorties et comment réaliser les tâches. En résumant, les processus indiquent ce qu'il faut faire, tandis que les méthodologies montrent comment le faire.

2.1 État de l'art des modèles de processus

Un des problèmes identifiés par Qiang Yang et Xindong Wu dans [21] est l'importance de construire une méthodologie qui automatise la composition des opérations de Data Mining afin d'éviter les erreurs faites par les utilisateurs dans ce domaine. Une fois que certaines étapes vont être automatisées dans le processus de Data Mining, le travail humain va être beaucoup réduit. Il y a une palette assez grande de processus et de méthodologies dans le domaine de DM, mais aucune de ces méthodologies ou de ces processus ne donne le meilleur résultat. Une « bonne »méthodologie devrait guider l'utilisateur vers la solution optimale, sans avoir besoin de refaire certaines étapes pour obtenir les résultats attendus.

Le premier processus de base a été proposé par Fayyad (Knowledge Discovery in Databases - KDD)[22], processus amélioré ultérieurement. Ce processus est constitué de plusieurs étapes qui s'exécutent de manière séquentielle. Chaque étape commence après que les précédentes sont bien finalisées, parce qu'elle utilise les sorties des étapes précédentes et dans le cas où les résultats ne sont pas satisfaisants, certaines étapes peuvent être refaites.

En partant du modèle de Fayyad, dans les dernières années, les efforts ont été orientés pour trouver d'autres processus ou méthodologies de Data Mining. Même si tous les modèles ont été conçus en isolation, des progrès peuvent être observés. Beaucoup d'efforts sont faits pour trouver des standards dans le domaine de Data Mining, par exemple, ceux de SEMMA [23] et CRISP-DM [16]. Les deux sont utilisés dans le domaine industriel et définissent les étapes pour guider l'implémentation des applications de Data Mining. Dans la suite les différences et les similarités entre SEMMA, CRISP-DM et KDD seront présentées.

La plupart des modèles ne sont pas spécifiquement liés aux besoins industriels ou académiques. La limite entre les modèles utilisés dans la recherche et ceux utilisés dans l'industrie devient de plus en plus floue. Dans [40], Cios et Kurgan essaient de diviser les modèles en : modèles orientés vers la recherche et modèles orientés vers l'industrie. Dans la suite seront décrits brièvement les modèles les plus importants de chaque catégorie.

2.2 Modèles orientées vers la recherche

2.2.1 *Fayyad et al. (KDD)*

En 1996, Osama Fayyad a proposé un processus pour la fouille de données qui a bien répondu aux besoins d'entreprises, et qui est devenu rapidement très populaire. Knowledge Discovery in Databases (KDD) a comme but l'extraction des connaissances, des motifs valides, utiles et exploitables à partir des grandes quantités de données, par des méthodes automatiques ou semi-automatiques [12]. Comme nous pouvons observer dans l'image 1, ci-dessous , Data Mining(DM) représente seulement une étape dans le processus de KDD et implique l'interférence de plusieurs algorithmes d'exploitation des données, développant le modèle et découvrant des motifs inconnus.

Le processus de KDD est itératif et interactif. Le processus est itératif, ce qui signifie que parfois il peut être nécessaire de refaire les pas précédents. Le problème de ce processus, comme pour les autres présentés dans la section suivante, est le manque de guidage de l'utilisateur, qui ne choisit pas à chaque étape la meilleure solution adaptée pour ses données.

Les neuf étapes de KDD sont les suivantes [18] :

1. **Développer et comprendre le domaine de l'application** - c'est le pas initial de ce processus. Il prépare la scène pour comprendre et développer les buts de l'application.
2. **La sélection et la création d'un ensemble de données sur lequel va être appliqué le processus d'exploration.**
3. **Le prétraitement et le nettoyage des données** - cette étape inclut des opérations comme l'enlèvement du bruit et des valeurs aberrantes -si nécessaire, des décisions sur les stratégies qui vont être utilisées pour traiter les valeurs manquants, etc.

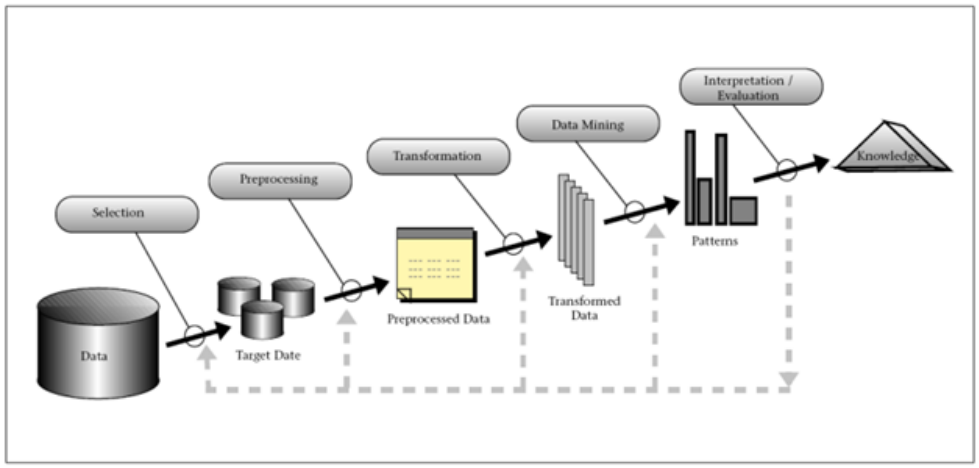

FIGURE 1 – Le processus de Knowledge Discovery in Databases [17]

4. **La transformation des données** - cette étape est très importante pour la réussite du projet et doit être adaptée en fonction de chaque base de données et des objectifs du projet. Dans cette étape nous cherchons les méthodes correctes pour représenter les données. Ces méthodes incluent la réduction des dimensions et la transformation des attributs.

