

**République Algérienne Démocratique et Populaire**  
**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

Université HADJ LAKHDAR – BATNA  
Faculté des sciences de l'ingénieur  
Département d'informatique

N° d'ordre : .....  
Série : .....

*Mémoire*

*En vue de l'obtention du diplôme de Magister en informatique*

*Spécialité : Sciences et Technologies de l'Information et de la Communication (STIC)*

*Option: Systèmes d'Information et de Connaissance (SIC)*

**Une plate forme orientée agent pour le data mining**

*Par :*

*Melle.* **CHAMI Djazia**

Présenté le : / /

*Devant le jury composé de :*

*Président* : Dr. BELATTAR Brahim, Maître de Conférence à l'université de Batna.

*Rapporteur* : Dr. KAZAR Okba, Maître de Conférence à l'université de Biskra.

*Examineur* : Dr. ZIDANI Abdelmadjid, Maître de Conférence à l'université de Batna.

*Examineur* : Dr. BILAMI Azeddine, Maître de Conférence à l'université de Batna.

**2009-2010**

**République Algérienne Démocratique et Populaire**  
**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

Université HADJ LAKHDAR – BATNA  
Faculté des sciences de l'ingénieur  
Département d'informatique

N° d'ordre : .....  
Série : .....

*Mémoire*

*En vue de l'obtention du diplôme de Magister en informatique*

*Spécialité : Sciences et Technologies de l'Information et de la Communication (STIC)*

*Option: Systèmes d'Information et de Connaissance (SIC)*

**Une plate forme orientée agent pour le data mining**

*Par :*

*Melle.* **CHAMI Djazia**

Présenté le : / /

*Devant le jury composé de :*

*Président* : Dr. BELATTAR Brahim, Maître de Conférence à l'université de Batna.

*Rapporteur* : Dr. KAZAR Okba, Maître de Conférence à l'université de Biskra.

*Examineur* : Dr. ZIDANI Abdelmadjid, Maître de Conférence à l'université de Batna.

*Examineur* : Dr. BILAMI Azeddine, Maître de Conférence à l'université de Batna.

**2009-2010**

## Résumé :

Le Data Mining est une technologie dont le but est la valorisation de l'information et l'extraction de connaissances d'un grand nombre de données. Cette technologie qui est devenue un outil important pour améliorer les revenus des entreprises rencontre des problèmes dus à la quantité énorme de données à exploiter. Pour y parvenir, nous avons fait appel au paradigme de système multi agents pour distribuer la complexité sur plusieurs entités autonomes appelés agents. Le résultat de notre recherche est une modélisation d'un système de clustering basé agents. Notre approche qui est une approche cognitive se base sur la connaissance pour assurer le bon fonctionnement du système et sa fiabilité.

Mots clés: data mining, système multi-agents.

## ملخص:

تنقيب البيانات هي تكنولوجيا الهدف منها تجميع المعلومات و استخراج المعرفة من عدد كبير من البيانات. هذه التكنولوجيا التي أصبحت أداة هامة لتحسين أرباح الشركات تعاني من مشاكل نظرا للكمية الهائلة من البيانات الواجب تنقيبها. من اجل معالجة هذه المشاكل استعنا بالأنظمة متعددة الوكلاء من أجل توزيع المهام على عدد من الكيانات المستقلة. نتيجة بحثنا هي بناء نظام متعدد الوكلاء لتجميع البيانات. منهجنا منهج إدراكي قائم على المعرفة لضمان الأداء السليم للنظام وموثوقيته .

الكلمات الرئيسية: تنقيب البيانات، الأنظمة متعددة الوكلاء.

# SOMMAIRE

<b><u>INTRODUCTION GENERALE.....</u></b>	<b><u>1</u></b>
--	-----------------

## **CHAPITRE I: GENERALITE SUR LE DATA MINING**

<b><u>1 INTRODUCTION .....</u></b>	<b><u>4</u></b>
------------------------------------	-----------------

<b><u>2 DEFINITION DU DATA MINING .....</u></b>	<b><u>4</u></b>
---	-----------------

<b><u>3 MOTIVATIONS .....</u></b>	<b><u>5</u></b>
-----------------------------------	-----------------

<b><u>4 DATA MINING SUR QUELS TYPES DE DONNEES ?.....</u></b>	<b><u>8</u></b>
---	-----------------

<b>4.1 LES FICHIERS PLATS .....</b>	<b>8</b>
-------------------------------------	----------

<b>4.2 LES BASES DE DONNEES RELATIONNELLES.....</b>	<b>8</b>
---	----------

<b>4.3 LES DATA WAREHOUSES .....</b>	<b>8</b>
--------------------------------------	----------

<b>4.4 LES BASES DE DONNEES TRANSACTIONNELLES .....</b>	<b>10</b>
---	-----------

<b>4.5 LES BASES DE DONNEES MULTIMEDIA .....</b>	<b>11</b>
--	-----------

<b>4.6 LES BASES DE DONNEES SPATIALES .....</b>	<b>11</b>
---	-----------

<b>4.7 LES BASES DE DONNEES DE SERIES TEMPORELLES .....</b>	<b>11</b>
---	-----------

<b>4.8 LE WORLD WIDE WEB .....</b>	<b>12</b>
------------------------------------	-----------

<b><u>5 LES TACHES DU DATA MINING.....</u></b>	<b><u>13</u></b>
--	------------------

<b>5.1 LA CLASSIFICATION .....</b>	<b>13</b>
------------------------------------	-----------

<b>5.2 L'ESTIMATION .....</b>	<b>14</b>
-------------------------------	-----------

<b>5.3 LA PREDICTION .....</b>	<b>14</b>
--------------------------------	-----------

<b>5.4 LE GROUPEMENT PAR SIMILITUDE.....</b>	<b>15</b>
--	-----------

<b>5.5 L'ANALYSE DES CLUSTERS .....</b>	<b>15</b>
---	-----------

<b>5.6 LA DESCRIPTION.....</b>	<b>15</b>
--------------------------------	-----------

<b><u>6 LES ETAPES DU PROCESSUS DE DATA MINING .....</u></b>	<b><u>16</u></b>
--	------------------

<b><u>7 TECHNIQUES DU DATA MINING.....</u></b>	<b><u>17</u></b>
--	------------------

<b>7.1 LES RESEAUX DE NEURONES .....</b>	<b>17</b>
--	-----------

<b>7.1.1 AVANTAGES ET INCONVENIENTS .....</b>	<b>18</b>
---	-----------

<b>7.2 LES ARBRES DE DECISION .....</b>	<b>18</b>
---	-----------

<b>7.2.1 LES ALGORITHMES D'INDUCTION DES ARBRES DE DECISION .....</b>	<b>19</b>
---	-----------

<b>7.2.2 AVANTAGES ET INCONVENIENTS .....</b>	<b>19</b>
---	-----------

<b>7.3 LES ALGORITHMES GENETIQUES .....</b>	<b>19</b>
---	-----------

<b>7.3.1 PRINCIPE DE BASE DES ALGORITHMES GENETIQUES .....</b>	<b>20</b>
--	-----------

<b>7.3.2 CODAGE D'UN ALGORITHME GENETIQUE .....</b>	<b>20</b>
---	-----------

7.3.3	AVANTAGES ET INCONVENIENTS .....	21
<b>7.4</b>	<b>LES REGLES ASSOCIATIVES.....</b>	<b>22</b>
7.4.1	LES ALGORITHMES D'INDUCTION DES REGLES ASSOCIATIVES .....	22
7.4.2	AVANTAGES ET INCONVENIENTS .....	22
<b>7.5</b>	<b>L'ALGORITHME DES K-PLUS PROCHES VOISINS .....</b>	<b>23</b>
7.5.1	ALGORITHME DE CLASSIFICATION PAR K-PPV.....	23
7.5.2	COMMENT CELA MARCHE-T-IL ? .....	23
7.5.3	AVANTAGES ET INCONVENIENTS .....	24
<b>7.6</b>	<b>L'ALGORITHME DES K-MOYENNES (K-MEANS).....</b>	<b>25</b>
7.6.1	PRINCIPE DE FONCTIONNEMENT .....	25
<b>7.7</b>	<b>ALGORITHME DE CLUSTERING PAR K-MEANS .....</b>	<b>25</b>
7.7.1	EVALUATION .....	25
7.7.2	AVANTAGES ET INCONVENIENTS .....	26

## **8 CATEGORISATION DES SYSTEMES DU DATA MINING.....26**

## **9 DOMAINES D'APPLICATION DU DATA MINING .....27**

9.1	LE DATA MINING DANS LE SECTEUR BANCAIRE.....	27
9.2	LE DATA MINING DANS LA BIO-INFORMATIQUE ET LA BIOTECHNOLOGIE .....	28
9.3	LE DATA MINING DANS LE MARKETING DIRECT ET LE COLLECTE DE FONDS .....	28
9.4	LE DATA MINING DANS LA DETECTION DE FRAUDE.....	29
9.5	LE DATA MINING DANS LA GESTION DE DONNEES SCIENTIFIQUES.....	29
9.6	LE DATA MINING DANS LE SECTEUR DES ASSURANCES .....	30
9.7	LE DATA MINING DANS LA TELECOMMUNICATION .....	30
9.8	LE DATA MINING DANS LA MEDECINE ET LA PHARMACIE .....	30
9.9	LE DATA MINING DANS LE COMMERCE AU DETAIL .....	31
9.10	LE DATA MINING DANS LE E-COMMERCE ET LE WORLD WIDE WEB .....	31
9.11	LE DATA MINING DANS LE MARCHE BOURSIER ET L'INVESTISSEMENT.....	31
9.12	LE DATA MINING DANS L'ANALYSE DE CHAINE D'APPROVISIONNEMENT .....	32

## **10 CONCLUSION .....32**

### **CHAPITRE II: ÉTUDE DES TRAVAUX EXPLOITANT LES SMAS POUR LE DATA MINING**

## **1 INTRODUCTION .....34**

## **2 L'INTELLIGENCE ARTIFICIELLE DISTRIBUEE (IAD).....34**

## **3 CONCEPT D'AGENT .....35**

3.1	DEFINITIONS .....	35
3.2	DIFFERENCE ENTRE OBJET ET AGENT .....	36
3.3	TYPES D'AGENTS.....	36
3.3.1	LES AGENTS COGNITIFS.....	36
3.3.2	LES AGENTS REACTIFS .....	37

## **4 LES SYSTEMES MULTI-AGENTS.....37**

<b>4.1</b>	<b>DEFINITIONS .....</b>	<b>37</b>
<b>4.2</b>	<b>QUAND UTILISER UN SMA? .....</b>	<b>38</b>
<b>5</b>	<b><u>QUELQUES TRAVAUX EXPLOITANT LES SMAS POUR LE DM .....</u></b>	<b><u>38</u></b>
<b>5.1</b>	<b>APPROCHE SMA POUR LA SEGMENTATION MARKOVIENNE DES TISSUS ET STRUCTURES PRESENTS DANS LES IRM CEREBRALES .....</b>	<b>38</b>
5.1.1	IMPLEMENTATION SMA .....	39
5.1.2	EVALUATION SUR IMAGES REELLES ACQUISES A 3T.....	43
<b>5.2</b>	<b>UNE APPROCHE SMA DE L'AGREGATION ET DE LA COOPERATION DES CLASSIFIEURS.....</b>	<b>43</b>
5.2.1	UNE APPROCHE FONDEE SUR LES SMA .....	44
5.2.2	FONCTIONNEMENT.....	44
5.2.3	SYNTHESE.....	46
<b>5.3</b>	<b>UNE APPROCHE POUR L'EXTRACTION DES REGLES D'ASSOCIATION SPATIALES BASEE MULTI-AGENT : RASMA</b>	<b>47</b>
5.3.1	ARCHITECTURE DE RASMA .....	50
5.3.2	PRESENTATION DES AGENTS.....	51
5.3.3	EXPERIMENTATION, RESULTATS ET PERFORMANCES .....	55
5.3.4	EVALUATION .....	56
<b>6</b>	<b><u>CONCLUSION .....</u></b>	<b><u>57</u></b>

### CHAPITRE III: MODÉLISATION

<b>1</b>	<b><u>INTRODUCTION .....</u></b>	<b><u>59</u></b>
<b>2</b>	<b><u>PRESENTATION GENERALE DU MODELE .....</u></b>	<b><u>59</u></b>
<b>3</b>	<b><u>DESCRIPTION DES COMPOSANTS DE L'ARCHITECTURE.....</u></b>	<b><u>61</u></b>
<b>3.1</b>	<b>AGENT INTERFACE .....</b>	<b>61</b>
3.1.1	ARCHITECTURE DE L'AGENT INTERFACE .....	61
3.1.2	FONCTIONNEMENT DE L'AGENT INTERFACE .....	62
3.1.3	LE SAVOIR DE L'AGENT INTERFACE.....	62
<b>3.2</b>	<b>AGENT INIT-CLUSTER.....</b>	<b>63</b>
3.2.1	ARCHITECTURE DE L'AGENT INIT-CLUSTER.....	63
3.2.2	FONCTIONNEMENT DE L'AGENT INIT-CLUSTER .....	64
3.2.3	LE SAVOIR DE L'AGENT INIT-CLUSTER.....	64
<b>3.3</b>	<b>AGENT AFFECT-CLUSTER .....</b>	<b>64</b>
3.3.1	ARCHITECTURE DE L'AGENT AFFECT-CLUSTER .....	65
3.3.2	FONCTIONNEMENT DE L'AGENT AFFECT-CLUSTER.....	66
3.3.3	LE SAVOIR DE L'AGENT AFFECT-CLUSTER.....	67
<b>3.4</b>	<b>AGENT CALC-CENTROÏDE.....</b>	<b>68</b>
3.4.1	ARCHITECTURE DE L'AGENT CALC-CENTROÏDE.....	68
3.4.2	FONCTIONNEMENT DE L'AGENT CALC-CENTROÏDE .....	69
3.4.3	LE SAVOIR DE L'AGENT CALC-CENTROÏDE .....	69
<b>3.5</b>	<b>AGENT CLAC-DISTANCE.....</b>	<b>70</b>
3.5.1	ARCHITECTURE DE L'AGENT CALC-DISTANCE.....	70
3.5.2	FONCTIONNEMENT DE L'AGENT CALC-DISTANCE .....	71
3.5.3	LE SAVOIR DE L'AGENT CALC-DISTANCE .....	71

<b>4</b>	<b><u>LA COMMUNICATION INTER-AGENTS</u></b> .....	<b>72</b>
<b>5</b>	<b><u>FONCTIONNEMENT DU SYSTEME</u></b> .....	<b>73</b>
5.1	DIAGRAMME DE CLASSES .....	73
5.2	DIAGRAMME DE SEQUENCE .....	73
<b>6</b>	<b><u>CONCLUSION</u></b> .....	<b>75</b>

#### **CHAPITRE IV: ÉTUDE DE CAS**

<b>1</b>	<b><u>INTRODUCTION</u></b> .....	<b>77</b>
<b>2</b>	<b><u>LES ETAPES DU PROCESSUS ADOPTE</u></b> .....	<b>77</b>
2.1	PREPARATION DES DONNEES .....	77
2.2	NETTOYAGE ET TRANSFORMATION DES DONNEES.....	79
2.3	DATA MINING .....	84
<b>3</b>	<b><u>OUTILS DE PROGRAMMATION</u></b> .....	<b>90</b>
3.1	POURQUOI JAVA? .....	90
3.2	LA PLATEFORME JADE .....	90
3.2.1	L'ENVIRONNEMENT D'EXECUTION JADE .....	92
3.2.2	LA COMMUNICATION ENTRE LES AGENTS JADE .....	92
<b>4</b>	<b><u>CONCLUSION</u></b> .....	<b>93</b>
	<b><u>CONCLUSION GENERALE</u></b> .....	<b>95</b>
	<b><u>REFERENCES</u></b> .....	<b>97</b>

# **Introduction générale**

# *Introduction générale*

Toutes les entreprises de nos jours collectent et stockent des quantités de données très grandes et qui ne cessent d'augmenter jour après jour. Ces mégabases de données sont des mines d'informations, elles cachent des connaissances décisives face au marché et à la concurrence, mais elles restent peu exploitées. Pour combler ce besoin une nouvelle industrie est en train de naître : le Data Mining (en français fouille de données) qui propose d'utiliser un ensemble d'algorithmes issus de différentes disciplines scientifiques tel que les statistiques, l'intelligence artificielle et les bases de données afin de construire des modèles à partir des données, autrement dit trouver des schémas intéressants selon des critères fixés au départ et d'extraire un maximum de connaissances utiles à l'entreprise.

Fouiller de très grandes bases de données de l'ordre de petabytes et en extraire des informations utiles ainsi que la distribution des données dans certain cas de bases de données distribuées représentent des difficultés dans la mise en œuvre d'un système de data mining qui nécessite un énorme volume de travail et un temps d'exécution très grand.

Une branche de l'Intelligence Artificielle Distribuée consiste à ce que des entités possédants une certaine autonomie, doivent être dotés de capacités de perception et d'action sur leur environnement, nous parlons ici d'agents et par conséquent de systèmes multi agents. Ces systèmes deviennent indispensables dans plusieurs domaines d'applications dans la mesure où ils résolvent les problèmes de complexité et de distribution surtout quand il s'agit de grand système tel que ceux d'extraction de connaissances.

L'objectif de ce travail est d'intégrer le paradigme multi agents dans les systèmes de data mining pour manier à ses problèmes. Pour ce fait notre thèse commence par un chapitre qui introduit la notion de data mining par donner quelques définitions de ce terme ainsi que les motivations qui ont mené à l'apparition de ce genre de systèmes, ensuite nous avons illustré les différents types de données qu'un système de data mining peut s'appliquer sur, après ça nous avons expliqué expliquer toutes les taches du data mining et que chacune d'elles représente une façon d'extraction de connaissances. Les étapes du processus de data mining ont été aussi expliquées tout comme les techniques et les algorithmes utilisées dans le domaine de data mining et la catégorisation des systèmes de data mining. Pour conclure ce chapitre nous avons abordées avec un peu de détailles les différents domaines d'application de data mining.

## **Introduction Générale**

---

Le deuxième chapitre est destiné aux systèmes multi agent, il est composé de deux parties : la première constitue une vue générale sur les systèmes multi agent et le concept d'agent en commençant par une définition de l'intelligence artificielle. Dans la deuxième partie nous avons exposé quelques travaux exploitant les systèmes multi agents pour le data mining.

Le troisième chapitre contient les détails du modèle que nous avons proposée pour une approche d'intégration d'agent dans les systèmes de data mining, il contient aussi l'architecture que nous avons adoptée pour chaque composant de notre système.

Le quatrième chapitre contient la validation du modèle proposé, et la présentation des différents outils utilisés dans ce but.

Nous finissons notre mémoire par une conclusion générale qui contient les perspectives possibles qui constituent une suite de recherche pour notre projet.

## **Chapitre I**

---

### ***Généralités Sur Le Data Mining.***

### 1 Introduction

Les données brutes, malgré leur quantité qui augmente d'une façon exponentielle, n'ont presque aucune valeur, ce qui est le plus important en fait c'est les connaissances pour lesquelles nous sommes tous assoiffés et qui sont obtenus par la compréhension de ces données, mais plus on a de données plus ce processus devient difficile.

De nos jours, les changements de notre environnement sont dénotés par des capteurs qui sont devenus de plus en plus nombreux. Par conséquent, la compréhension de ces données est très importante. Et comme il est dit par Piatestky-Shapiro, « [...] *as long as the world keeps producing data of all kinds [...] at an ever increasing rate, the demand for data mining will continue to grow* » [1]. D'où la fouille de données devient une nécessité.

### 2 Définition du data mining

Selon le Groupe Gartner, le Data Mining appelé aussi fouille de données est le processus de découverte de nouvelles corrélations, modèles et tendances en analysant une grande quantité de données, en utilisant les technologies de reconnaissance des formes ainsi que d'autres techniques statistiques et mathématiques[2].

Ils existent d'autres définitions :

- Le Data Mining est l'analyse de grandes ensembles de données observationnelles pour découvrir des nouvelles relations entre elles et de les reformuler afin de les rendre plus utilisables de la part de ses propriétaires [3].
- Le Data Mining est un domaine interdisciplinaire utilisant dans le même temps des techniques d'apprentissage automatiques, de reconnaissance des formes, des statistiques, des bases de données et de visualisation pour déterminer les manières d'extraction des informations de très grandes bases de données [4].
- Le Data Mining est un processus inductif, itératif et interactif dont l'objectif est la découverte de modèles de données valides, nouveaux, utiles et compréhensibles dans de larges Bases de Données [5].

### 3 Motivations

La nécessité est la mère de l'invention. — Platon

Le Data Mining a eu une grande attention ces derniers temps dans l'industrie de l'information, ça est due principalement à la disponibilité de grandes quantités de données et au besoin de les transformés en connaissances utiles surtout dans les applications concernant l'analyse du marché, la détection de fraudes, la conservation de client, le contrôle de productions et l'exploration scientifique.

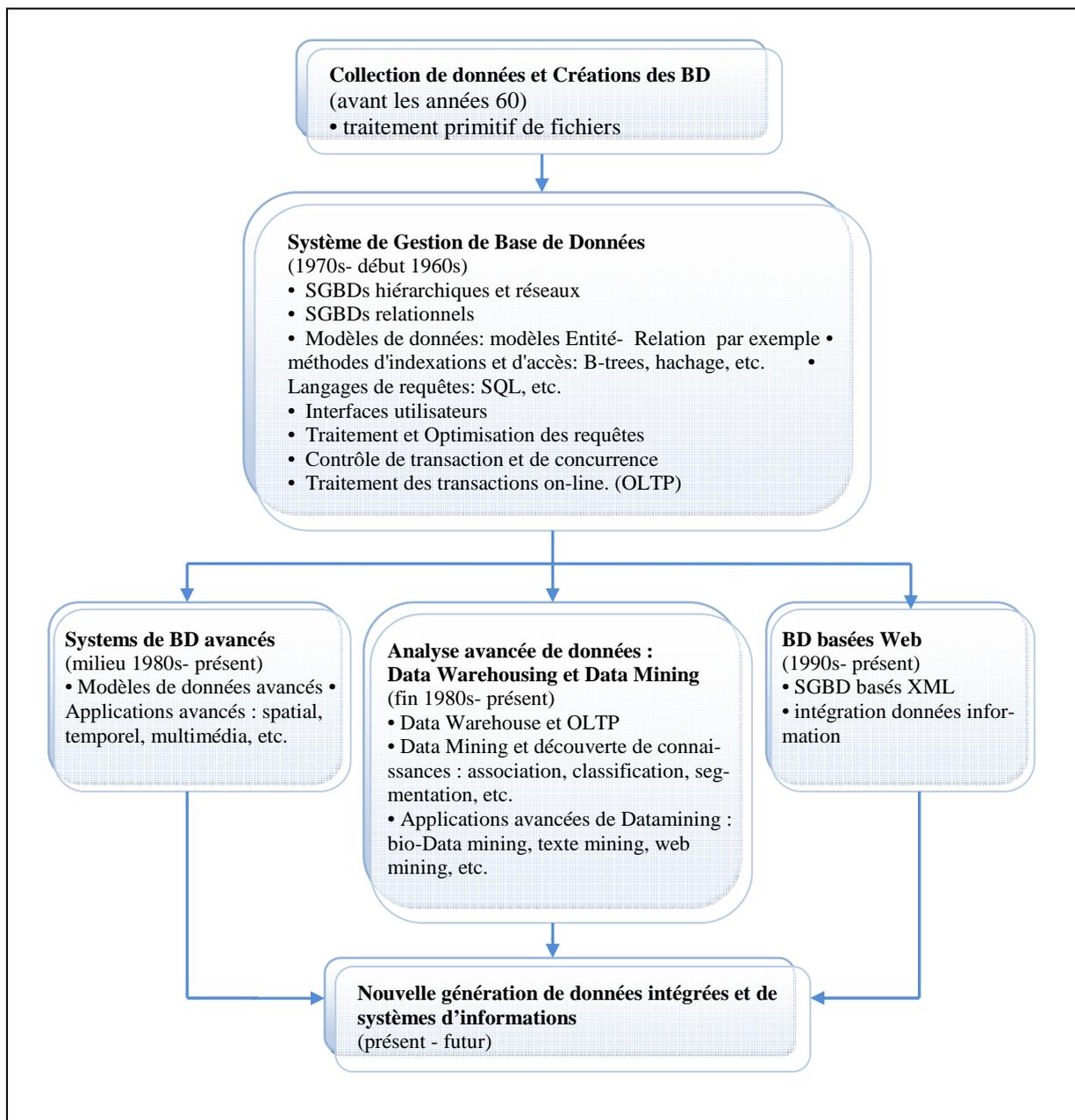


Figure 1: L'évolution de la technologie système de BD.

## Chapitre I : Généralités Sur Le Data Mining

---

Le Data Mining peut être vue comme l'évolution naturelle de la technologie d'information (Figure 1). Depuis les années 60s, la technologie des informations et des Bases de Données a systématiquement évoluée des systèmes de gestion de fichiers aux systèmes de Bases de Données plus puissant et plus sophistiqués. La recherche et le développement dans le domaine de systèmes de Bases de Données depuis les années 70s a menée au développement des systèmes de Bases de Données relationnels, des outils de modélisation des données et des techniques d'indexation et d'organisation des données.

Depuis le milieu des années 80s, la technologie des Bases de Données a été caractérisée par l'adoption de la technologie relationnelle et la croissance des recherches et des activités de développement d'un nouveau et puissant système de Bases de Données. Ceux-ci favorisent des modèles avancés de données comme le modèle relationnel étendu, orienté-objet, relationnel-objet et des systèmes de Bases de Données déductifs. Des issues de distribution, de diversification et de partage de données ont été étudiées intensivement. Des systèmes de BD hétérogènes et des systèmes d'informations basés sur l'internet comme le World Wide Web (WWW) ont aussi émergés à jouer un rôle vital dans l'industrie de l'information.

La progression de la technologie du matériel informatique pendant les trois décennies passées nous a amené à la puissance accrue des ordinateurs, des équipements de collection de données et du stockage du media. Cette technologie a donnée un grand coup de puce à l'industrie de bases de données et de l'information et a permet la gestion des transactions, la recherche d'informations et l'analyse de données de très grandes Bases de Données.

Les données peuvent maintenant être stockées dans des différentes Bases de Données. L'une de Bases de Données la plus récente est le Data Warehouse, qui regroupe une multitude de sources de données hétérogènes, organisées sous un unie schéma et un seul emplacement pour faciliter la gestion. La technologie du Data Warehousing incluse l'integration de données et le OLAP (On-Line Analytical Processing) qui est un ensemble de techniques d'analyse avec des fonctionnalités tels que la consolidation et l'agrégation.

L'abondance de données couplées avec le besoin de bons outils d'analyse de données ont été décrits par "*Riche en données mais Pauvre en informations*". La croissance rapide et l'énorme quantité de données collectées et stockées dans de grandes et nombreuse Bases de Données ont excédé notre habilité humaine de vouloir comprendre sans utilisé aucun outils (Figure 2). Et comme résultat à tout ça, les grandes Bases de Données collectant de grandes quantités de données sont devenues des *tombes de données* (les données archivées sont rarement revisitées). Par conséquence, des décisions importantes sont souvent prises par

l'intuition des décideurs et non pas par les données riches en informations stockées dans les Bases de Données, simplement parce que les décideurs n'ont pas les outils pour extraire les connaissances précieuses embarquées dans les grandes quantités de données.



**Figure 2: Riche en données mais Pauvre en informations.**

Les technologies des systèmes experts actuels s'appuient typiquement sur les utilisateurs ou les experts d'un domaine pour faire entrer des connaissances dans des bases de connaissances manuellement. Malheureusement, cette procédure est coûteuse, elle peut causer des erreurs et elle prend du temps. Les outils du Data Mining qui exécutent l'analyse des données peuvent découvrir d'important modèles de données contribuent énormément aux stratégies du business, aux bases de connaissances et aux recherches scientifiques et médicales. L'élargissement de l'écart entre données et informations fait l'appel au développement systématique des outils du Data Mining qui vont transformer les tombes de données en pépites d'or de connaissances [6] [7].

### 4 Data mining sur quels types de données ?

Le Data Mining n'est pas spécifique à un type de médias ou de données. Il est applicable à n'importe quel type d'information. Le Data Mining est utilisé et étudié pour les Bases de Données incluant les Bases de Données relationnelles et les Bases de Données Orientées-Objets, les data warehouses, les Bases de Données transactionnelles, les supports de données non structurés et semi-structurés comme le World Wide Web, les Bases de Données avancés comme les Bases de Données spatiales, les Bases de Données multimédia, les Bases de données de séries temporelles et les Bases de Données textuelles et même fichiers plats.

#### 4.1 Les fichiers plats

Les fichiers plats sont actuellement la source de données la plus commune pour les algorithmes du Data Mining et particulièrement dans le niveau de recherches. Les fichiers plats sont des fichiers de données simples dans le format texte ou binaire avec une structure connue par l'algorithme du Data Mining qui va être y appliqué.

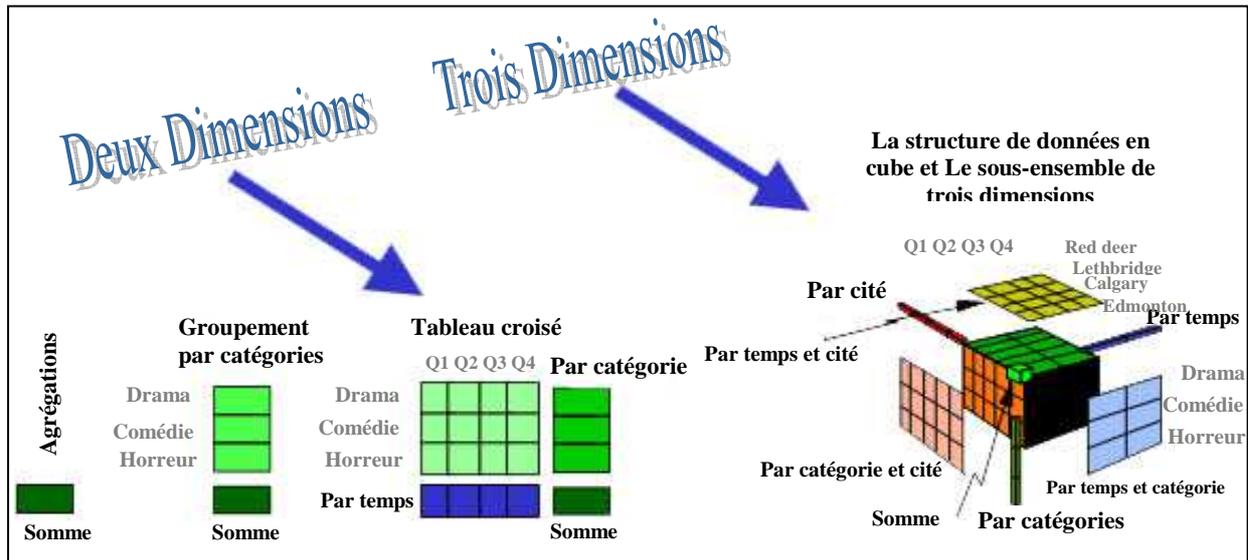
#### 4.2 Les bases de données relationnelles

Les algorithmes du Data Mining appliqués sur des Bases de Données relationnelles sont plus polyvalents que les algorithmes spécifiquement faits pour les fichiers plats puisqu'ils peuvent profiter de la structure inhérente aux bases de données relationnelles. Le Data Mining peut profiter du SQL pour la sélection, la transformation et la consolidation, il passe au-delà de ce que le SQL pourrait fournir, comme la prévision, la comparaison, la détection des déviations, etc. [13].

#### 4.3 Les data warehouses

Un Data Warehouse est un support de données rassemblées de multiples sources de données (souvent hétérogènes) et est destinée à être utilisé dans l'ensemble sous le même schéma unifié. Supposons une entreprise OurVideoStore qui a des contrats d'exclusivité dans toute l'Amérique du Nord. La plupart des magasins de vidéos appartenant à l'entreprise OurVideoStore peuvent avoir des bases de données et des structures différentes. Si le directeur de l'entreprise veut accéder aux données de tous les magasins pour prendre des décisions stratégiques, il serait plus approprié si toutes les données étaient stockées dans un seul emplacement avec une structure homogène qui permet l'analyse interactive des données. Autrement dit, les données de différents magasins peuvent être chargées, nettoyées, transformées et intégrées ensemble. Pour faciliter la prise de décisions et les vues

multidimensionnelles, les Data Warehouses sont souvent modélées par une structure de données multidimensionnelle. La figure 3 montre un exemple d'un sous-ensemble de trois dimensions d'une structure de données en cube utilisée pour le Data Warehouse de OurVideoStore.



**Figure 3: La structure de données en cube utilisée généralement pour les données du Data Warehouses.**

La figure montre les résumés des locations groupées par les catégories de film, et une table croisée des résumés de locations groupées par les catégories de film et du temps (en trimestres). Le cube de données donne les résumés des locations selon des trois dimensions : catégorie, temps et ville. Un cube contient des cellules qui stockent les valeurs de quelques ensembles de mesures (dans notre cas les comptes de location) et des cellules spéciales contenant les sommes le long de dimensions. Chaque dimension du cube de données contient pour un attribut une hiérarchie de valeurs [13] [14].

Les cubes de données, grâce à leur structure, des résumés de données pré-calculés qu'ils contiennent et des valeurs hiérarchiques d'attributs de leurs dimensions, conviennent bien aux requêtes interactives rapides et l'analyse de données à différents niveaux conceptuels connus par **OLAP (On-Line Analytical Processing)**. Les opérations de L'OLAP permettent la navigation des données aux différents niveaux d'abstraction, tel que drill-down qui consiste à donner un niveau de détails sur les données et roll-up qui est le contraire de drill-down qui consiste à faire de l'agrégation (ou résumé) des données. La figure 4 montre les opérations de drill-down (sur la dimension temps) et roll-up (sur la dimension emplacement) [13][15].