   Une fois que toutes ces étapes seront terminées, les étapes suivantes seront liées à la partie de Data Mining, avec une orientation sur l'aspecte algorithmique.
5. **Choisir la meilleure tâche pour Data Mining** - nous devons choisir quel type de Data Mining sera utilisé, en décidant le but du model (par exemple : classification, régression, regroupement, etc.).
6. **Choisir l'algorithme de Data Mining** - dans cette étape nous devons choisir la méthode spécifique pour faire la recherche des motifs, en décidant quels modèles et paramétrés sont appropriés.
7. **Implémenter l'algorithme de Data Mining** - dans cette étape nous implémentons les algorithmes de Data Mining choisis dans l'étape antérieure. Peut être il sera nécessaire d'appliquer l'algorithme plusieurs fois pour avoir le résultat attendu.
8. **Evaluation** - inclut l'évaluation et l'interprétation des motifs découverts. Cette étape donne la possibilité de retourner à une des étapes précédentes, mais aussi d'avoir une représentation visuelle des motifs, d'enlever les motifs redondants ou non-représentatifs et de les transformer dans des termes compréhensibles pour l'utilisateur.
9. **Utiliser les connaissances découvertes** - inclut l'incorporation de ces connaissances dans des autres systèmes pour d'autres actions. Nous devons aussi mesurer l'effet de ces connaissances sur le système, vérifier et résoudre les conflits possibles avec les connaissances antérieures.

Applications du model : Le processus est itératif et il peut avoir plusieurs boucles qui s'exécutent entre n'importe laquelle des deux étapes. Le KDD est devenu lui-même un modèle pour les nouveaux modèles. Le modèle est incorporé dans le système commercial MineSet [19] et a été utilisé dans plusieurs domaines différentes : ingénierie, médicine, e-business, production, développement du logiciel, etc.

2.2.2 *Anand et Buchner*

Anand et Buchner ont développé une méthodologie hybride qui a été appliquée pour résoudre les problèmes de ventes croisées [28] et pour analyser les données de marketing sur Internet [29]. Cette méthodologie a huit étapes :

1. **Human Resources Identification** : Identifie les ressources humaines et leur rôle.
2. **Problem Specification** : Divise le projet en plusieurs tâches et chaque tâche sera résolue par une méthode particuliere de DM.
3. **Data Prospecting** : Analyse l'accessibilité et la disponibilité des données et sélectionne les attributs et un model de stockage.
4. **Domain Knowledge Elicitation**
5. **Methodology Specification** : Séléctionne la meilleur méthode de Data Mining ou utilise plusieurs méthodes de Data Mining.

6. **Data Preprocessing** : Suppression des valeurs aberrantes, des données bruitées, transformation et codage, etc.
7. **Pattern Discovery** : Découvre des motifs dans les données prétraitées.
8. **Knowledge Post-Processing** : Validation et visualisation de connaissances découvertes.

Application du model : Anand et Buchner fournit une analyse détaillée pour les étapes initiales du processus, mais n'inclut pas les activités nécessaires pour utiliser les connaissances découvertes et la documentation du projet.

2.2.3 *Les 5 A's*

Les 5 A's est une méthodologie développée par SPSS [27] pour donner une vision plus générale de l'analyse des données et le processus de Data Mining. L'importance de ce modèle est donnée par l'état « Automate », qui automatise le processus de Data Mining afin d'aider les utilisateurs non-experts à appliquer des méthodes antérieures à des nouvelles données. Cette méthode est très utile parce que les utilisateurs non-experts peuvent obtenir plus facilement de nouvelles connaissances.

Les cinq étapes de cette méthodologie sont : Asses, Acces, Analyse, Act et Automate. Le désavantage principal est que le 5 A's ne contient pas l'étape de « Data Understanding »qui est considérée très importante pour comprendre et tester la qualité des données pour éviter des éventuels problèmes qui peuvent apparaitre pendant le développement du projet [32].

Applications du model : Le modèle a été abandonné en 1999 par SPSS pour participer après au développement du processus CRISP-DM.

2.2.4 *Cios et al.*

Le modèle de Cios [30] a été proposé en adaptant le model CRISP-DM pour répondre aux besoins académiques. Ces processus utilisent des technologies comme : XML, PMML, SOAP UDDI et OLE BD-DM. Le modèle est constitué de six étapes :

1. **Understanding the problem domain** : pour apprendre la technologie et les solutions courantes, et déterminer les objectifs du domaine et du Data Mining.
2. **Understanding the data** : comprendre les mécanismes pour collectionner, explorer et vérifier les données.
3. **Preparation of the data** : décider quels vont être les algorithmes utilisés ; nettoyer et reformater la base de données. Cette étape est la plus gourmande en temps, mais sa qualité peut déterminer le succès ou l'échec du projet.
4. **Data Mining** : L'implémentation de différents algorithmes.
5. **Evaluation of the discovered knowledge** : interprétation des résultats et la recherche des améliorations possibles pour les algorithmes.
6. **Using the discovered knowledge** : La création d'un plan pour superviser l'implémentation des connaissances découvertes, la documentation du projet, l'extension de l'application dans d'autres domaines.

Cette méthode souligne et décrit explicitement les aspects itératifs et interactifs du processus. S'il y a des changements dans les étapes antérieurs, les étapes suivantes vont être aussi affectées, donc il est nécessaire d'avoir des boucles de rétroaction.