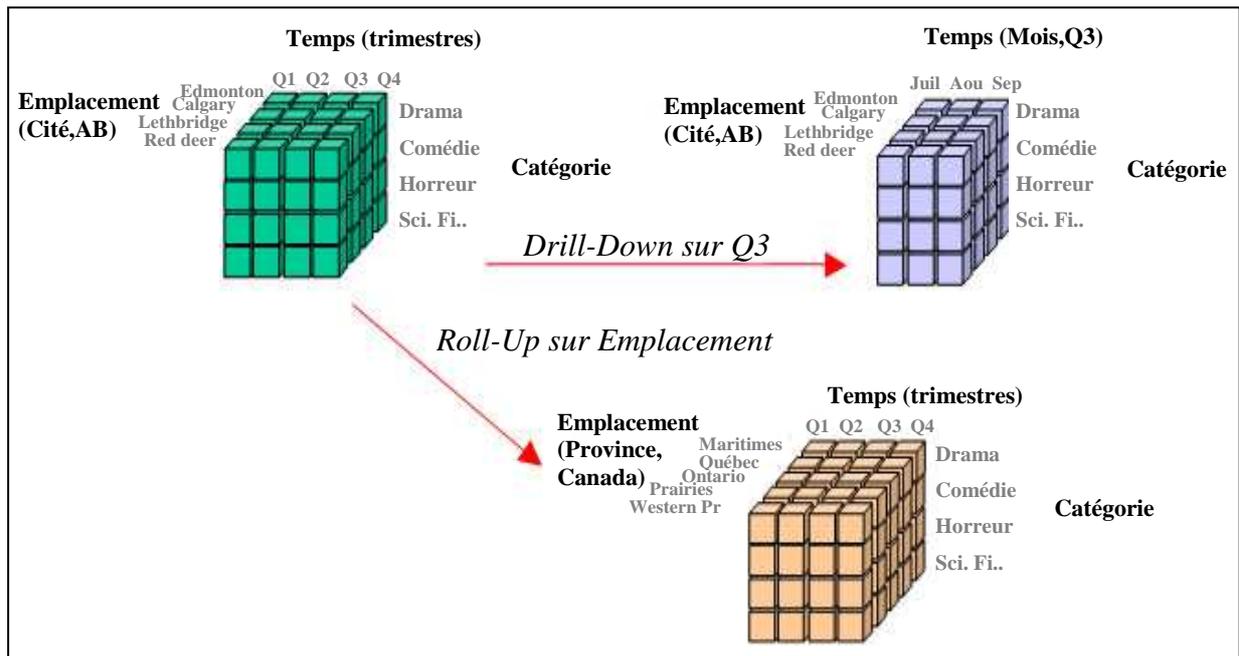


Figure 4: Les données résumées de *OurVideoStore* avant et après les opérations *drill-down* et *roll-up* [13].

#### 4.4 Les bases de données transactionnelles

En général, une Base de Données transactionnelle est un fichier où chaque enregistrement représente une transaction. Une transaction contient un identifiant unique de transaction (*transactionID*) et une liste d'items composant la transaction (les achats d'un client lors d'une visite). Les bases de données transactionnelles peuvent contenir d'autres informations tels que la date de la transaction, l'identifiant du consommateur, l'identifiant de la personne qui a vendu, et ainsi de suite. [6] Par exemple, dans le cas du magasin de vidéos, la table de locations qui est illustrée par la figure 5 représente la base de données transactionnelle où chaque enregistrement est un contrat de location contenant un identifiant du consommateur, une date et une liste des items loués (cassettes vidéo, jeux, VCR, etc.). Pour le Data Mining sur ce type de données nous utilisons souvent les règles d'association (appelées aussi Analyse du panier de la ménagère) dans lesquelles les associations des items arrivant ensemble ou séquentiellement soient étudiées.

Locations					
<i>transactionID</i>	<i>Date</i>	<i>temps</i>	<i>consommateurID</i>	<i>itemList</i>	
T12345	99/09/06	19:38	C1234	{ I2, I6, I10, I45 ... }	
...					

**Figure 5: Fragment de la base de données transactionnelle des locations à OurVideoStore [13].**

### 4.5 Les bases de données multimédia

Les bases de données multimédia comportent des documents sonores, des vidéos, des images et des médias en textes et audio. Elles peuvent être stockées sur des bases de données orientées objets ou objets relationnelles ou simplement sur un fichier système. Le multimédia est caractérisé par sa haute dimension ce qui rend le datamining sur ce type de données très difficile. Le data mining sur les supports des multimédias requiert exige la vision par ordinateur, l'infographie, l'interprétation des images et les méthodologies de traitement de langages naturels [13] [16].

### 4.6 Les bases de données spatiales

Ce sont des bases de données, qu'en plus de leurs données usuelles, elles contiennent des informations géographiques comme les cartes et les positionnements mondiaux ou régionaux. De telles bases de données présentent de nouveaux défis aux algorithmes de data mining [13].

### 4.7 Les bases de données de séries temporelles

Les bases de données de séries temporelles contiennent des données relatives au temps, comme les données du marché boursier ou les activités enregistrées. Ces bases de données ont couramment un flux continu de nouvelles données entrantes, qui parfois rend l'analyse en temps réel un besoin exigeant. Le data mining pour ce genre de bases de données est généralement l'étude des tendances et des corrélations entre les évolutions des différentes variables, aussi bien que la prédiction des tendances et des mouvements des variables par rapport au temps. Par exemple, une base de données du trafic automobile qui stocke une description symbolique de séries temporelles de ce dernier, il sera possible de répondre à la requête : « définir les grand axes où le commerce est fluide le week-end ». La figure 5 montre quelques exemples de données en séries temporelles [13] [17].



Figure 6: Quelques exemples de données en séries temporelles [12].

### 4.8 Le World Wide Web

Le World Wide Web est le support de données le plus hétérogène et le plus dynamique disponible. Un grand nombre d'auteurs et d'éditeurs contribuent sans arrêt à son accroissement et évolution, et chaque jour un énorme nombre d'utilisateurs accède à ses ressources. Les données dans le World Wide Web sont organisées dans des documents interconnectés. Ces documents peuvent être des textes, audio, vidéos, données brutes et même des applications. Conceptuellement, le World Wide Web est composé de trois grands composants : le contenu du Web, qui englobe les documents disponibles ; la structure du Web, qui garantit les hyperliens et les relations entre documents ; et l'usage du Web, en décrivant quand et comment les ressources seront accédées. Une quatrième dimension peut être ajoutée concernant la nature dynamique ou l'évolution des documents. Le data mining pour le World Wide Web, ou le web mining, essaie d'aborder toutes ces questions et il est souvent divisé en contenu Web mining, la structure Web mining et l'usage Web mining [13] [18].

### 5 Les tâches du data mining

Beaucoup de problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en termes des six tâches suivantes :

- ❖ La classification
- ❖ L'estimation
- ❖ La prédiction
- ❖ Le groupement par similitude
- ❖ L'analyse des clusters
- ❖ La description [8].

Les trois premières tâches sont des exemples du Data Mining supervisé dont le but est d'utiliser les données disponibles pour créer un modèle décrivant une variable particulière prise comme but en termes de ces données. Le groupement par similitude et l'analyse des clusters sont des tâches non-supervisées où le but est d'établir un certain rapport entre toutes les variables [9].

La description appartient à ces deux catégories de tâche, elle est vue comme une tâche supervisée et non-supervisée en même temps [8].

#### 5.1 La classification

La classification est la tâche la plus commune du Data Mining et qui semble être une obligation humaine. Afin de comprendre notre vie quotidienne, nous sommes constamment classifiés, catégorisés et évalués [8].

La classification consiste à étudier les caractéristiques d'un nouvel objet pour lui attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant. L'objectif est de créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées [9]

Quelques exemples de l'utilisation des tâches de classification dans les domaines de recherche et commerce sont les suivants :

- Déterminer si l'utilisation d'une carte de crédit est frauduleuse.
- Diagnostiquant si une certaine maladie est présente [10].
- Déterminer quels numéros de téléphone correspondent aux fax [8].

- Déterminer quelles lignes téléphoniques sont utilisées pour l'accès à Internet [9].

### 5.2 L'estimation

L'estimation est similaire à la classification à part que la variable de sortie est numérique plutôt que catégorique. En fonction des autres champs de l'enregistrement l'estimation consiste à compléter une valeur manquante dans un champ particulier. Par exemple on cherche à estimer La lecture de tension systolique d'un patient dans un hôpital, en se basant sur l'âge du patient, son genre, son indice de masse corporelle et le niveau de sodium dans son sang. La relation entre la tension systolique et les autres données vont fournir un modèle d'estimation. Et par la suite nous pouvons appliquer ce modèle dans d'autres cas [8] [10].

Quelques exemples de l'utilisation des tâches d'estimation dans les domaines de recherche et commerce sont les suivants :

- Estimer le nombre d'enfants dans une famille [8].
- Estimant le montant d'argent qu'une famille de quatre membres choisis aléatoirement dépensera pour la rentrée scolaire [10].
- Estimer la valeur d'une pièce immobilière [9].

Souvent, la classification et l'estimation sont utilisés ensemble, comme quand le Data Mining est utilisée pour prévoir qui va probablement répondre à une offre de transfert d'équilibre(de solde) de carte de crédit et aussi évaluer la taille de l'équilibre(du solde) à être transféré.

### 5.3 La prédiction

La prédiction est la même que la classification et l'estimation, à part que dans la prédiction les enregistrements sont classés suivant des critères (ou des valeurs) prédites (estimées). La principale raison qui différencie la prédiction de la classification et l'estimation est que dans la création du modèle prédictif on prend en charge la relation temporelle entre les variables d'entrée et les variables de sortie [8].

Quelques exemples de l'utilisation des tâches de prédiction dans les domaines de recherche et commerce sont les suivants :

- Prévoir le prix des actions dans les trois prochains mois [10].
- Prévoir le champion de la coupe du monde en football en se basant sur la comparaison des statistiques des équipes.

- Prévoir quels clients va déménager dans les 6 mois qui suivent [8].

### 5.4 Le groupement par similitude

Le groupement par similitude consiste à déterminer quels attributs “vont ensemble”. La tâche la plus répandue dans le monde du business, où elle est appelée l’analyse d’affinité ou l’analyse du panier du marché, est l’association des recherches pour mesurer la relation entre deux et plusieurs attributs. Les règles d’associations sont de la forme “Si antécédent, alors conséquent”.

Quelques exemples de l’utilisation des tâches du groupement par similitude dans les domaines de recherche et commerce sont les suivants :

- Trouver dans un supermarché quels produits sont achetés ensemble et quels sont ceux qui ne s’achètent jamais ensemble.
- Déterminer la proportion des cas dans lesquels un nouveau médicament peut générer des effets dangereux [10].

### 5.5 L’analyse des clusters

Le clustering (ou la segmentation) est le regroupement d’enregistrements ou des observations en classes d’objets similaires; un cluster est une collection d’enregistrements similaires l’un à l’autre, et différents à ceux existants sur les autres clusters. La différence entre le clustering et la classification est que dans le clustering il n’y a pas de variables sortantes. La tâche de clustering ne classifie pas, n’estime pas, ne prévoit pas la valeur d’une variable sortantes. Au lieu de cela, les algorithmes de clustering visent à segmenter la totalité de données en des sous groupes relativement homogènes. Ils maximisent l’homogénéité à l’intérieur de chaque groupe et la minimisent entre ces derniers [10].

Les algorithmes du clustering peuvent être appliqués dans des différents domaines, tel que :

- Découvrir des groupes de clients ayants des comportements semblables.
- Classification des plantes et des animaux étant donné leurs caractéristiques.
- Segmentation les observations des épicentres pour identifier les zones dangereuses [21].

### 5.6 La description

Parfois le but du Data Mining est simplement de décrire se qui se passe sur une Base de Données compliquée en expliquant les relations existantes dans les données pour en

premier lieu comprendre le mieux possible les individus, les produit et les processus présents sue cette base. Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Dans la société Américaine nous pouvons prendre comme exemple comment une simple description, «les femmes supportent le parti Démocrate plus que les hommes», peut provoquer beaucoup d'intérêt et promouvoir les études de la part des journalistes, sociologues, économistes et les spécialistes en politiques [8] [11].

## 6 Les étapes du processus de data mining

- 1) **Collecte des données** : la combinaison de plusieurs sources de données, souvent hétérogènes, dans une base de données [13] [19].
- 2) **Nettoyage des données** : la normalisation des données : l'élimination du bruit (les attributs ayant des valeurs invalides et les attributs sans valeurs) [13] [19].
- 3) **Sélection des données** : Sélectionner de la base de données les attributs utiles pour une tâche particulière du data mining [20].
- 4) **Transformation des données** : le processus de transformation des structures des attributs pour être adéquates à la procédure d'extraction des informations [21].
- 5) **Extraction des informations (Data mining)**: l'application de quelques algorithmes du Data Mining sur les données produites par l'étape précédente (*Knowledge Discovery in Databases*, ou KDD) [19] [20].
- 6) **Visualisation des données** : l'utilisation des techniques de visualisation (histogramme, camembert, arbre, visualisation 3D) pour exploration interactive de données (la découverte des modèles de données) [21] [19].
- 7) **Evaluation des modèles** : l'identification des modèles strictement intéressants en se basant sur des mesures données [13].

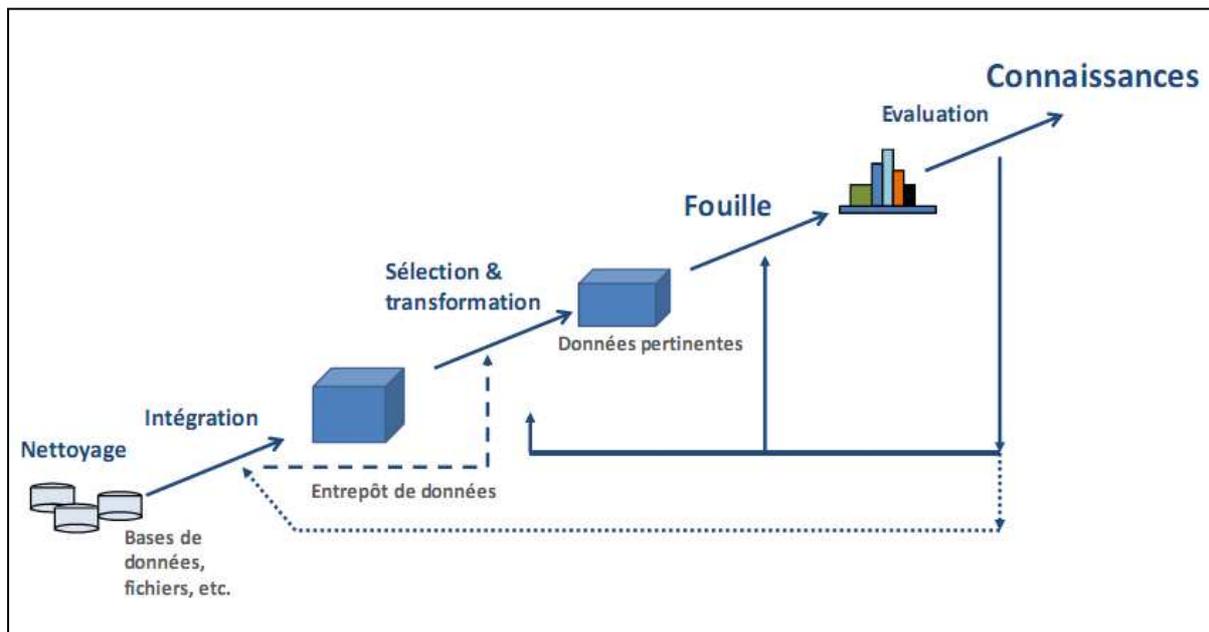


Figure 7: Le processus de data mining [24].

## 7 Techniques du data mining

Pour effectuer les tâches du Data Mining il existe plusieurs techniques issues de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) afin de faire apparaître des corrélations cachées dans des gisements de données pour construire des modèles à partir de ces données. Dans ce chapitre, nous présentons les techniques du data mining les plus connues.

### 7.1 Les réseaux de neurones

Un réseau de neurones est un modèle de calcul dont le fonctionnement vise à simuler le fonctionnement des neurones biologiques, il est constitué d'un grand nombre d'unités (neurones) ayant chacune une petite mémoire locale et interconnectées par des canaux de communication qui transportent des données numériques. Ces unités peuvent uniquement agir sur leurs données locales et sur les entrées qu'elles reçoivent par leurs connections. Les réseaux de neurones sont capables de prédire de nouvelles observations (sur des variables spécifiques) à partir d'autres observations (soit les mêmes ou d'autres variables) après avoir exécuté un processus d'apprentissage sur des données existantes.

La phase d'apprentissage d'un réseau de neurones est un processus itératif permettant de régler les poids du réseau pour optimiser la prédiction des échantillons de données sur

lesquelles l'apprentissage été fait. Après la phase d'apprentissage le réseau de neurones devient capable de généraliser [21].

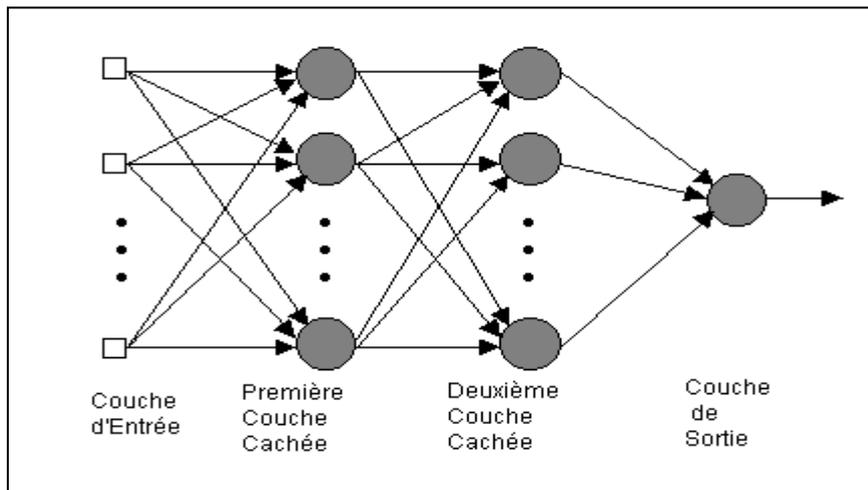


Figure 8 : Réseau de neurones.

### 7.1.1 Avantages et inconvénients

- **Avantages**

Les réseaux de neurones sont théoriquement capables d'approximer n'importe quelle fonction continue et ainsi le chercheur n'a pas besoin d'avoir aucunes hypothèses du modèle sous-jacent [21].

- **Inconvénients**

Généralement les réseaux de neurones ne sont pas souvent utilisées dans les tâches du data mining parce qu'ils produisent des modèles souvent incompréhensibles et demande un longtemps d'apprentissage [25].

### 7.2 Les arbres de décision

Les arbres de décisions sont des outils d'aide à la décision qui permettent selon des variables discriminantes de répartir une population d'individus en groupes homogènes en fonction d'un objectif connu. Les arbres de décision sont des outils puissants et populaires pour la classification et la prédiction. Un arbre de décision permet à partir des données connues sur le problème de donner des prédictions par réduction, niveau par niveau, du domaine des solutions.

Chaque nœud interne d'un arbre de décision permet de répartir les éléments à classifier de façon homogène entre ses différents fils en portant sur une variable discriminante de ces éléments. Les branches qui représentent les liaisons entre un nœud et ses fils sont les valeurs

discriminantes de la variable du nœud. Et en fin, les feuilles d'un arbre de décision représentent les résultats de la prédiction des données à classifier [25].

### 7.2.1 Les algorithmes d'induction des arbres de décision

<i>Nom de l'algorithme</i>	<i>Développeur</i>	<i>Année</i>
CHAID	Kass	1980
CART	Breiman, et al.	1984
ID3	Quinlan	1986
C4.5	Quinlan	1993
SLIQ	Agrawal, et al.	1996
SPRINT	Agrawal, et al.	1996

Tableau 1: Les algorithmes d'inductions des arbres de décision [21].

### 7.2.2 Avantages et inconvénients

- **Avantages**

- Les arbres de décision sont capables de produire des règles compréhensibles.
- Les arbres de décision effectuent la classification sans exiger beaucoup de calcul
- Les arbres de décision sont en mesure de manipuler à la fois les variables continues et catégorielles [21].

- **Inconvénients**

- Manque de performance dans le cas de plusieurs classes; les arbres deviennent très complexes et ne sont pas nécessairement optimaux.
- Demande beaucoup de temps de calcul lors de la construction (le choix du meilleur partitionnement) et l'élagage (la comparaison de sous-arbres).
- Moins bonnes performances concernant les prédictions portant sur des valeurs numériques [26].

### 7.3 Les algorithmes génétiques

Un algorithme génétique se constitue d'une catégorie de programmes dont le principe est la reproduction des mécanismes de la sélection naturelle pour résoudre un problème donné. L'optimisation des problèmes combinatoires et surtout les problèmes dits NP-complets (dont le temps de calcul croît de façon non polynomiale avec la complexité du problème) est l'objectif principal des algorithmes génétiques, ils sont particulièrement adaptés à ce type de

problèmes. Ces algorithmes constituent parfois une alternative intéressante aux réseaux de neurones mais sont le plus souvent complémentaires.

### 7.3.1 Principe de base des algorithmes génétiques

Le principe de fonctionnement d'un algorithme génétique est le suivant :

1. Codage du problème sous forme d'une chaîne binaire.
2. Génération aléatoire d'une population. Celle-ci contient un pool génétique qui représente un ensemble de solutions possibles.
3. Calcul d'une valeur d'adaptation pour chaque individu. Elle sera fonction directe de la proximité des différents individus avec l'objectif, on parle ici d'évaluation (fitness).
4. Sélection des individus doit se reproduire en fonction de leurs parts respectives dans l'adaptation globale.
5. Croisement des génomes des parents.
6. Sur la base de ce nouveau pool génétique, on repart à partir du point 3.

### 7.3.2 Codage d'un algorithme génétique

Avant de pouvoir utiliser un algorithme génétique pour résoudre un problème, il faut trouver un moyen pour encoder une solution potentielle à ce problème (les chromosomes), le codage consiste alors à choisir les structures de données qui coderont les gènes. Il existe différentes manières de le faire :

- **Codage binaire** : ce type de codage se base sur le principe de coder la solution selon une chaîne de bits (0 ou 1).
- **Codage à caractères multiples** : les chromosomes d'un algorithme génétique peuvent être codés d'une autre manière qui est le codage à l'aide de caractères. Souvent, ce type de codage est plus naturel que son précédent.
- **Codage sous forme d'arbre** : il utilise une structure arborescente avec une racine (parent) de laquelle peuvent être issus un ou plusieurs fils (descendants).

### 7.3.3 Avantages et inconvénients

- **Avantages**

- Ils utilisent l'évaluation de la fonction objective sans prendre en compte sa nature ce qui lui donne plus de souplesse et un large domaine d'application.
- Ils sont dotés de parallélisme car ils travaillent sur plusieurs points en même temps il s'agit des individus de la population.
- L'utilisation de règles de transition probabilistes de croisement et de mutation permet dans certains cas d'éviter des optimums locaux et d'aller vers un optimum global [27].

- **Inconvénients**

- Temps de calcul très élevé car ils nécessitent de nombreux calculs particulièrement au niveau de la fonction d'évaluation.
- Difficiles à mettre en œuvre à cause
  - i. des paramètres parfois difficiles à déterminer comme la taille de la population ou le taux de mutation. Ce qui implique la nécessité de plusieurs essais car le succès de l'évolution en dépend, ce qui limite encore l'efficacité de l'algorithme.
  - ii. du choix de la fonction d'évaluation qui est critique, elle doit prendre en compte les bons paramètres du problème. Elle doit donc être choisie avec soin.
- Il est impossible d'être sûr que la solution obtenue après un nombre fini de générations soit la meilleure, on peut seulement être sûr que l'on s'est approché de la solution optimale.
- Problème des optimums locaux : lorsqu'une population évolue, il se peut que certains individus deviennent majoritaires. À ce moment, il se peut que la population converge vers cet individu et s'écarte ainsi d'individus plus intéressants mais trop éloignés de l'individu vers lequel la population converge [18].

### 7.4 Les règles associatives

Les règles associatives sont des règles extraites d'une base de données transactionnelles et qui décrivent des associations entre certains éléments. Elles sont fréquemment utilisées dans le secteur de la distribution des produits où la principale application est *l'analyse du panier de la ménagère* (Market Basket Analysis) dont le principe est l'extraction d'associations entre produits sur les tickets de caisse. Le but de la méthode est l'étude de ce que les clients achètent pour obtenir des informations sur qui sont les clients et pourquoi ils font certains achats. La méthode recherche quels produits tendent à être achetés ensemble. La méthode peut être appliquée à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services : services bancaires, services de télécommunications, par exemple. Elle peut être également utilisée dans le secteur médical pour la recherche de complications dues à des associations de médicaments ou à la recherche de fraudes en recherchant des associations inhabituelles.

Une *règle d'association* est de la forme : Si **condition** alors **résultat**. Dans la pratique, nous nous limitons généralement à des règles où la condition se présente sous la forme d'une conjonction d'apparition d'articles et le résultat se constitue d'un seul article. Par exemple, une règle à trois articles sera de la forme : Si X et Y alors Z ; règle dont la sémantique peut être énoncée : Si les articles X et Y apparaissent simultanément dans un achat alors l'article Z apparaît [28] [25].

#### 7.4.1 Les algorithmes d'induction des règles associatives

<i>Nom de l'algorithme</i>	<i>Développeur</i>	<i>Année</i>
APRIORI	Agrawal, et al.	1993
FP-GROWTH	Han, et al.	2000
ECLAT	Zaki	2000
SSDM	Escovar, et al.	2005
KDCI	Orlando, et al.	2003

Tableau 2 : Les algorithmes d'inductions des règles associatives [25].

#### 7.4.2 Avantages et inconvénients

- **Avantages**
  - Résultats clairs : règles faciles à interpréter.
  - Simplicité de la méthode et des calculs (calculs élémentaires des fréquences d'apparition).

- Aucune hypothèse préalable (Apprentissage non supervisé).
- Méthode facile à adopter aux séries temporelles.
- **Inconvénients**
  - la méthode est coûteuse en temps de calcul.
  - Qualité des règles : production d'un nombre important de règles triviales (des règles évidentes qui, par conséquent, n'apportent pas d'information) ou inutiles (des règles difficiles à interpréter provenant de particularités propres à la liste des achats ayant servi à l'apprentissage).
  - Méthode non efficace pour les articles rares [28].

### 7.5 L'algorithme des *k-Plus proches voisins*

L'algorithme des *k* plus proches voisins (K-PPV, *k* nearest neighbor en anglais ou *k*NN) est un algorithme de raisonnement à partir de cas qui est dédié à la classification qui peut être étendu à des tâches d'estimation. Le but de cet algorithme est de prendre des décisions en se basant sur un ou plusieurs cas similaires déjà résolus en mémoire. Dans ce cadre, et Contrairement aux autres méthodes de classification (arbres de décision, réseaux de neurones, algorithmes génétiques, ...etc.) l'algorithme de *k*NN ne construit pas de modèle à partir d'un échantillon d'apprentissage, mais c'est l'échantillon d'apprentissage, la fonction de distance et la fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constituent le modèle [28] [29].

#### 7.5.1 Algorithme de classification par *k-PPV*

**Paramètre** : le nombre *k* de voisins

**Donnée** : un échantillon de *m* exemples et leurs classes

La classe d'un exemple *X* est  $c(X)$

**Entrée** : un enregistrement *Y*

1. Déterminer les *k* plus proches exemples de *Y* en calculant les distances
2. Combiner les classes de ces *k* exemples en une classe *c*

**Sortie** : la classe de *Y* est  $c(Y)=c$  [30].

#### 7.5.2 Comment cela marche-t-il ?

Nous supposons avoir une base de données d'apprentissage constituée de *N* couples « entrée-sortie ». Pour estimer la valeur de sortie d'une nouvelle entrée *x*, la méthode des *k*

plus proches voisins consiste à prendre en compte (de façon identique) les  $k$  échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée  $x$ , selon une distance à définir [18].

Si nous prenons une base d'apprentissage de 100 éléments, Dès que nous recevons un nouvel élément que nous souhaitons classifier, l'algorithme calcule sa distance à tous les éléments de la base. Si cette base comporte 100 éléments, alors il va calculer 100 distances et donc obtenir 100 nombres réels. Si  $k = 25$  par exemple, il cherche alors les 25 plus petits nombres parmi ces 100 nombres qui correspondent donc aux 25 éléments de la base qui sont les plus proches de l'élément que nous souhaitons classifier. La classe attribuée à l'élément à classifier est la classe majoritaire parmi ces 25 éléments [31].

### 7.5.3 Avantages et inconvénients

- **Avantages**

- La qualité de la méthode s'améliore en introduisant de nouvelles données sans nécessiter la reconstruction d'un modèle. Ce qui représente une différence majeure avec des méthodes telles que les arbres de décision et les réseaux de neurones.
- La clarté des résultats : la classe attribuée à un objet peut être expliquée en exhibant les plus proches voisins qui ont amené à ce choix.
- La méthode peut s'appliquer à tout type de données même les données complexes tels que des informations géographiques, des textes, des images, du son. C'est parfois un critère de choix de la méthode PPV car les autres méthodes traitent difficilement les données complexes. Nous pouvons noter, également, que la méthode est robuste au bruit.
- Facile à mettre en œuvre.

- **Inconvénients**

- temps de classification : la méthode ne nécessite pas d'apprentissage ce qui implique que tous les calculs sont effectués lors de la classification. Contrairement aux autres méthodes qui nécessite un apprentissage (éventuellement long) mais qui sont rapides en classification.
- méthode donnera de mauvais résultats Si le nombre d'attributs pertinents est faible relativement au nombre total d'attributs, car la proximité sur les attributs pertinents sera noyée par les distances sur les attributs non pertinents.

- Les performances de la méthode dépendent du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins [28] [32].

### 7.6 L'algorithme des *k*-moyennes (K-Means)

L'algorithme des K-moyennes est dédié aux tâche de clustering, il permet de diviser une population donnée en K groupes homogènes appelés clusters. Le nombre de clusters K est déterminé par l'utilisateur selon ses attentes.

#### 7.6.1 Principe de fonctionnement

Après avoir déterminé un nombre K de clusters nous positionnons les K premiers points (appelés graines) au hasard (nous utilisons en général les K premiers enregistrements). Chaque enregistrement est affecté à la graine dont il est plus proche (en utilisant la fonction de distance). A la fin de la première affectation, la valeur moyenne de chaque cluster est calculée et la graine prend cette nouvelle valeur. Le processus est répété jusqu'à stabilisation des clusters [33] [34].

### 7.7 Algorithme de clustering par K-Means

L'algorithme *k-means* est en 4 étapes :

1. Choisir k objets formant ainsi k clusters
2. (Ré) affecter chaque objet O au cluster  $C_i$  de centre  $M_i$  tel que  $\text{dist}(O, M_i)$  est minimal
3. Recalculer  $M_i$  de chaque cluster (le barycentre)
4. Aller à l'étape 2 s'il faut faire une affectation

#### 7.7.1 Evaluation

Le but de la technique des k-means est le regroupement par similitude des populations statiques, ce qui nécessite de déterminer la qualité des différents clusters. Une évaluation de la qualité pourrait consister à l'étude de la variance de cette population. Alors, nous pouvons dire qu'un cluster possédant d'une part population et d'autre part une variance faible est un cluster solide.

Il est nécessaire d'effectuer d'autres évaluations dans les cas suivants :

- Si la population d'un cluster est trop faible, il sera préférable de grouper ce cluster avec un autre.

- Si un cluster est trop dominant, il pourrait être valable de diviser la population en deux (dans et hors cluster) et de relancer le processus pour chaque sous groupe [33].

### 7.7.2 Avantages et inconvénients

- **Avantages**
  - La méthode résolve une tâche non supervisée, donc elle ne nécessite aucune information sur les données.
  - Technique facile à mettre en œuvre.
  - La méthode est applicable à tout type de données (mêmes textuelles), en choisissant une bonne notion de distance.
- **Inconvénients**
  - La difficulté de trouver une bonne fonction de distance.
  - Un bon choix du nombre  $k$  est nécessaire, un mauvais choix de  $k$  produit de mauvais résultats.
  - La difficulté d'expliquer certains clusters (i.e. attribuer une signification aux groupes constitués) [33] [28].

Les techniques de fouille de données présentées ci-dessus représentent une partie des techniques existantes pour l'extraction des données, et ce grand nombre de techniques est dû qu'ils n'ont pas tous le même objet, et qu'aucun n'est optimal dans tous les cas et qu'ils s'avèrent en pratique complémentaires les uns des autres et qu'en les combinant intelligemment (en construisant ce que l'on appelle des modèles de modèles ou métamodèles) il est possible d'obtenir des gains de performance très significatifs.

## 8 Catégorisation des systèmes du data mining

Les systèmes de data mining peuvent être catégorisés selon plusieurs critères. Parmi les catégorisations existantes nous citons :

- **Classification selon le type de données à explorés :** dans cette classification les systèmes de data mining sont regroupés selon le type des données qu'ils manipulent tel que les données spatiales, les données de séries temporelles, les données textuelles et le Word Wide Web, etc.
- **Classification selon les modèles de données avancés :** cette classification catégorise les systèmes de data mining en se basant sur les modèles de données avancés tel que

les bases de données relationnelles, les bases de données orientées objets, les data warehouses, les bases de données transactionnelle, etc.

- **Classification selon le type de connaissance à découvrir :** cette classification catégorise les systèmes de data mining en s'appuyant sur le type de connaissance à découvrir ou les tâches de data mining tel que la classification, l'estimation, la prédiction, etc.
- **Classification selon les techniques d'exploration utilisées :** cette classification catégorise les systèmes de data mining suivant l'approche d'analyse de données utilisés la reconnaissance des formes, les réseaux neurones, les algorithmes génétiques, les statistiques, la visualisation, orienté-base de données ou orienté-data warehouse, etc. [13].