2.3 Modèles orientés vers l'industrie

2.3.1 *SEMMA*

L'Institute SAS [24] divise la fouille de données en cinq étapes représentées pas l'acronyme SEMMA [25] - « Sample, Explore, Modify, Model, Asses ». Cette méthodologie facilite pour les analystes de données la manière d'appliquer des techniques d'exploration statistique et de visualisation des données, de sélection et de transformation des variables prédictives les plus importantes, de modélisation des variables pour prédire les résultats et de confirmation de la précision du model.

1. **Sample (Echantillon des données)** - extrait des échantillons d'un vaste ensemble de données, en nombre suffisamment grand pour contenir l'information importante, mais assez petit pour être manipulé rapidement.
2. **Explore (Exploitation des données)** - cette étape consiste dans l'exploration des données en recherchant les tendances et les anomalies imprévues afin de mieux comprendre les données.

3. **Modify (Modifier)** - modifie les données en créant, en sélectionnant et en transformant les variables afin de s'axer sur le processus de sélection de modèles.
4. **Model (Modélisation)** - modèlise les données en permettant au logiciel de rechercher automatiquement une combinaison des données qui prédit de façon fiable le résultat souhaité. Il y a plusieurs techniques de modélisation : les réseaux de neurones, arbres de décision, modèles statistiques - l'analyse en composantes principales, l'analyse de séries temporelles, etc.
5. **Assess (Evaluer)** - évalue l'utilité et la fiabilité des résultats du processus de Data Mining et estime comment il va s'exécuter.

En évaluant les résultats obtenus à chaque étape du processus de SEMMA, nous pouvons déterminer le façon de modéliser les nouveaux problèmes déterminés par les résultats précédents, et donc de refaire la phase d'exploration supplémentaire pour le raffinement des données.

SEMMA est un cycle lui-même ; les étapes internes peuvent être réalisées de manière aléatoire si c'est nécessaire. La figure 2 illustre les tâches du projet de Data Mining et fait le lien avec les étapes de la méthodologie SEMMA [25].

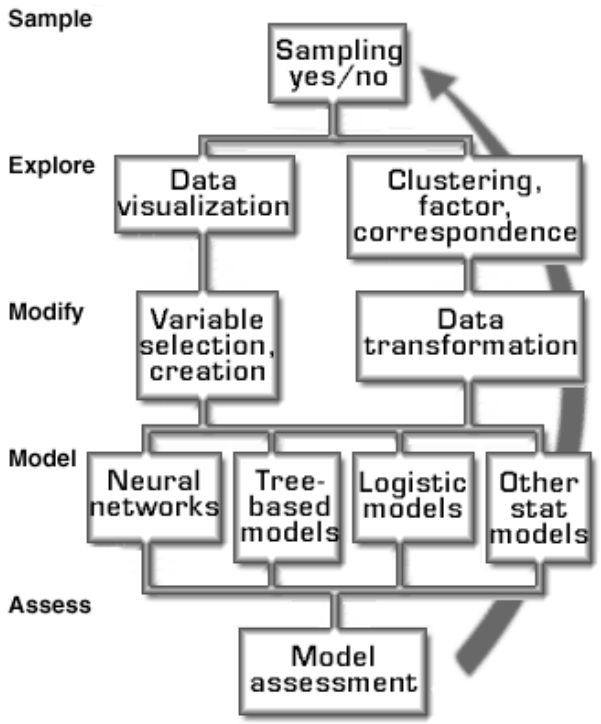

FIGURE 2 – Les étapes dans la méthodologie SEMMA [25]

Application du model : SAS est définie par SEMMA comme l'organisation logique d'outil SAS Entreprise Miner pour la réalisation des tâches de Data Mining. Enterprise Miner peut être utilisé comme une partie de n'importe quelle méthodologie itérative de Data Mining adoptée par le client. Une des différences entre KDD et SEMMA est que SEMMA est intégré dans l'outil Enterprise Miner et ils n'utilisent pas d'autres méthodologies, tandis que le KDD est un processus ouvert qui peut être appliqué dans plusieurs environnements.

2.3.2 *CRISP-DM*

CRISP-DM [16] a été développé en 1996 pour répondre aux besoins des projets industriels de Data Mining. CRISP-DM est décrit comme un processus hiérarchique constitué par plusieurs tâches, avec quatre niveaux d'abstraction : la phase, la tâche générique, la tâche spécialisée et l'instance du processus [16]. Peu de temps après son apparition, CRISP-DM est devenu le processus le plus utilisé pour le développement des projets de Data Mining, d'après les sondages faits par KDnuggets en 2002 [13], 2004 [14], 2007 [15].

CRISP-DM [16] est l'acronyme de **CR**oss-**I**ndustry **S**tandard for **D**ata **M**ining et contient un cycle de six étapes : **Business understanding** (La compréhension du business), **Data understanding** (La com-

préhension des données), **Data preparation** (La préparation des données), **Modeling** (La modèlisation), **Evolution** (L'évolution) et **Deployment** (Déploiement).

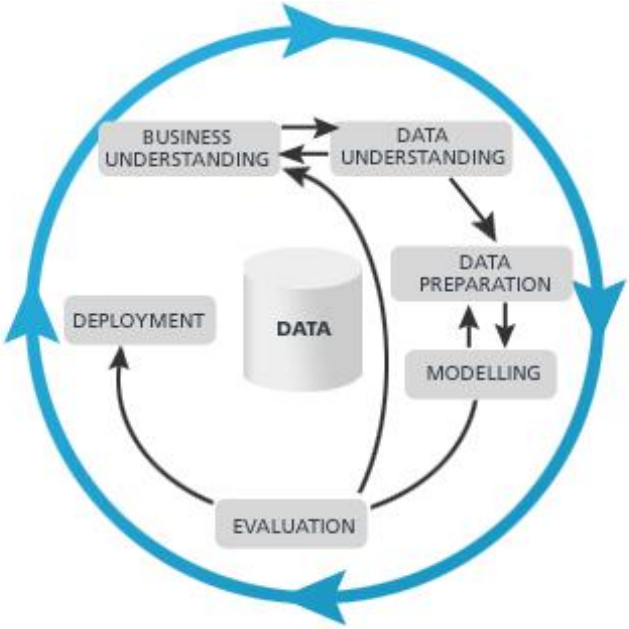

FIGURE 3 – Le cycle de vie du CRISP-DM [16]