## 9 Domaines d'application du data mining

La technologie de data mining a une grande importance économique grâce aux possibilités qu'elle offre pour optimiser la gestion des ressources (humaines et matérielles). Les domaines d'application actuels du data mining sont les suivants :

### 9.1 *Le data mining dans le secteur bancaire*

Le secteur bancaire est à la tête de tous les autres domaines industriels grâce son utilisation des techniques du Data Mining dans ses grandes bases de données clients. Bien que les banques ont employées des outils d'analyse statistiques avec un peu de succès pendant plusieurs années, les modèles précédemment invisibles des comportements des clients deviennent maintenant plus clair à l'aide des nouveaux outils du data mining.

Quelques applications du data mining dans ce domaine sont :

- Prédire la réaction des clients aux changements des taux d'intérêt.
- Identifier les clients qui seront les plus réceptifs aux nouvelles offres de produits.
- Identifier les clients "fidèles".
- Déterminer les clients qui pausent le risque le plus élevé de manquer à leurs engagements aux prêts.
- Détecter les activités frauduleuses dans les transactions par cartes de crédit.
- Prédire les clients qui sont susceptibles de changer leurs cartes d'affiliation au cours du prochain trimestre.

- Déterminez les préférences des clients pour les différents modes de transaction à savoir par le biais de guichets ou par l'intermédiaire de cartes de crédit, etc.

### **9.2 Le data mining dans la bio-informatique et la biotechnologie**

La Bio-informatique est un domaine de recherche en développement rapide, qui a des racines aussi bien dans la biologie que dans la technologie d'informations.

Quelques applications du data mining dans ce domaine sont :

- la prédiction les structures de différentes protéines.
- la détermination de la complexité des structures de plusieurs médicaments.

### **9.3 Le data mining dans le marketing direct et le collecte de fonds**

Le marketing direct est un ensemble de techniques permettant d'identifier les consommateurs (particuliers et entreprises) d'un produit stockés dans une base de données, de leur adresser directement et individuellement une proposition commerciale, afin d'obtenir une réponse directe, à laquelle l'entreprise répondra tout aussi directement [22] [23]. Le Marketing Direct est le premier domaine qui a employé les outils du data mining pour son profit.

Les différentes façons dont lesquelles le data mining peut aider dans le marketing direct sont :

- *Les règles d'association* (Analyse du panier de la ménagère) permettent de décider l'emplacement et la promotion des produits dans un magasin et de comparer les résultats entre les différents magasins, entre les clients dans les différents groupes démographiques, entre les différents jours de la semaine, les différentes saisons de l'année, etc. pour augmenter ses ventes.
- Les règles d'association prédictives peuvent être utilisées pour identifier les séries des produits qui s'achètent ensembles.

Les data mining peut être utilisé dans la collecte de fonds (par exemple : les organisations bénévoles, collecte de fonds dans les élections) par exemple par :

- Rassembler des données illimitées sur les différents donateurs, les bénévoles, les perspectives, les agences et les membres des différentes fiducies dans une base de données centrale. Ces données sont gérées et utilisées par des outils du data mining pour construire et améliorer les relations à long terme avec eux.

- Définition des classes des récompenses, les critères d'attribution et permettre au système de se référer aux données pour générer automatiquement des prix comme une reconnaissance des efforts exceptionnels des donateurs.

### 9.4 Le data mining dans la détection de fraude

La fouille de données est largement appliquée dans des processus de détection de fraude divers tel que :

- Détection de fraude de cartes de crédits.
- Détection de fraude dans les listes des électeurs en utilisant les réseaux de neurones en combinaison avec le data mining.
- La détection des fraudes dans les demandes de passeport par la conception d'un système de diagnostic par apprentissage en ligne.
- Détection de fausses demandes de remboursement médicale.

### 9.5 Le data mining dans la gestion de données scientifiques

Quelques exemples de la fouille de données dans l'environnement scientifique sont :

- **Les études sur les changements climatiques du globe:** Il s'agit d'un domaine de recherche *chaud* et est essentiellement un exercice de vérification axée sur l'exploitation. À travers les données climatiques qui ont été recueillies au fil des siècles et qui sont en train d'être étendues dans le passé lointain et, en même temps, à travers des activités telles que l'analyse des échantillons de carottes de glace de l'Antarctique, des différents modèles de prédiction ont été proposées pour les futures conditions climatiques.
- **Les études sur les bases de données de géophysique à l'Université d'Helsinki:** Ils ont publié une analyse scientifique des données sur l'agriculture et l'environnement. En conséquence, ils ont optimisé le rendement des cultures, tout en réduisant au minimum les ressources fournies. Afin de réduire au minimum les ressources, ils ont identifié les facteurs qui influent sur le rendement des cultures, comme les engrais chimiques et les additifs (phosphate), Le contenu d'humidité et le type du sol.

### **9.6 Le data mining dans le secteur des assurances**

Les compagnies d'assurance peuvent bénéficier des méthodes du data mining, qui aident les entreprises à réduire les coûts, augmenter les profits, de conserver les clients actuels, d'acquérir de nouveaux clients, et développer de nouveaux produits.

Cela peut être fait par le biais de:

- Evaluation du risque d'un bien assuré prenant en compte les caractéristiques du bien et de son propriétaire.
- Formulation des modèles statistiques des risques d'assurance.
- Utilisation du modèle de l'exploitation de Poisson / Log-normale afin d'optimiser les polices d'assurance.

### **9.7 Le data mining dans la télécommunication**

À nos jours, toute activité de télécommunication a utilisé une la technique de data mining.

- Analyse des achats de services de télécommunications.
- Prédiction de modèles d'appels téléphoniques.
- Gestion des ressources et de trafic réseau.
- Automatisation de la gestion du réseau et de la maintenance en utilisant l'intelligence artificielle pour diagnostiquer et réparer les problèmes de transmission du réseau, etc.

### **9.8 Le data mining dans la médecine et la pharmacie**

Quelques exemples de l'usage médicaux et pharmaceutiques des techniques de Data Mining pour l'analyse de bases de données médicales.

- Prédiction de présence de maladies et/ou de complications.
- Le choix d'un traitement pour le cancer.
- Choix des antibiotiques pour des infections.
- Le choix d'une technique particulière (de sutures, matériel de suture, etc.) dans une des procédures chirurgicales.
- Approvisionnement des médicaments les plus fréquemment prescrits.

### **9.9 Le data mining dans le commerce au détail**

Les techniques du data mining ont été très utiles pour le CRM (Customer Relationship Marketing) en développant des modèles pour :

- La prédiction de la propension du client à acheter.
- L'évaluation des risques pour chaque transaction.
- Connaître la distribution et l'emplacement géographique des clients.
- L'analyse de la fidélité des clients dans les opérations à base de crédit.
- Évaluation de la menace concurrentielle dans une région.

### **9.10 Le data mining dans le e-commerce et le World Wide Web**

Quelques façons d'utilisation des outils du data mining dans le e-commerce sont :

- En formulant des tactiques du marché dans les opérations de business.
- En automatisant des interactions d'affaires avec des clients, pour que les clients puissent traiter avec tous les acteurs dans la chaîne d'approvisionnement.

La détermination de la taille d world Wide Web est extrêmement difficile. En 1999, elle a été estimée à 350 milliards de pages avec un taux de croissance de 1 million de pages / jour. En considérant le World Wide Web comme la plus grande collection de bases de données, le Web mining peut être fait par les façons susdites.

### **9.11 Le data mining dans le marché boursier et l'investissement**

L'évolution rapide de la technologie informatique au cours des dernières décennies a facilité l'investissement par des professionnels (et des amateurs), avec la possibilité d'accéder et d'analyser d'énormes quantités de données financières. Les outils du data mining sont utilisés pour :

- Aider les spécialistes du marché boursier à prédire les mouvements du prix des actions.
- Le data mining des anciens des prix et des variables liées aide à découvrir des anomalies de marché boursier comme le scandale hawala.

### **9.12 Le data mining dans l'analyse de chaîne d'approvisionnement**

Les techniques du data mining ont trouvé une large application dans l'analyse des chaînes d'approvisionnements. Des exemples d'utilisation du data mining par les fournisseurs sont :

- Analyser le processus des données pour gérer l'évaluation d'acheteur.
- L'extraction des données de paiement avec l'utilité de mettre à jour la politique tarifaire [21].

## **10 Conclusion**

Le data mining est l'extraction d'informations prédictives cachés dans de grandes base de données. C'est une technologie nouvelle et puissante qui donne la possibilité aux entreprises de se concentrer sur les informations les plus importantes dans leurs data warehouses. Les outils du data mining peuvent prédire les futurs tendances et actions, permettant de prendre les bonnes décisions. C'est ce qui rend le data mining la technologie la plus importantes.

Le chapitre suivant comprendra une étude des travaux exploitant les systèmes multi agents pour le data mining.

## **Chapitre II**

---

*Étude des travaux exploitant les SMA pour le data mining.*

### 1 Introduction

La technologie des systèmes multi-agents a suscité beaucoup d'excitations dans les dernières années grâce aux promesses qu'elle donne commettant un nouveau paradigme de conception et d'implémentation des systèmes logiciels. Ces promesses sont particulièrement attrayantes pour la création des logiciels fonctionnant dans des environnements distribués et ouverts [35].

Dans ce chapitre nous présentons la contribution des systèmes multi-agents dans le domaine de data mining. Avant de commencer, il serait intéressant de comprendre quelques concepts comme l'intelligence artificielle distribuée, nous présentons par la suite l'évolution de l'aspect individuel (le comportement d'un agent seul) vers l'aspect collectif (son comportement dans une société d'agents).

### 2 L'Intelligence Artificielle distribuée (IAD)

L'*intelligence artificielle* est une science dédiée à résoudre les problèmes qui ne peuvent être résolus par l'informatique traditionnelle, elle réunit beaucoup de sciences tel que l'informatique, la psychologie, et la philosophie afin de produire des machines (ordinateurs) qui peuvent être qualifié d'intelligentes.

Au fil du temps, l'Intelligence Artificielle (IA) classique a montré des limites à résoudre des problèmes complexes. Dans le but de combler ces limites les chercheurs ont senti le besoin de passer du comportement individuel aux comportements collectifs et la nécessité de distribuer l'intelligence sur plusieurs entités.

L'ensemble des études couvrant les comportements collectifs constituent le domaine de l'Intelligence Artificielle Distribuée (IAD).

Techniquement, l'IAD est une branche de l'IA qui propose de remplacer les logiciels conçus de manière centralisée par des logiciels basés sur l'interaction de composants logiciels plus élémentaires. Elle repose sur le principe de '*diviser pour régner*' qui facilite la mise au point et le test des systèmes de résolution de problèmes et permettre une meilleure réutilisabilité des composants.

L'IAD s'articule autour de trois axes :

- RDP: La Résolution Distribuée des Problèmes qui s'intéresse à la manière de diviser un problème en un ensemble d'entités distribuées et coopérantes et à la manière de partager la connaissance d'un problème afin d'en obtenir la solution.
- IAP: L'IA Parallèle qui développe des langages et des algorithmes parallèles visant à l'amélioration des performances des systèmes d'IA.
- SMA : Les Systèmes Multi Agents qui privilégient une approche décentralisée de la modélisation et mettent l'accent sur les aspects collectifs des systèmes [36].

### 3 Concept d'agent

#### 3.1 Définitions

D'un point de vue purement informatique, un agent peut être défini comme un objet (au sens des langages objets) dont le comportement est décrit par un "script" (fonction principale *main*), disposant de ses propres moyens de calcul, et qui peut se déplacer de places en places (une place pouvant être un site informatique distant du site originel de l'agent) pour communiquer avec d'autres agents. De par son "script", l'agent est capable de suivre un comportement de vie qui lui sera inculqué au moment de l'implémentation et qui lui permettra d'avoir comme principale caractéristique d'être entièrement autonome [37].

Une autre définition est la suivante (d'après Jacques Ferber): un agent est une entité physique ou virtuelle

- capable d'agir dans un environnement,
- peut communiquer directement avec d'autres agents,
- mue par un ensemble de tendances (sous la forme d'objectifs individuels ou d'une fonction de satisfaction, voire de survie, qu'elle cherche à optimiser),
- possédant des ressources propres,
- capable de percevoir (mais de manière limitée) son environnement,
- ne disposant que d'une représentation partielle de cet environnement (et éventuellement aucune),
- possédant des compétences et offrant des services,
- qui peut éventuellement se reproduire,

- dont le comportement tend à satisfaire ses objectifs, en tenant compte des ressources et des compétences dont elle dispose, et en fonction de sa perception, de ses représentations et des communications qu'elle reçoit [38].

### **3.2 Différence entre objet et agent**

Les concepts objet et agent distribué sont très proche mais il n'en est pas de même de l'agent comme entité intentionnelle. En effet les objets n'ont pas but de recherche de satisfaction et le mécanisme d'envoi de messages se résume à un simple appel de procédure. Il n'y a pas de langage de communication à proprement parler. Les mécanismes d'interaction sont donc à la charge du programmeur.

La différence essentielle entre un objet et un agent est qu'un objet est défini par l'ensemble des services qu'il offre (ses méthodes) et qu'il ne peut refuser d'exécuter si un autre objet le lui demande. Par contre, les agents disposent d'objectifs leur donnant une autonomie de décision vis à vis des messages qu'ils reçoivent. Par ailleurs, ils établissent des interactions complexes faisant intervenir des communications de haut niveau.

Du fait, un agent peut être considéré comme un objet doté de capacités supplémentaires : recherches de satisfactions (intentions, pulsions) d'une part, et d'autre part la communication à base de langages plus évolués (actes de langages pour les agents cognitifs, propagation de stimuli pour des agents réactifs). Inversement, un objet peut être considéré comme un agent "dégénéré" devenu un simple exécutant, tout message étant considéré comme une requête [37].

### **3.3 Types d'agents**

Une des caractéristiques discriminantes des agents est la représentation et le raisonnement sur l'environnement ( le monde extérieur et les autres agents ), suivant cette caractéristique , nous trouvons deux classes différentes.

#### **3.3.1 Les agents cognitifs**

Un agent cognitif est un agent qui possède une représentation explicite de son objectif et de son environnement. Les actions qu'il effectue pour atteindre son objectif sont le résultat d'un raisonnement sur l'état de l'environnement. Généralement un système cognitif

comprend un petit nombre d'agents, chacun est assimilable à un système expert plus au moins complexe. Dans ce cas on parle d'agent de forte granularité.

### 3.3.2 Les agents réactifs

Un agent réactif est un agent dont le comportement répond uniquement à la loi stimulus/action, le stimulus étant un élément de l'environnement ( action, message, situation, etc). Généralement un système réactif comprend un grand nombre d'agents de faible granularité. Ces agents n'ont pas forcément un but explicite à atteindre. Par contre, ils peuvent mettre en œuvre un raisonnement complexe sur leur état interne pour réaliser leurs actions [39].

Systèmes d'agents cognitifs	Systèmes d'agents réactifs
Représentation explicite de l'environnement	Pas de représentation explicite
Peut tenir compte de son passé	Pas de mémoire de son histoire
Agents complexes	Fonctionnement Stimulus/action
Petit nombre d'agents	Grand nombre d'agents

Tableau 1 : comparaison entre agents cognitifs et agents réactifs

## 4 Les Systèmes Multi-Agents

### 4.1 Définitions

Les systèmes multi-agents (SMA) mettent en œuvre un ensemble de concepts et de techniques permettant à des logiciels hétérogènes, ou à des parties de logiciels, appelés "agents" de coopérer suivant des modes complexes d'interactions. Ils apportent une nouvelle solution au concept de modèle et de simulation dans les sciences de l'environnement, en proposant de représenter directement les individus, leurs comportements et leurs interactions sous la forme d'agents, et la quantité d'individus d'une espèce donnée sera le résultat de la confrontation (coopération, lutte, reproduction) des comportements de tous les individus représentés dans le système [37].

Et selon Jacques Ferber, un SMA est système composé de:

- Un environnement E, c'est-à-dire un espace disposant généralement d'une métrique.

- Un ensemble  $O$  d'objets situés, c'est-à-dire que, pour tout objet, il est possible, à un moment donné, d'associer une position dans  $E$ . Ces objets sont passifs, c'est-à-dire qu'ils peuvent être perçus, créés, détruits et modifiés par les agents.
- Un ensemble  $A$  d'agents, qui sont des objets particuliers ( $A \subseteq O$ ), lesquels représentent les entités actives du système.
- Un ensemble de relations  $R$  qui unissent des objets (et donc des agents) entre eux.
- Un ensemble d'opérations  $Op$  permettant aux agents de  $A$  de percevoir, produire, consommer, transformer et manipuler des objets de  $O$ .
- Des opérateurs appelés les lois de l'univers chargés de représenter l'application de ces opérations et la réaction du monde à cette tentative de modification [38].