1. **Business understanding** - cette phase initiale porte sur la compréhension des objectifs et des exigences du projet. Les connaissances acquises vont définir la problématique et le plan préliminaire pour accomplir ces objectifs.
2. **Data understanding** - elle commence avec une collection de données initiales et continue avec des activités afin de se familiariser avec les données, d'identifier les problèmes de qualité des données, de détecter des sous ensembles afin de construire des hypothèses pour l'information cachée.
3. **Data preparation** - cette phase contient toutes les activités nécessaires afin de construire la base de données finale.
4. **Modeling** - dans cette phase sont sélectées et appliquées plusieurs techniques sur les données et leur paramètres sont calibrés avec des valeurs optimales. Comme il y a plusieurs formes spécifiques de construction des données, la plupart du temps il est nécessaire de revenir une étape en arrière.
5. **Evaluation** - à ce niveau le(s) modèle(s) sont évalué(s) et les étapes suivies pour la construction du modèle sont réévaluées pour s'assurer que le projet respecte les objectifs du business, définis au début du projet.
6. **Deployment** - La création du modèle ne représente pas la fin du projet. Même si le but initial du projet est d'augmenter les connaissances de données, les connaissances acquises ont besoin d'être organisées et présentées d'une manière utilisable par le client.

Cette séquence de phases n'est pas obligatoire. On peut aller entre les phases, fait suggéré dans la figure 3 par la flèche qui indique les plus importantes et fréquentes dépendances entre les phases. Le problème de ce processus est le fait qu'il est itératif, nous devons revenir sur certaines étapes pour obtenir le résultat voulu. CRISP-DM ne guide pas l'utilisateur sur comment les tâches doivent être réalisées, mais le modèle est facile à comprendre et très bien documenté. En général, CRISP-DM est un processus extensivement appliqué dans le domaine industriel.

Après l'apparition de la première version de CRISP-DM, les applications de Data Mining ont beaucoup changées. Les nouveaux enjeux et exigences sont caractérisés par :

- Plusieurs types des données disponibles et aussi de nouvelles techniques pour manipuler, analyser et combiner les données.
- Des exigences concernant le facteur d'échelle et le déploiement dans des environnements en temps réel.
- Le besoin d'explorer des bases des données de grande échelle, etc.

Pour répondre aux nouveaux besoins du Data Mining, le processus CRISP-DM sera amélioré dans une nouvelle version CRISP-DM 2.0 [16]. Normalement cette version devait apparaitre en 2007, mais elle n'était pas encore finalisée.

Application du model : CRISP-DM a été utilisé dans de domaines comme : ingénierie, médicine, ventes, marketing, etc. et a été inclus dans le système de KDD Clementine [5].

2.3.3 *Cabena et al.*

Cabena et al. [32] définissent Data Mining comme un processus d'extraction des connaissances non valides et utiles à partir des larges bases de données afin d'utiliser ces informations pour prendre des décisions importantes. La structure de Cabena contient 5 étapes, mais il n'y a pas une grande différence entre ce processus et le processus KDD. Les étapes sont les suivantes :

1. **Business objectives determination** : Comprendre les problèmes de business et définir les objectifs de Data Mining.
2. **Data preparation** : Identifier les ressources des données externes ou internes et sélectionner un échantillon de données pour une certaine tâche. A ce niveau la qualité des données va être vérifiée et améliorée et les méthodes de DM qui seront utilisées dans les pas suivants vont être déterminées .
3. **Data Mining** : Application des méthodes sélectionnées sur des données prétraitées.
4. **Analysis of results** : Interprétation des résultats obtenus en utilisant des techniques de visualisation.
5. **Assimilation of knowledge** : Utilisation des connaissances obtenues dans les systèmes de l'organisation en expliquant comment ces connaissances seront exploitées.

Application du model : Cabena est le première modèle orienté vers le business, mais les auteurs n'ont pas donné assez des détails sur les boucles de rétroaction, d'où l'incomplétude de ce modèle. Ce processus est plutôt utilisé dans le domaine de marketing et des ventes.

2.4 Comparaison les méthodologies et processus

Pour synthétiser ce qui a été présenté jusqu'à maintenant, dans cette sous-section nous proposons une comparaison entre les processus et les méthodologies mentionnés dans les sous-chapitres antérieurs. Cette synthèse sous forme de tableaux est à la base des études faites par Kurgan [31] et Marban [32]. La comparaison est faite en respectant certains critères comme : le domaine d'application, le nombre d'étapes, les étapes les unes en fonction des autres - on peut visualiser dans les tableaux 1 et 2 la correspondance/non-correspondance entre les étapes, les approches liées, le support logiciel, les avantages et les désavantages de chacun de processus ou méthodologies.

Comme le tableau 1 l'indique, il y a plusieurs caractéristiques communes entre les processus. La plupart des processus suivent la même séquence d'étapes et parfois ils utilisent des pas similaires ; les modèles impliquent des tâches de « Data Preparation »trop complexes et lourdes. Une autre similitude importante entre les processus est l'aspect itératif, les boucles nécessaires entre certaines ou toutes les étapes. Les processus se différentient plutôt pour la première étape de « Business Understanding »et la dernière étape de « Using discovery knowledge », qui n'existent pas dans tous les processus (SEMMA, 5A's, Anand et Buchner). Le modèle 5A's contient une étape en plus, d'automatisation du processus de DataMining, pour que les utilisateurs non-expérimentés puissent appliquer des modèles antérieurs sur des nouvelles données.

Nous pouvons conclure que CRISP-DM et KDD sont les processus les plus utilisés pour développer des projets de Data Mining (on peut voir dans le tableau 1 que les approches liées pour les processus présentés sont CRISP-DM et KDD), mais leur utilisation est diminuée a cause des autres méthodes internes, développées dans le cadre des équipes (comme par exemple SEMMA). Le fait que CRISP-DM et KDD ne sont plus assez utilisés est dû au fait qu'ils définissent ce qu'il faut faire et pas comment le faire. A cause de ça, plusieurs équipes ont commencé à développer leur propre méthodologie de leur côté.

	KDD (1996)	**SEMMA (1996)**	**5A's (1996)**	**Cabena et al. (1997)**	**Anand et Buchner(1998)**	**CRISP-DM (2000)**	**Cios et al. (2000)**
Domaine	Recherche	Industriel	Recherche	Industriel	Recherche	Industriel	Recherche
Nombre d'étapes	9	5	5	5	8	6	6
	1. Learning application domain		1. Asses	1.Business Objectives Determination	1.Human resources identification	1.Business understanding	1.Understanding the problem's domain
					2.Problem specification		
	2. Creating a target data set	1. Sample		2. Data preparation	3. Data prospecting	2. Data understanding	2. Understanding the data
					4. Domain knowledge elicitation		
	3. Data cleaning and prepocessing	2. Explore	2. Acces		5. Methodology identification	3. Data preparation	3. Preparation of the data
	4. Data reduction and projection	3. Modify			6. Data preprocessing		
	5. Choosing the function of Data Mining	4. Model	3. Analyse	3. Data Mining	7. Pattern discovery	4. Modeling	4. Data Mining
	6. Choosing the Data Mining algorithm						
	7. Data Mining						
	8. Interpretation	5. Asses	4. Act	4. Analysis of results	8. Knowledge post-processing	5. Evaluation	5. Evaluation of the discovered knowledge
	8. Using discovered knowledge	-	-	5. Assimilation of knowledge	-	6. Deployment	6. Using the discovered knowledge
			5. Automate				
Approches liées	KDD	KDD	-	KDD	KDD	CRISP-DM	CRISP-DM
Support logiciel	MineSet, KDB2000, VidaMine	SAS Enterprise Miner	-	-	-	Clementine	Grid Miner-Core

TABLE 1 – Comparaison entre les étapes des processus et méthodologies [31], [32]

	KDD (1996)	**SEMMA (1996)**	**5A's (1996)**	**Cabena et al. (1997)**	**Anand et Buchner(1998)**	**CRISP-DM (2000)**	**Cios et al. (2000)**
Avantages	-Le processus est interactif et itérative -peut traiter des grandes quantités de données -description technique détaillé sur l'analyse des données	-on peut procéder de retour à la phase d'exploration pour un raffinement supplémentaire des données.	-l'automatisation du processus de Data Mining -les utilisateurs non-expérimentés peuvent appliquer des modèles antérieurs sur des nouvelles données	- le modèle est facile à comprendre même par les personnes non-spécialistes	- fournie une analyse détaillée sur les étapes initiales du processus -met l'accent sur la nature itérative du modèle, où les experts examinent les connaissances après la dernière étape et ils peuvent décider d'affiner et de relancer une partie ou l'ensemble du processus	-modèle hiérarchique - la séquence des phases n'est pas stricte -est la plus utilisé méthodologie et processus de Data Mining -a une perspective plus large sur les aspects du business -est plus proche du du concept réel du projet -est le plus utilisé par les utilisateurs novices (facile de comprendre) -très bien documenté	- inclut une description pour les premières étapes plus générale et orientée sur la recherche -l'introduction des plusieurs mécanismes de feedback -on peut aller en arrière du « Data Preparation »au « Data Understanding » -met l'accent et décrit explicitement la nature itérative et interactive du processus
Inconvénients	-omet des aspects importants dans cadre du business -parfois c'est indispensable de faire beaucoup des boucles de retour non-nécessaires -les auteurs ne donnent pas assez des informations sur les itérations entre les étapes	-n'inclut pas une étape explicite pour utiliser les connaissances découverts	-n'établie pas des alternatifs pour appliquer les modèles construits ou les connaissances obtenues -n'inclut pas l'étape de « Data Understanding »	-il n'y a pas beaucoup d'informations sur la nature itérative du processus -n'inclut pas la documentation sur les connaissances obtenus	-n'inclut pas les activités nécessaires pour pour utiliser les connaissances obtenues	-des informations limitées sur les boucles de retour -le model a besoin d'être mis à jour -décrie ce qu'on doit faire et pas comment on doit faire	-orienté sur la recherche

TABLE 2 – Les avantages et les inconvénients de processus et méthodologies [31], [32]

3 De nouvelles perspectives

Comme la taille et la complexité des projets actuels sont incomparables à celles d'il y a dix ans, les solutions présentées dans les chapitres antérieures ne sont plus adaptées pour les nouveaux problèmes. Dans cette section, nous avons pris en considération de nouvelles approches pour construire une méthodologie standardisée pour les projets de Data Mining.