### **4.2 Quand utiliser un SMA?**

- Quand le problème est trop complexe mais peut être décomposé.
- Quand il n'y a pas de solution générale ou lorsque celle-ci est trop coûteuse en CPU.
- A des fins de modélisation (populations, structures moléculaires, tas de sables...)
- Quand on peut paralléliser le problème (gain de temps).
- Quand on veut une certaine robustesse (redondance).
- Quand l'expertise provient de différentes sources.
- Quand les données, contrôles, ressources sont distribués.
- Quand le système doit être adaptatif [40].

## **5 Quelques travaux exploitant les SMAs pour le DM**

### **5.1 Approche SMA pour la Segmentation Markovienne des Tissus et Structures Présents dans les IRM Cérébrales**

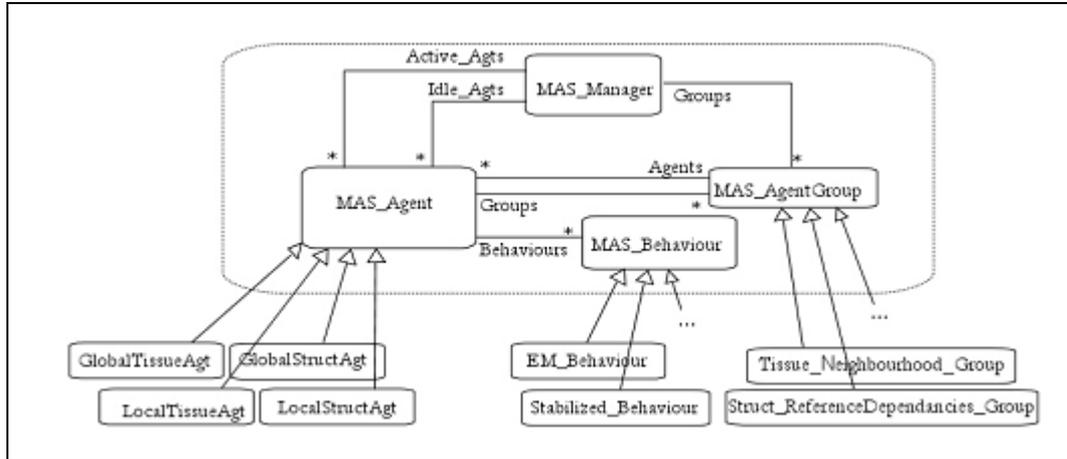
La segmentation précise des tissus et des structures présents dans des IRM (Imagerie par Résonance Magnétique) anatomique est indispensable à de nombreuses applications. La majorité des approches existantes considèrent en premier lieu la segmentation des tissus puis ensuite celle des structures de manière indépendante. L'approche de segmentation markovienne coopérative des tissus et des structures se fonde sur le raffinement mutuel des segmentations en tissus et en structures. La connaissance a priori nécessaire à la segmentation des structures est apportée par une description floue de l'anatomie cérébrale. Elle est intégrée dans le cadre markovien via l'expression d'un champ externe. La segmentation des tissus intègre dynamiquement l'information structure via un autre champ externe, permettant de

raffiner l'estimation des modèles d'intensités. Cette approche est implémentée dans un environnement multi-agents. L'évaluation est réalisée à la fois sur des images fantômes et sur des images réelles acquises par l'appareil 3 Tesla.

### 5.1.1 Implémentation SMA

L'approche est implémentée dans un système multi-agents spécifique à l'application. Il se fonde sur une approche décentralisée et à partage de mémoire entre agents, et s'inspire du modèle conceptuel Agent/Groupe/Comportement (voir la Figure ci-dessus) de MadKit (Multi-Agents Développement Kit: conçue selon le modèle d'organisation Alaadin AGR (Agent/Group/Rôle) [41]).

Dans ce modèle un agent est une entité qui possède plusieurs comportements et exécute un seul à un moment donné suivant son état. Le groupe est le rassemblement virtuel d'agents communiquant et coopérant entre eux afin de réaliser leur tâche. Dans la notion de groupe deux caractéristiques importantes doivent être respectées (1) un agent peut appartenir à plusieurs groupes et (2) les groupes peuvent se recouvrir.



**Figure 9: Modèle Agent/Groupe/Comportement du système Multi-Agents.**

Nous définissons quatre types d'agents et cinq types de groupes décrits dans les sections suivantes.

### 5.1.1.1 Agents du système

- a) **Agent Global des Tissus** : une seule instance de cet agent est créée en possédant un seul comportement qui est la réalisation du partitionnement cubique régulier du volume et la création d'un agent local de segmentation par contexte. Il calcule aussi un modèle global d'intensité des tissus avec l'algorithme Fuzzy C-Mean (FCM) afin d'agencer l'ordre d'exécution des agents locaux. Une fois son traitement réalisé, il se met en sommeil définitivement.
- b) **Agent Global des Structures** : pareil que l'agent global des tissus, une seule instance de cet agent est créée qui se charge de construire les  $L = 9$  agents structures. Il leur fournit la description des relations spatiales stables entre structures puis se met en sommeil définitivement.
- c) **Agent Local de Segmentation des Tissus** : Chacun de ces agents est associé à un unique contexte de segmentation tissu. Il possède 3 comportements dont les enchaînements possibles sont décrits sur la Figure 10 :

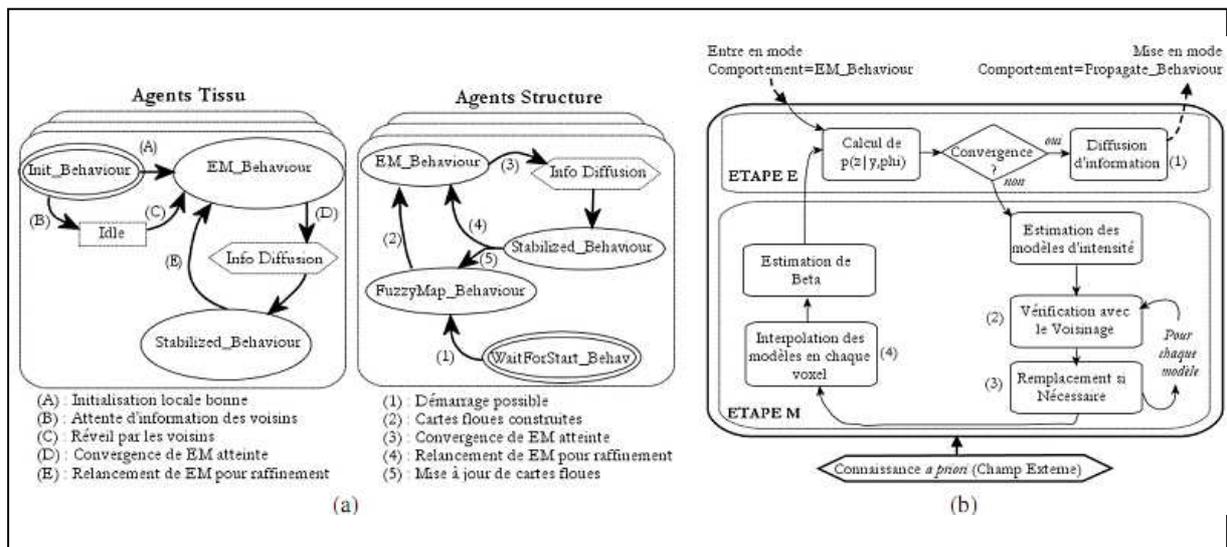


Figure 10: Diagrammes fonctionnels : enchaînements possibles des comportements des agents tissu et structure (a) et détails du comportement EM Behaviour des agents locaux (b) qui correspond à un EM classique avec ajout des étapes (1), (2), (3) et (4) pour réaliser les mécanismes de coopération.

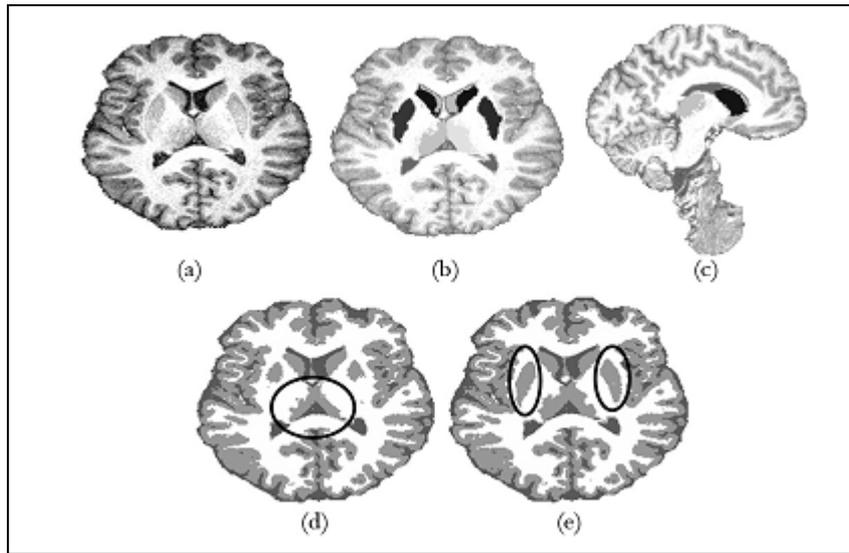
- **Init Behaviour** : lancé à la création de l'agent, ce comportement réalise une première estimation des modèles d'intensités locaux nécessaire à l'initialisation de EM. Nous avons choisi d'utiliser l'algorithme FCM pour cette première estimation. Les agents dont les modèles d'intensités locaux sont trop éloignés des modèles globaux se mettent en sommeil (transition B). Les autres rentrent directement en mode EM Behaviour (transition A).
  - **EM Behaviour** : c'est le comportement principal qui réalise la segmentation Markovienne locale et coopérative. La Figure 10b détaille son fonctionnement: vérification des modèles avec le voisinage, interpolation en chaque voxel (contraction de « volumetric pixel » est un pixel en 3D [18]) et intégration d'un champ externe. Le paramètre du modèle Markovien  $\beta$  (un paramètre ajustant la force des interactions spatiales) n'est pas estimé mais considéré comme  $\beta = 1/T$  avec T une température décroissante comme proposé dans [42]. la convergence de EM est considérée atteinte lorsque la modification relative de la log-vraisemblance entre deux étapes E devient suffisamment petite (expérimentalement  $10^{-5}$  donne de bons résultats). L'agent diffuse alors l'information : les agents voisins (tissu et structure) partageant le territoire sont réveillés afin de propager l'éventuelle modification des modèles d'intensité.
  - **Stabilized Behaviour** : le comportement qui correspond à un état de sommeil de l'agent. Quand-t-il sera réveillé (par un agent voisin ou un agent structure), l'agent tissu vérifie ses modèles avec le voisinage et relance la segmentation si nécessaire (transition E) ou se remet en mode sommeil.
- d) Agent local de Segmentation des Structures** : pour segmenter chaque structure un agent de segmentation structure est défini. Il est dynamiquement localisé grâce à la connaissance anatomique floue et opportuniste : il commence la segmentation le plus tôt possible et enrichit sa connaissance au fur et à mesure. Le fonctionnement se base sur les 4 comportements suivants (voir Figure 10a) :
- **WaitForStart Behaviour** : ce comportement est lancé à la création de l'agent, il synchronise le début de la segmentation : il vérifie premièrement que les agents tissus partageant le territoire avec l'agent structure ont déjà estimé leur modèle d'intensité et deuxièmement que les cartes des relations spatiales indispensables au démarrage de l'agent sont calculables.

- FuzzyMap Behaviour : l'agent construit ou met à jour les cartes 3D floues de ses relations spatiales, pour ensuite passer au comportement EM Behaviour (transition 2).
- EM Behaviour : c'est le comportement qui étiquette la structure cible grâce aux modèles d'intensités estimés par les tissus et à l'introduction de l'information anatomique des relations spatiales. Après convergence de EM il diffuse l'information : il réveille si nécessaires d'autres agents structures en mode FuzzyMap Behaviour afin de mettre à jour leur connaissance anatomique et réveille les agents tissu qui partagent le territoire pour leur permettre de mettre à jour leur champ externe.
- Stabilized Behaviour : le comportement qui correspond à un état de sommeil de l'agent. Au réveil, si des structures identifiées comme référence dans des relations spatiales ont mise à jour leur segmentation, il passe en comportement FuzzyMap Behaviour (transition 5). Si les modèles d'intensité des tissus ont été modifiés, l'agent repasse en comportement EM Behaviour (transition 4). Sinon il se remet en sommeil.

Les agents locaux tissu et structure coopèrent via le mécanisme de Groupe du système multi-agents.

### 5.1.1.2 Les Groupes

- Tissue Neighbourhood Group* : chaque agent tissu local détermine le groupe de ses agents voisins avec qui il va pouvoir coopérer : réveil des voisins et vérification des modèles locaux.
- Struct ReferenceDependancies Group* : chaque agent structure local détermine le groupe des agents structures qui présentent des référence pour ses relations spatiales et qui doivent donc fournir leur segmentation pour le calcul des cartes floues.
- Struct DependanciesToWakeUp Group* : chaque agent structure local A définit le groupe des agents utilisant A comme référence, donc les agents à réveiller après une mise à jour de la segmentation.
- Struct TissueDependancies Group* : chaque agent structure local détermine le groupe des agents tissus avec qui il partage le territoire et qui doivent lui fournir les modèles d'intensités.
- Tissue StructDependancies Group* : chaque agent tissu local détermine le groupe des agents structures avec qui il partage le territoire et qui doivent lui fournir la segmentation de ces structures pour calculer le champ externe.



**Figure 11: Evaluation visuelle : Image réelle 3T (a), Segmentation des structures (b-c), Segmentation des tissus sans coopération (d) et avec l'approche SMA (e).**

### 5.1.2 Evaluation sur images réelles acquises à 3T

L'approche a été évaluée visuellement sur des images réelles acquises à 3 Tesla. La Figure 11 montre la segmentation structure obtenue (b-c) et la nette amélioration de la segmentation des tissus obtenue grâce à la coopération tissu-structure, particulièrement au niveau des putamens et thalamus (d-e) [43].

## 5.2 Une approche SMA de L'agrégation et de la coopération des classifieurs

L'objectif général des méthodes d'apprentissage supervisé (précisément la tâche de classification) est la construction à partir de la base d'apprentissage, des classifieurs permettant de reconnaître la classe d'un individu d'une population donnée. Le but sera alors la détermination d'une procédure de prévision permettant de prédire la classe de tous les éléments d'une population en ayant comme point de départ une fonction de classement construite sur un échantillon. L'échantillonnage permet donc de tirer des conclusions au sujet d'un tout en y examinant une partie. Cette approche propose un système de classification supervisé fondé sur le groupement de plusieurs classifieurs qui sont amenés à contribuer à la même tâche de classification. Les Systèmes Multi-Agents fournissent une plate-forme adéquate pour mettre en place ce système de classification et définir les différentes stratégies, en se fondant sur les interactions et la communication entre les classifieurs. Ces interaction-coopérations se font à travers les échantillons qui sont permutées ou qui sont mélangées. Les

résultats des expérimentations effectuées sur une base de données de type benchmark sont très satisfaisants pour les méthodes qui mélangent les échantillons et permettent de tirer des conclusions intéressantes sur des méthodes qui les permutent.

### **5.2.1 Une approche fondée sur les SMA**

La combinaison différents systèmes pour une même tâche est une des voies explorées depuis plusieurs années pour améliorer les capacités de généralisation des systèmes d'apprentissage.

Cette approche d'agrégation-coopération entre algorithmes de classification consiste à faire communiquer entre eux des agents (chacun représentant une méthode de classification) qui ont pour but final de se mettre d'accord entre eux (à quelque chose près) dans l'espoir de minimiser le plus possible l'erreur de classification.

Chaque méthode de classification a son propre algorithme de construction de classifieur, et l'échantillonnage aléatoire donne lieu à un échantillon propre à chaque méthode au départ. Il est possible aussi de supposer que chaque méthode a son propre algorithme d'échantillonnage.

La philosophie de la méthode d'agrégation-coopération consiste à dire que les méthodes ont intérêt à coopérer en échangeant leurs échantillons (et en particulier en donnant naissance à de nouveaux échantillonnage) afin d'améliorer leurs performances.

Le principal but de cette méthode est de trouver la base d'échantillonnage la plus adaptée à chacun des classifieurs pour obtenir le meilleur résultat possible.

### **5.2.2 Fonctionnement**

Deux méthodes d'échantillonnage sont mises en place pour optimiser la phase d'apprentissage du système. Ces méthodes défendent la thèse de la coopération et de l'échange des connaissances.

La première méthode (Méthode A : permutation des échantillons) se fonde sur l'idée que s'il y avait de mauvais résultats, il fallait tout revoir : abandonner les connaissances et les méthodes utilisées et récupérer les connaissances et les méthodes des autres.

## Chapitre II : Étude Des Travaux Exploitant Les SMAs Pour Le Data Mining

A partir de la base d'échantillonnage initiale, N bases sont générées par un tirage aléatoire simple avec remise. Les N classifieurs sont entraînés sur chacune des N bases obtenues. Chaque classifieur va communiquer ses résultats aux autres et, en fonction de quelques critères déjà définis (par exemple : l'erreur de classification), une permutation ou un échange des bases entre les différents classifieurs sera effectuée.

La deuxième méthode (Méthode B : mélange des échantillons) se fonde sur l'idée que s'il y avait de mauvais résultats, c'est parce qu'il y avait quelques méthodes ou connaissances qui doivent être amélioré ou échanger avec d'autres.