3.1 Software Engineering vs. Data Mining

L'histoire du développement du KDD et du Software Engineering (SE) est similaire. Après l'introduction du processus de Knowledge Discovery in Databases au début des années '90, il y avait une explosion des implémentations des algorithmes pour faire de la recherche des connaissances dans des larges bases de données. L'année 2000 a été l'année d'apparition du CRISP-DM qui est devenu en un court laps de temps la méthodologie la plus utilisée, un standard « de facto ». Mais le nombre de projets dans le domaine de Data Mining augmente rapidement et les processus et méthodologies existantes ne sont pas appropriés pour les projets actuels de Data Mining. Au début, Software Engineering a été aussi orienté sur le développement des algorithmes et des langages de programmation. Mais après la crise du software déclenchée à cause de l'inaccomplissement des objectifs du client, de la productivité faible, etc., la communauté en charge avec le développement du software a décidé d'investir plus pour trouver des standards dans le domaine de l'ingénierie et de les appliquer dans le domaine du software. Cette approche a amélioré beaucoup le développement du software et a baissé le taux d'échec des projets d'environ 15-20% [34].

Maintenant, le Data Mining est le domaine dans lequel les entreprises investissent le plus, même si pas tous les projets sont bien finis ; le taux d'échec des projets en Data Mining est très élevé : environ 60%. Même si CRISP-DM est une variante améliorée, face aux autres méthodologies existantes, il n'est pas assez mûr pour faire face à la complexité des projets. Mais, si nous appliquons les mêmes principes de l'ingénierie dans le domaine de Data Mining comme dans le Software Engineering, nous attendons à avoir des résultats similaires concernant l'évolution du Data Mining. O. Marban, J. Segovia, E. Menesalvas et C. Fernandez-Baizan proposent un modèle de processus de Data Mining Engineering [33] (voir aussi la figure 4) qui est à la base l'idée de Software Engineering et le processus CRISP-DM.

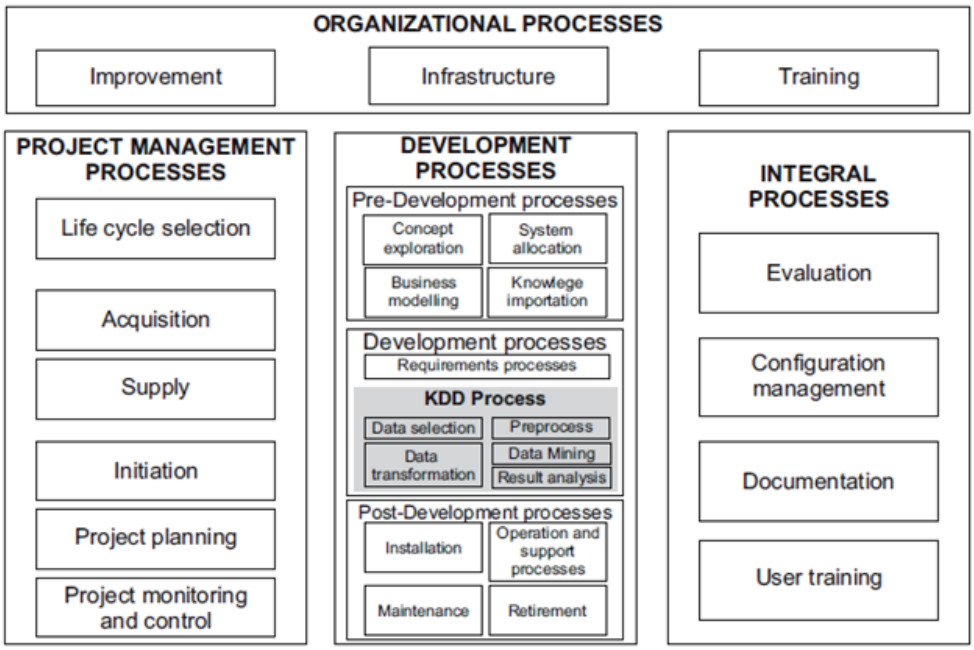

FIGURE 4 – Data Mining Engineering process [33]

Nous pouvons constater qu'il y a plusieurs étapes qui existent dans le SE et qui sont importantes pour le développement des projets de DM, mais elles n'existent pas dans la méthodologie CRISP-DM. Les auteurs [33] ont proposé de considérer les processus et les tâches inclus dans CRISP-DM et de les organiser comme les processus de SE et finalement d'ajouter les activités clés (les activités qui manquent du CRISP-DM : les processus de projet management, les processus d'intégration et les processus organisationnels) nécessaires

pour le développement d'un nouveau processus. Cette comparaison montre que le processus KDD n'est pas très bien organisé et qu'il lui manque plusieurs étapes qui sont importantes pour la réussite du projet.

Cette nouvelle méthodologie basée sur SE peut être un point de départ pour le futur développement du DM. Il existe aussi d'autres exemples des interactions entre les techniques de SE et DM comme les Service Oriented Architecture (SOA) et Model-View-Controller (MVC) qui sont utilisés pour le soutien des projets de DM. La méthodologie proposée par les auteurs est plutôt une base pour la nouvelle génération des projets de DM « Agile »ou « Extreme ».

3.2 Agile Discovery Environment

Agile software développement est un group de méthodologies de software basées sur le développement itératif et incrémental. Les demandes et les solutions évoluent à travers de la collaboration entre les équipes. Les plus importantes valeurs des méthodologies agiles pour le développement du software, ont été exprimées dans le « Manifesto for Agile Software Development »[35] en 2001. Pour développer plus le domaine de Data Mining, l'idée d'Agile Data Mining peut aider à améliorer le domaine de Data Mining. Il y a deux principes de base qui peuvent contribuer :

- Des interactions à travers des processus et des outils
- Répondre aux changements à la suite d'un plan

« Extreme Programming »(XP) est une des méthodologies agiles les plus utilisées. Comme pour le processus de Software Engineering, les méthodologies agiles, le XP en particulier, ont beaucoup aidé, mais il reste à trouver une manière de les appliquer en Data Mining Engineering aussi.

Les éléments essentiels pour avoir un « Agile Discovery Environment (ADE) »[38] :

1. L'ADE doit offrir un support pour la responsabilité humaine (le modèle doit être compris par l'explorateur, permettre à l'explorateur de modifier n'importe quel modèle comme il veut, offrir un support pour l'affinement du modèle de manière incrémentale).
2. L'environnement doit stimuler l'intuition de l'explorateur (beaucoup de techniques pour la visualisation des données, la visualisation du modèle, la comparaison visuelle des modèles).
3. L'ADE doit offrir un support pour l'apprentissage humain (pour l'évolution des connaissances de l'explorateur)
4. L'ADE doit offrir un support pour la communauté d'explorateurs (offrir un support pour la communication et le partage des connaissances)
5. L'ADE doit permettre à l'explorateur de garder sont ouverture d'esprit (offrir un support pour plusieurs modèles et aider l'utilisateur en choisissant le type approprié).

Les outils modernes pour faire le Data Mining, comme SPSS Clementine[5], Weka[7], RapidMiner[11], Tanagra[8], SAS[23], ont beaucoup d'aspects positifs (l'automatisation pour faire la répétition de certains pas dans le processus de Data Mining, des bonnes techniques de visualisation des résultats et des données, support pour l'exploration interactive du modèle, support pour l'interchangement des modèles, etc.), mais aussi des aspects négatifs (le processus de Data Mining est terminé après que le modèle a été construit ; les modèles sont limités a une seule représentation de la connaissance ; la modification des modèles à la main par l'utilisateur est limité) ce qui démontrent que ces outils sont très loins d'être agiles.

Comme R. Rokotomalala a dit en [36], pour le moment « il n'y a pas de bons ou mauvais logiciels. Il y a des logiciels qui répondent ou non à des spécifications ». Ainsi, l'utilisateur est obligé d'adapter ses besoins aux possibilités des logiciels mis à sa disposition. Donc, il faut construire une méthodologie qui permet la conception des logiciels qui ne contraint pas l'utilisateur, mais qui lui offre du support pour ses besoins.

3.3 Tendances dans le domaine de DM

Le domaine de Data Mining et Knowledge Discovery est un domaine qui est devenu un des plus répandus domaines de l'informatique à cause de ce grand besoin d'outils qui font le traitement des données pour des larges bases des données. Il y a des versions de processus de KDD qui ont à la base des technologies comme XML (Extended Markup Language), PMML (Predictive Model Markup Language), SOAP (Simple Object Access Protocol), UDDI (Universal Description Language), XML-RPC (XML Remote Procedure Call). Ces technologies sont utilisées pour offrir des solutions pour l'automatisation du processus KDD [37], pour construire un modèle plus facile à utiliser, semi-automatique et flexible. Elles permettent l'interopérabilité entre plusieurs outils de DM, des bases des données et des entrepôts des connaissances, l'automatisation et l'intégration des plusieurs tâches de DM.

La communauté orientée vers le business essaie d'intégrer ces technologies dans le processus de KDD. Il y a plusieurs outils qui ont à la base ces technologies : Intelligent Miner d'IBM, Oracle Darwin d'Oracle,

Taverna, etc. dans le but d'automatiser les étapes du processus. Mais ils ne résolvent que partiellement le problème de semi-automatisation du processus.

4 Conclusion

Dans la première partie de la bibliographie nous avons présenté un état de l'art des méthodologies qui sont encore utilisées (La deuxième section), même si elles sont moins adaptées aux nouveaux besoins des projets de DM. Après une synthèse plus approfondie sur les problèmes et les nouvelles tendances dans le domaine de DM, nous avons identifié de nouvelles perspectives pour construire une méthodologie standard. Le but de construire une méthodologie qui va être suivi dans les projets de DM, est d'être capable de modéliser à travers ces projets, le plus facilement possible, les problèmes du monde réel. Cette réalisation va constituer un pas important vers la maturation du domaine de DM.

Au début de mon stage, en mars, mon travail sera de proposer une cartographie des qualités et besoins d'une "bonne" méthodologie qui sera mise en regard des processus et workflows proposés au sein des logiciels les plus populaires. Ma mission, dans la deuxième partie de mon stage, au sein de l'équipe DECIDE, est de mettre en évidence les points de convergence et de divergence entre les logiciels et méthodologies, et en utilisant les nouvelles technologies , de proposer, si nécessaire, une méthodologie, a périmètre réduit, et un démonstrateur à base de workflow supportant cette méthodologie.

Références

[1] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining.*, AAAI/MIT Press,editors, 1996.

[2] Kodratoff, Y., Napoli, A., and Zighed, D., *Bulletin de l'association française d'intelligence artiffcielle, extraction de connaissances dans des bases de données*, 2001.

[3] Parsaye, K., Chignell, M., Khosha
an, S., and Wong, H., *Intelligent Databases ; Object-Oriented, Deductive Hypermedia Technologies*, John Wiley & Sons, 1989.

[4] Piatetsky-Shapiro, G. : *Report on the AAAI-91 Workshop on Knowledge Discovery in Databases.*, IEEE Expert, Vol. Technical Report 6, 1991

[5] Khabaza, T., Shearer, C. : *Data Mining with Clementine*, IEEE Colloquium on Knowledge Discovery in Databaes. IEEE Digest No. 1995/021(B), London.

[6] IBM. DB2 Intelligent Miner
http ://www-01.ibm.com/software/data/iminer/ [Cited 16.09.2010]

[7] Witten, I. H. and Frank, E. : *Data Mining : Practical Machine Learning Tools with Java implementations*,Morgan Kaufmann, 2005.