A partir de la base d'échantillonnage initiale, N bases (voir Figure 12) sont générées par un tirage aléatoire simple avec remise. Les N classifieurs sont entraînés sur chacune des N bases obtenues. Chaque classifieur va communiquer ses résultats aux autres et, en fonction de quelques critères déjà définis entre deux classifieurs, un tirage aléatoire avec remise sera effectué sur l'ensemble des deux bases de chacun des classifieurs. La nouvelle base obtenue servira comme base d'apprentissage à l'un des deux classifieurs.

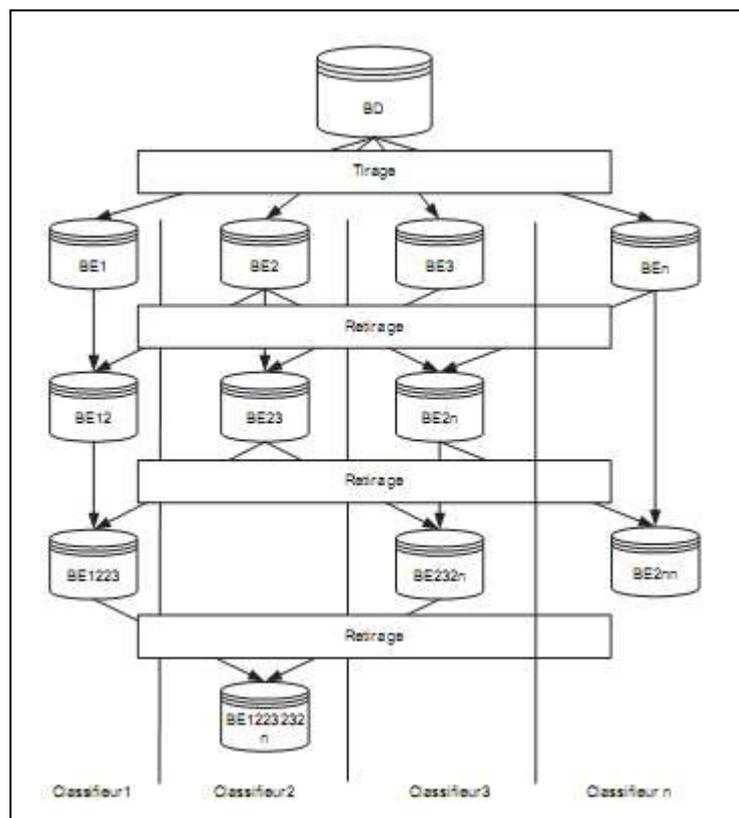


Figure 12: Evolution de la base d'échantillonnage (BE).

### 5.2.3 Synthèse

En résumé, la coopération entre les différents agents (classifieurs) peut induire une réduction au niveau de l'espace de représentation et, par conséquent, sur la base d'apprentissage. Le mixage ou la permutation suivant certains critères (exemple: taux de classification, ...) des bases d'apprentissage entre les différents classifieurs est une des méthodes utilisées par ce système (voir figure 13).

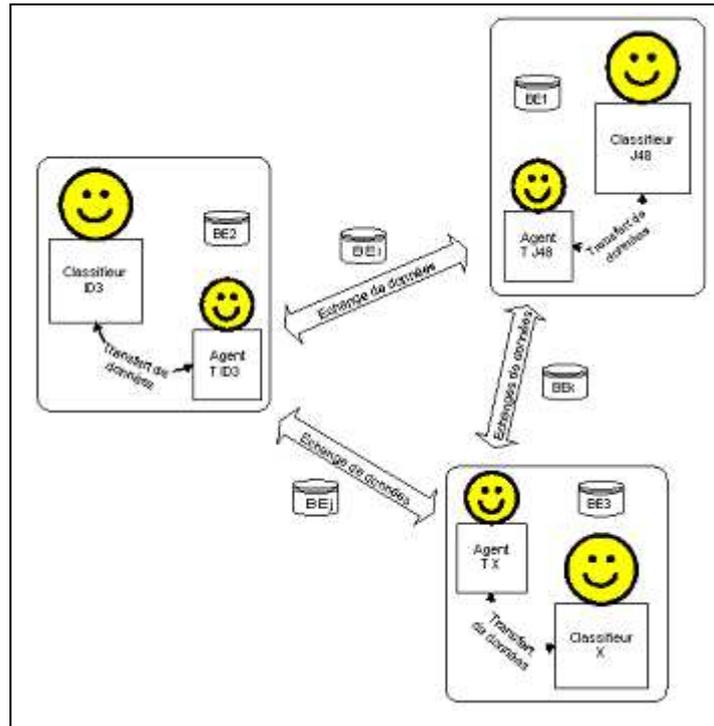


Figure 13: Communication et transfert des données entre classifieurs.

Ce système de classification est basé sur le modèle Aalaadin et repose donc sur les concepts d'agent, de groupe et de rôle (voir figure 14).

Un seul groupe a été défini dont tous les agents y appartiennent. Quatre types d'agents ont été établis :

- L'agent superviseur :** qui contrôle et régule les différents échanges d'informations, le début et la fin du traitement.
- Les agents classifieurs :** qui effectuent une classification particulière (ex : ID3, C4.5, NaiveBayes, VotedPerceptron, etc.).
- Les agents de traitements :** qui effectuent des traitements sur différentes données reçues de la part des autres agents (échange de données, ré-échantillonnage, etc.).

d) *L'agent de calcul* : qui récupère et stocke tous les résultats des différents agents de classification [44].

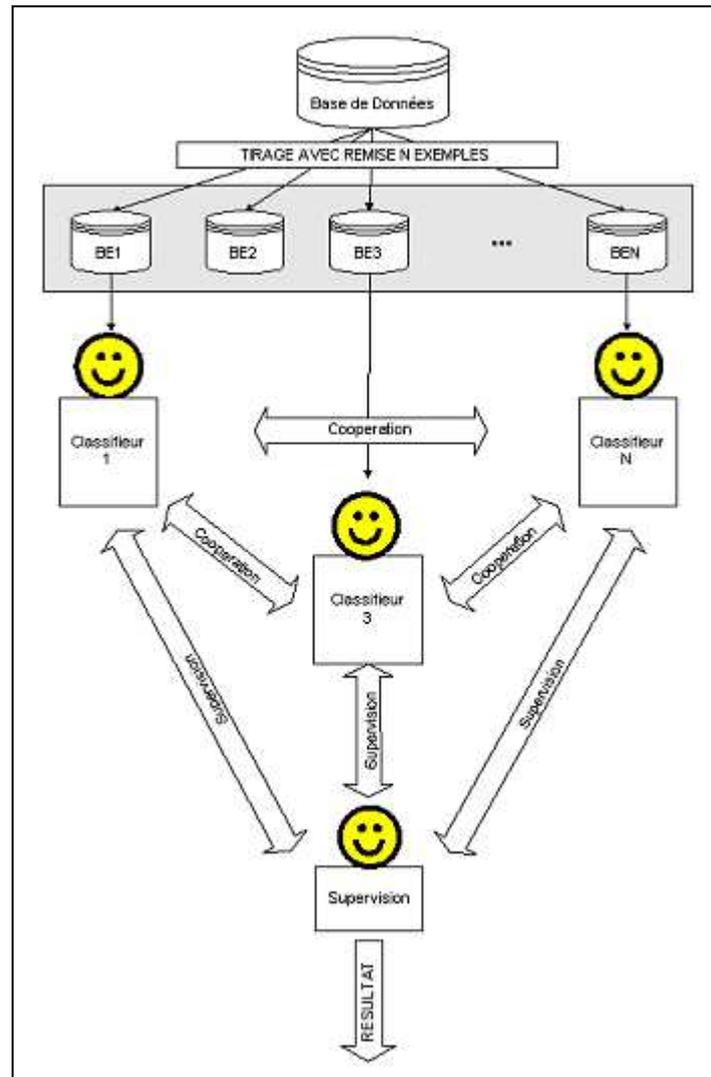


Figure 14: Architecture du système proposé [44].

### 5.3 Une approche pour l'extraction des règles d'association spatiales basée Multi-Agent : RASMA

Les techniques de Fouille de Données ont été étendues aux données spatiales ce qui a donné naissance à la Fouille de Données Spatiales (FDS). La FDS est l'extraction des connaissances implicites, les rapports spatiaux et d'autres modèles qui ne sont pas explicitement stockés dans une base de données géographiques.

Le premier algorithme d'extraction des règles d'association spatiales (RAS), a été proposé par Koperski [45]. Des algorithmes visant à améliorer les performances de cet

## Chapitre II : Étude Des Travaux Exploitant Les SMAs Pour Le Data Mining

algorithmes ont été présentés dans la littérature. Ceci de point de vue type ou nombre de règles extraites, mais l'algorithme d'extraction de règles d'association spatiales nécessite toujours un temps d'exécution important. L'algorithme des Règles d'Association Spatiales basé Multi-Agent (RASMA) vise à améliorer les performances de l'algorithme RAS en terme de temps d'exécution. Une description détaillée des agents et de leurs rôles est donnée. Les résultats retournés par chaque agent et les messages échangés entre les agents sont exposés. Les résultats expérimentaux, les tests effectués sur RAS et RASMA et l'évaluation de l'expérimentation sont ensuite présentés.

La figure 15 montre les étapes de l'extraction des règles d'association spatiales dans RAS. Les étapes de RAS ont été de manière à distinguer les étapes qui peuvent s'exécuter en même temps et les différentes interactions entre ces étapes. L'extraction des règles d'association spatiales se divise en deux étapes. L'étape de la recherche des données et l'étape de la génération des règles.

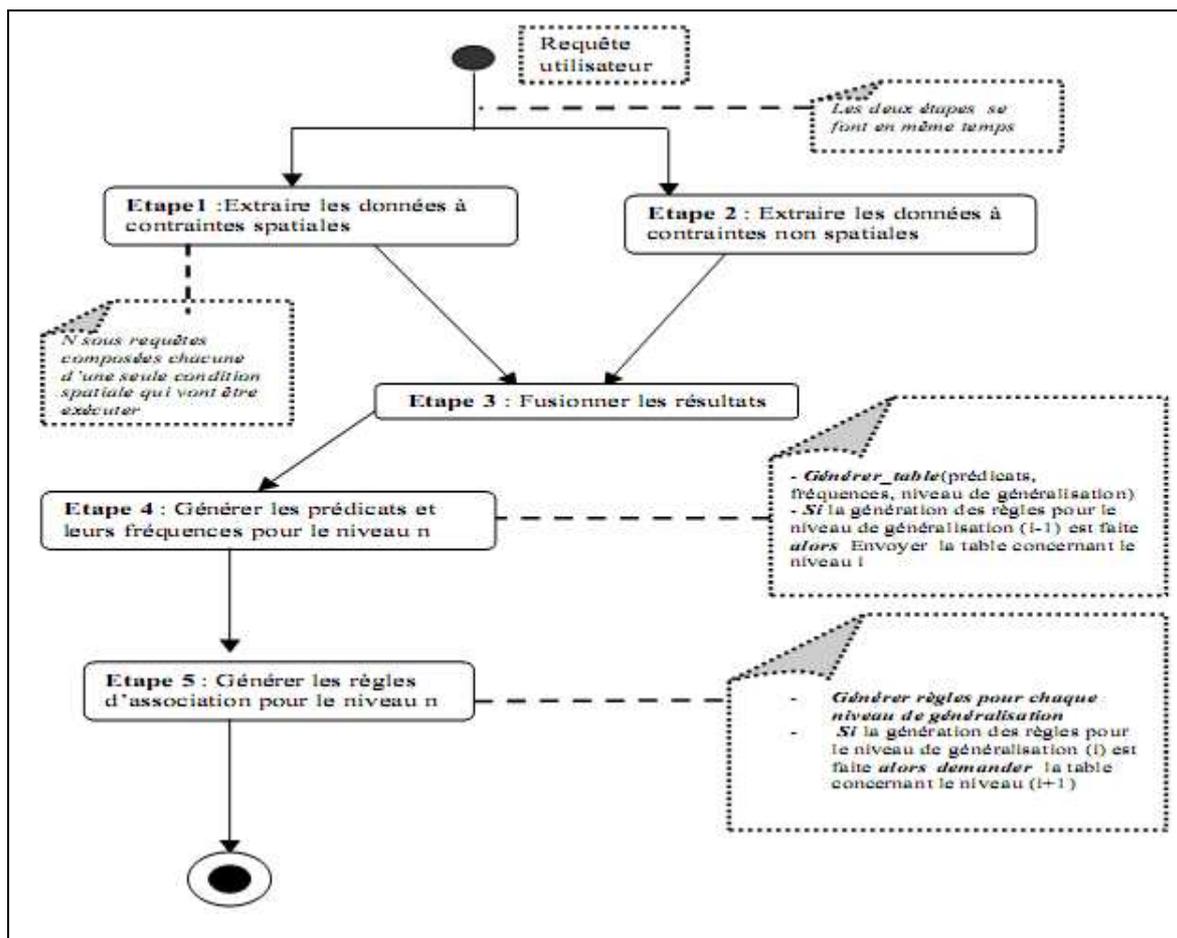


Figure 15: Algorithme d'extraction des règles d'association spatiales.

## **Chapitre II : Étude Des Travaux Exploitant Les SMAs Pour Le Data Mining**

---

La première étape se décompose en trois sous étapes :

1. La recherche des données sous conditions non spatiales ;
2. La recherche des données sous conditions spatiales ;
3. La jointure entre les données recherchées.

La deuxième étape, se fait pour chaque niveau de généralisation. Cette étape est composée de deux sous étapes :

4. La génération des  $k$ -prédicats fréquents ;
5. La génération des règles.

Les étapes 1 et 2 et les étapes 4 et 5 s'exécutent en même temps. Notons également que l'étape 1 peut se diviser en  $n$  sous étapes pour des données différentes et dont le traitement est le même.

Cette décomposition permet de dégager les traitements qui peuvent s'exécuter en même temps et les différentes interactions entre les étapes, ce qui donne l'idée de profiter de ce point pour améliorer le temps d'exécution de l'algorithme RAS. La technique qui offre la possibilité de réaliser ces tâches proposées par l'approche est le système multi-agent. Cette contribution a donné naissance à une approche d'extraction des Règles d'Association Spatiales basée Multi-Agent (RASMA).

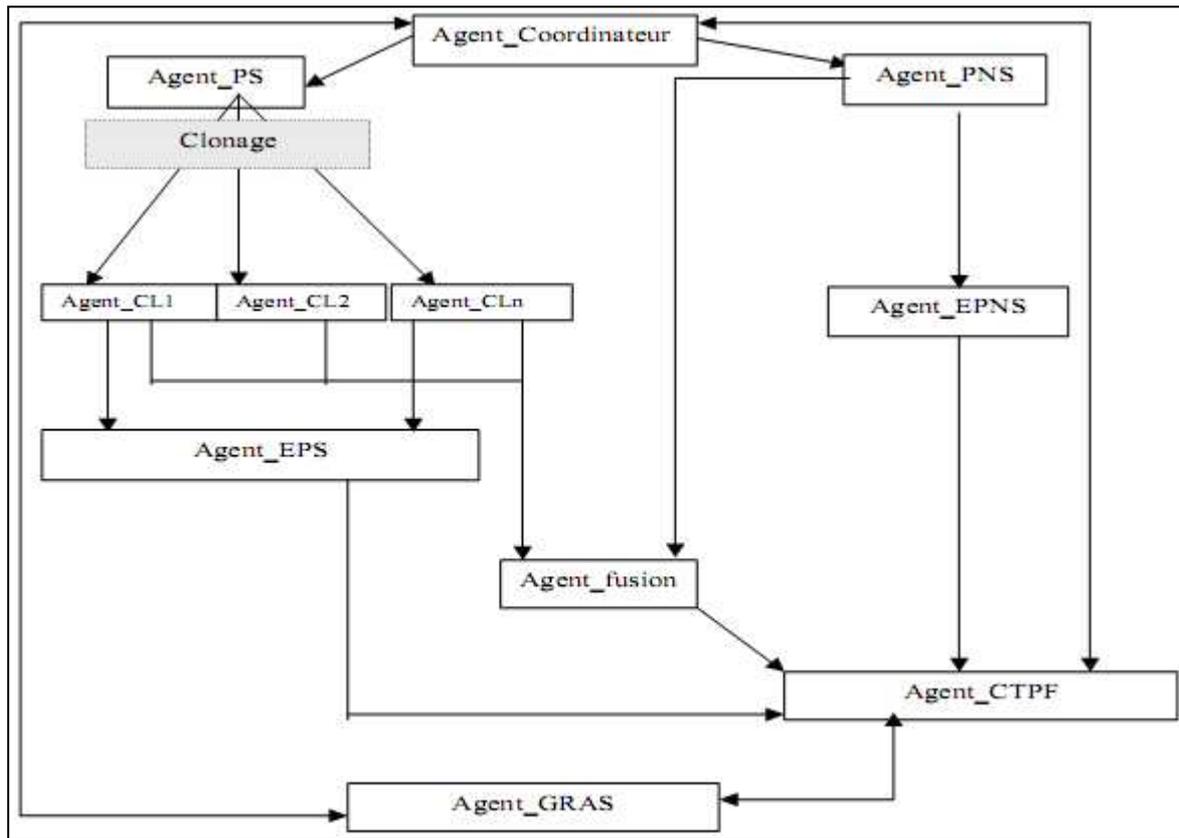


Figure 16: Architecture du système multi-agent RASMA.