[8] Rakotomalala, R., *TANAGRA : un logiciel gratuit pour l'enseignement et la recherche*, in Actes de EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005.

[9] Han, J., Fu, Y., Wang, W., Chiang, J., Zaiane, Osmar R. , Koperski, K. : *DBMiner : Interactive Mining of Multiple-Level Knowledge in Relational Databases*, SIGMOD 1996, 6/96, Montreal, Canada

[10] Orange
http ://www.ailab.si/orange/ [Citation : 11.01.2011.]

[11] Rapid Miner
http ://rapid-i.com/content/view/181/190/ [Citation : 11.01.2011.]

[12] Wikipedia
http ://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es [Citation : 06.09.2010.]

[13] KDnuggets Polls 2002
http ://www.kdnuggets.com/polls/2002/methodology.htm. [Cited : 06.12.2010.]

[14] KDnuggets Polls 2002
http ://www.kdnuggets.com/polls/2002/methodology.htm. [Cited : 06.12.2010.]

[15] KDnuggets Polls 2002
http ://www.kdnuggets.com/polls/2002/methodology.htm. [Cited : 06.12.2010.]

[16] Daimler-Chrysler Project Overview, *CRISP-DM*, 1996
http ://www.crisp-dm.org/Overview/index.html [Cited : 07.09.2010.]

[17] Fayyad, U.M : *Data Mining in the KDD Environment*, http ://www.data-mining-blog.com/data-mining/data-mining-kdd-environment-fayyad-semma-five-sas-spss-crisp-dm/ [Cited : 07.09.2010.]

[18] Maimon, O., Rokach, L. : *The Data Mining and Knowledge Discovery Handbook*, Springer, Tel Aviv , 2005.

[19] Brunk, C., James, K., Kohavi, Mountain Viez, R. : *MineSet : An integrated System for Data Mining*, KDD-97 Proceedings, AAAI, 1997.

[20] Ranjan, J., Bhatnagar, V. : *A Review of Data Mining Tools In Customer Relationship Management*, Journal of Knowledge Management Practice, Delhi, 2008, Vol. 9.

[21] Yang, Q., Wu, X. : *10 Challenging problems in Data Mining research*, Journal of Information Technology and Decision Making, World Scientific Publishing Company, Vol. 5, 2005

[22] Fayyad, U.M et al. :
From data mining to knowledge discovery : an overview, AAAI Press/The MIT Press, Vol. Advances in knowledge discovery and data mining, 1996.

[23] SAS Enterprise Miner,
http ://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html. [Cited : 07.09.2010.]

[24] SAS Institute Inc.,
http ://www.sas.com [Cited : 07.09.2010.]

[25] Data Mining and the Case for Sampling. SAS,
http ://sce.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf. [Cited : 07.09.2010.]

[26] IBM SPSS ModelerProfessional,
http ://www.spss.com/software/modeling/modeler-pro/ [Cited : 21.09.2010.]

[27] SPSS Website,
http ://www.spss.com [Cited : 21.09.2010.].

[28] Anand, S.S., Patrick, A.R., Hughes, J.G., Bell ,D.A. : *A Data Mining Methodology for Cross Sales*, Knowledge-Based Systems, 1998, Vol. 10.

[29] Anand, S. and Buchner. : *Decision Support using Data Mining*, Londres : Finacial Time Management, 1998.

[30] Kurgan, Lukasz A., Cios,Krzysztof J. , Tadeusiewicz, R., Ogiela, M. , Goodenday, L. : *Knowledge discovery approach to automated cardiac SPECT diagnosis*, Artificial Intelligence in Medicine, 2001, Vol. 23/2.

[31] Kurgan, Lukasz A. and Musilek, P. : *A survey of Knowledge Discovery and Data Mining process model*, The Knowledge Engineering Review, Cambridge University Press, Cambridge, 2006, Vol. 21. 1.

[32] Marban, O., Mariscal, G., Fernandez, C : *A survey of data mining and knowledge discovery process models and methodologies*, The Knowledge Engineering Review, Cambridge University Press, Cambridge, 2010, Vol. 25.

[33] Marban, O., Segovia, J., Menesalvas, E., Fernandez-Baizan, C : *Toward data mining engineering : A software engineering approach*, ScienceDirect, Information Systems, Madrid, 2008, Vol. 34.

[34] Incorporated, Galorath : Software Project Failure Costs Billions, Better Estimation & Planning Can Help. SEER
http ://www.galorath.com/wp/software-project-failure-costs-billions-better-estimation-planning-can-help.php. [Cited : 03.10.2010.]

[35] Beck, Kent ; et al. *Manifesto for Agile Software Development*, Agile Alliance, 2001, Retrieved 2010-06-14,
http ://agilemanifesto.org/iso/en/ [Cited : 03.10.2010.].

[36] Rokotomalala, R., *K-Means - Comparaison de logiciels*,
http ://tutoriels-data-mining.blogspot.com/search/label/Classification-Clustering, 2008, [Cited : 21.01.2011.]

[37] Kurgan, L., Cios, K. : *Trends in Data Mining and Knowledge Discovery*, In : Pal N.R., Jain, L.C. and Teoderesku, N. (Eds.), Knowledge Discovery in Advanced Information Systems, Springer, 2005

[38] Grigoriev, Peter A., Yevtushenko, Serhiy A. : *Elements of an Agile Discovery Envoronment*, Darmstadt : Springer-Verlag, 2003.

[39] Jackson, J. : *Data Mining : A Practical Perspective*, Americas Conference on Information Systems (AMCIS), South-Carolina, 2001.

[40] Cios, K.J., Pedrycz, W., Swinarski, R.W., Kurgan, L.A. : *Data Mining, A Knowledge Discovery Approach*, New York : Springer, pp. 11-24, 2007 .