### 5.3.1 Architecture de RASMA

L'architecture de RASMA englobe huit agents, comme le montre la figure 16. Dans RASMA l'agent peut réaliser le même traitement  $n$  fois et en même temps, mais sur des données différentes en profitant de la propriété du clonage d'un agent [46]. RASMA offre la possibilité de créer des agents qui coopèrent ensemble pour diminuer le temps d'exécution de RAS.

Les agents constituant le système RASMA sont les suivants:

- Un agent coordinateur ;
- Un agent prédicats spatiaux (Agent\_PS) ;
- Un agent prédicats non spatiaux (Agent\_PNS) ;
- Des agents clones (Agent\_CL1...Agent\_CLn) ;
- Un agent extraction prédicats spatiaux (Agent\_EPS) ;
- Un agent extraction prédicats non spatiaux (Agent\_EPNS) ;
- Un agent fusion ;
- Un agent construction table prédicats fréquents (Agent\_CTPF) ;

- Un agent génération règles association spatiales (Agent\_GRAS).

### 5.3.2 Présentation des agents

Cette section présente les rôles de chaque agent ainsi que les résultats retournés par chacun de ces agents à travers un exemple.

Soit la base de données décrivant les accidents suivante qui représente la base de données exemple, elle est composée d'un ensemble de données spatiales et non spatiales :

- Accident (Acc\_ID, cond\_ID, date\_ID, Gravité ...SDO\_GID),
- Ecoles (Ecole\_ID, ..., SDO\_GID),
- Espace-vert (Espace\_vert\_ID, Type, ....., SDO\_GID),
- Condition (Cond\_ID, Luminosité, Etat\_de\_Surface\_Route),
- Route (R\_ID, ..., SDO\_GID)

Ce jeu de données a été fourni par l'équipe BDG du laboratoire Prism il présente les accidents dans la commune urbaine de Lille. Ce jeu est réalisé sous le SIG ArcView de la firme ESRI. Les données sont relatives à une région d'une grande variété aussi bien au niveau de la morphologie urbaine que du risque routier. Nous disposons de données urbaines et suburbaines assez riches en matière de qualité et de contenu :

- Données d'accidents : les données descriptives sont fournies par les services de gendarmerie. Ces données sont enrichies par l'information spatiale qui représente les localisations des accidents en système de coordonnées universelles Lambert II. nous travaillons sur un fichier de 29810 accidents ayant eu lieu entre 1984 et le premier trimestre de 1998.
  - Données de voirie : le domaine routier (communal, départemental, etc.) sur le territoire, est divisé en tronçons élémentaires compris entre deux carrefours et caractérisé par un point de début et un point de fin et donc par une direction et une longueur et identifiés par un numéro.
  - Données sur le tissu urbain (Ecoles, espace\_vert, etc.).
- a) **L'Agent Coordinateur** : son rôle est d'assurer le dialogue entre l'utilisateur et le système en offrant à l'utilisateur une interface graphique. Cette interface a plusieurs rôles :

1) Offrir à l'utilisateur un assistant qui l'aide à spécifier la requête.

- 2) Permettre à l'utilisateur de modifier les configurations du système (la valeur de confiance minimale et la valeur du support minimum).
- 3) L'affichage des règles.
- 4) La décomposition de la requête spécifiée par l'utilisateur en deux sous requêtes. Une sous requête contenant les conditions spatiales, elle sera envoyée à l'Agent\_Prédicat\_Spatial (Agent\_PS). La deuxième sous requête sera envoyée à l'Agent\_ Prédicats\_Non\_Spatiaux (Agent\_PNS) en contenant les conditions non spatiales.
- 5) Fournir à l'Agent\_Constructeur\_Table\_Prédicats\_Fréquents (Agent\_CTPF) le seuil de généralisation pour être capable de généraliser les attributs selon le niveau de hiérarchie, ainsi que les deux autres seuils : le support minimum et la confiance minimale modifiés par l'utilisateur.

**b) L'Agent Prédicats Spatiaux (Agent\_PS) :** son rôle est de recevoir la sous requête spatiale de la part de l'Agent\_Coordonateur. Si cette sous requête contient n conditions spatiales, alors, il divise la sous requête en n sous requêtes dont chacune contient une seule condition spatiale. Ensuite, il se clone en n agents clone dont chacun reçoit une sous requête à exécuter. A la fin, chaque agent clone retourne un tableau (voir Tableau 1). Le résultat est envoyé à l'Agent\_Extraction\_ Prédicats\_Spatiaux (Agent\_EPS) et à l'Agent\_fusion.

Acc_ID	Relation_Spatial	Ecole.libelle
1	Proche_de	Emc A. Comte,
1	Proche_de	Emc A. Daudet,
2	Proche_de	Emc Application Jean Aicard
7	Proche_de	Emc B. Desrousseaux
64	Proche_de	Emc A. Franck
*****	*****	*****

**Tableau 3 : Résultat de l'Agent clone\_PS.**

**c) L'Agent Prédicats Non Spatiaux (Agent\_PNS) :** il exécute la sous requête non spatiale provenant de l'Agent\_Coordonateur et fournit le résultat sous forme d'un tableau (voir Tableau 2). Le résultat est envoyé à l'Agent\_Extraction\_ Prédi-cats\_Non\_Spatiaux (Agent\_EPNS) et à l'Agent\_fusion.

Acc_ID	Gravité	Luminosité	Etat_de_surface
1	Léger	Nuit éclairée	Sec normal
2	Grave	Jour	Sec normal
5	Léger	Jour	Sec normal
7	Grave	Jour	Humide
15	Léger	Nuit éclairée	Verglacée
24	Léger	Jour	Mouillée

**Tableau 4 : Résultat de l'Agent\_PNS.**

d) *L'Agent Extraction Prédicats Spatiaux (Agent\_EPS)* : il génère les prédicats spatiaux et leurs fréquences (voir Tableau 3) après avoir reçu les tableaux envoyés par l'Agent\_PS. Le résultat est envoyé à l'Agent\_Construction\_Table\_Prédicats\_Fréquents (Agent\_CTPF).

K	Prédicats	Fréquence
1	Proche_de (Emc A. Conte)	89
1	Proche_de (Emc A. Franck)	60
1	Proche_de (Emc B. Desrousseaux)	60
1	Proche_de (EV)	65

**Tableau 5 : Résultat de l'Agent\_EPS.**

e) *L'Agent Extraction Prédicats Non Spatiaux (Agent\_EPNS)* : il génère les prédicats non-spatiaux et leurs fréquences après avoir reçu le tableau envoyé par l'Agent\_PNS. Le résultat est envoyé à l'Agent\_Construction\_Table\_Prédicats\_Fréquents (Agent\_CTPF).

k	Prédicats	Fréquence
1	Gravité (sans gravité)	38
1	Gravité (léger)	23634
1	Gravité (grave)	5313
1	Gravité (mortel)	825
1	Luminosité (jour)	20444
1	Luminosité (demi-jour)	1483
1	Luminosité (Nuit éclairée)	6999
1	Luminosité (Nuit éclairée insuffisant)	188
1	Luminosité (Nuit sans éclairage)	696
1	Etat_de_surface (Humide)	4159
1	Etat_de_surface (Mouillée)	2699
1	Etat_de_surface (Enneigée)	99
1	Etat_de_surface (Verglacée)	221
1	Etat_de_surface (Gras boueux)	51
1	Etat_de_surface (Gravillons)	34
1	Etat_de_surface (Sec normal)	22547

**Tableau 6 : Résultat de l'Agent\_EPNS.**

## Chapitre II : Étude Des Travaux Exploitant Les SMAs Pour Le Data Mining

f) *L'Agent fusion* : il reçoit les tableaux 1 et 2 envoyés par les deux agents: Agent\_PS et Agent\_PNS. Il vérifie les identificateurs de chaque table et produit par la suite un tableau comme illustré par le tableau 5. Ensuite il génère les séquences des prédicats et leurs fréquences et produit le deuxième tableau (voir Tableau 6). Le résultat est envoyé à l'Agent\_Construction\_Table\_Prédicats\_Fréquents (Agent\_CTPF).

A	L	G	EDS	Ecole	EV
1	E-N	Léger	Sec Normal	Emc A. Comte, Emc A. Daudet,	EV
7	Jour	grave	Humide	Emc B. Desrousseaux	EV
11	Jour	Léger	Sec normal	Emc A. Franck, Emc Application Jean Aicard ,	EV
18	Jour	Léger	Mouillée	Emc A. Comte, Emc A. Daudet,	Null
....	...	....	.....	.....	.....

Où A: Acc\_ID, L: Luminosité, G : Gravité, EDS : Etat\_de\_surface, EV : Espace Vert; N-E: NUIT ECLAIRÉE

**Tableau 7 : Résultat intermédiaire de la fusion.**

K	Prédicats	Fréquence
2	Proche_de (Emc B. Desrousseaux) et Proche_de (EV)	36
2	Proche_de (Emc A. Franck) et Luminosité (demi-jour)	1
3	Proche_de (EMC A. FRANCK) et Gravité(léger) et Luminosité(Jour)	26
4	Proche_de (Emc A. Comte) et Proche_de (EV) et Gravité ( Léger) et Luminosité(Jour)	32
5	Proche_de (Emc A. Comte) et Proche_de (EV) et Gravité (Léger) et Luminosité(Jour) et Etat_de_surface (sec normal)	32
5	Proche_de (Emc Application Jean Aicard) et Gravité (Léger) et Luminosité(Jour) et Etat_de_surface (sec normal) et Proche_de (EV)	20

**Tableau 8 : Résultat de l'agent fusion.**

g) *L'Agent Construction Table Prédicats Fréquents (Agent\_CTPF)* : il reçoit les trois tableaux 3, 4 et 6, il génère les prédicats fréquents selon le support minimum dans un tableau (voir Tableau 7) en fusionnant les tableaux reçus. Ensuite, il envoie le tableau 7 à l'Agent\_Génération\_des\_Règles\_d'Association\_Spatiales (Agent\_GRAS). Ce processus se répète pour chaque niveau de hiérarchie en fonction des seuils fournis par l'utilisateur.

K	Prédicats	Fréquence
1	Proche_de (Emc A. Conte)	89
1	Proche_de (Emc A. Daudet)	99
1	Proche_de (EV)	65
1	Gravité (Léger)	69
1	Luminosité (Jour)	67
1	Etat_de_surface (SEC NORMAL)	81
2	Proche_de (Emc A. Conte) et Gravité (Léger)	69
3	Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger)	42
4	Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger) et Luminosité (Jour)	32
5	Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger) et Luminosité (Jour) et Etat_de_surface (SEC NORMAL)	32

**Tableau 9 : Résultat de l'Agent\_CTPF.**

*h) L'Agent Génération Règles Association Spatiales (Agent\_GRAS) :* après avoir reçu le tableau des prédicats fréquents, il génère les règles (voir Tableau 8) avec les deux mesures, le support et la confiance. Il est nécessaire de noter que l'utilisateur pour chaque niveau de généralisation introduit ces deux mesures. En fait, à chaque fois que l'Agent\_GRAS génère les règles pour un niveau donné de la hiérarchie, il demande à l'Agent\_CTPF de lui envoyer le tableau des prédicats fréquents pour le niveau suivant.

Prémisse	Conclusion	S%	C%
Accident (X) et Proche_de (Emc A. Conte)	Gravité (Léger)	69	70.4
Accident (X) et Proche_de (EV)	Etat_de_surface (Humide)	9	14
Accident (X) et Proche_de (Emc A. Conte) et Proche_de (EV) et Gravité (Léger)	Luminosité (Jour)	32	76

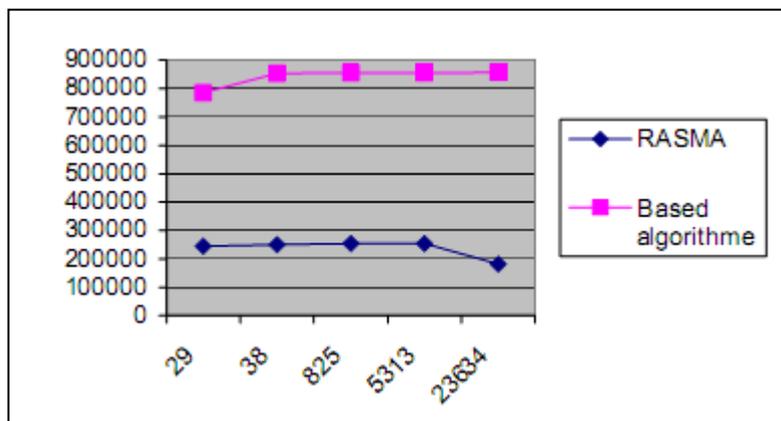
**Tableau 10 : Règles d'association spatiales.**

### 5.3.3 Expérimentation, résultats et performances

Les algorithmes 1 et 2 correspondent respectivement à l'approche multi-agent RASMA et l'approche classique RAS. Le tableau 9 résume les coûts d'exécution en milliseconde de chacun des deux algorithmes. Les tests visent à comparer les performances de chacun des deux approches. La figure 17 montre le temps d'exécution de deux algorithmes en fonction de la taille de la table cible. Nous pouvons remarquer l'énorme différence entre le temps d'exécution des deux algorithmes.

Nombre des objets cibles	Nombre des objets en relation	RASMA		RAS (Koperski et al.(1995))	
		Temps d'exécution (ms)	Temps d'exécution (ms)	Temps d'exécution (ms)	Temps d'exécution (ms)
29	3490	244656	785172		
38	58	250172	854188		
825	1084	254485	854687		
5313	7342	254250	855531		
23634	31765	182235	856891		

**Tableau 11 : Les temps d'exécution.**



**Figure 17 : temps d'exécution en fonction de la taille cible.**

### 5.3.4 Evaluation

Les résultats de l'expérimentation montrent une nette amélioration du temps d'exécution de RASMA par rapport à RAS. Le gain en temps d'exécution est dû à la distribution des tâches entre les agents et au parallélisme.

L'approche permet à l'utilisateur de configurer le système en introduisant la requête et les différents seuils d'une manière interactive en offrant une interface graphique. Concernant les règles extraites, RASMA permet de générer des règles qui englobent en même temps des prédicats spatiaux et des prédicats non spatiaux qui n'est pas le cas pour RAS. RASMA filtre les prédicats une seule fois par l'Agent-CTPF et génère un tableau des k\_prédicats fréquents au lieu de filtrer à deux reprises, une fois pour générer les 1\_prédicats fréquents et une deuxième fois pour générer les k\_prédicats fréquents, pour l'algorithme RAS [47].

### **6 Conclusion**

La combinaison des systèmes multi-agents avec des diverses applications dans de divers domaines a été un résultat naturel grâce aux avantages qu'ils donnent en permettant de résoudre des problèmes complexes en se basant sur les principes de base : la collaboration entre les agents et le parallélisme. Notre projet est un cas pareil, il s'intéresse à la contribution des SMAs dans les systèmes de fouille de données ce qui a été la cause de consacrer ce chapitre à déterminer les différentes possibilités de cette contribution.

Ce chapitre est présenté en deux parties, la première en étant des généralités sur le paradigme de systèmes multi-agents comportant une branche de l'intelligence artificielle distribuée. Ce système qui est un ensemble hétérogène d'applications encapsulé (agents) qui participent au processus de prise de décision, communiquent, collaborent et négocient afin de répondre chacun à ses propres objectifs de conception mais aussi à un objectif qui est partagé au sein de la communauté.

La deuxième partie du chapitre contient l'apport des systèmes multi-agents au data mining, et dans laquelle nous avons présenté trois approches visant à intégrer l'approche multi-agents dans des systèmes de data mining.

Le chapitre suivant va contenir les détails sur le modèle projeté.

## **Chapitre III**

---

*Modélisation.*

### 1 Introduction

Avec l'extraordinaire explosion de la quantité de données accumulées par les organisations d'aujourd'hui, il est extrêmement important que les techniques de data mining soient en mesure de traiter ces données de manière efficace ce qui rend la tâche plus complexe. Pour cela une plate forme multi agents est très efficace dans la mesure où la complexité sera distribuée sur un ensemble d'entités communicantes, autonomes, réactives et dotées de compétences appelées agents.

Pour notre projet nous avons choisi de modéliser par un système multi agents un système de data mining pour la tâche de clustering (ou la segmentation) dont le but est de regrouper des enregistrements ou des observations en classes d'objets similaires en utilisant la technique de k-moyennes (k-means. Voir chapitre 2). Les agents choisis pour cette approche sont des agents de type cognitif.

### 2 Présentation générale du modèle

Après avoir préparé et transformer les données, l'étape qui suit selon le processus de data mining (voir chapitre 1) est l'étape d'extraction de connaissances et dans notre cas le clustering qui représente l'objet de notre projet.

En se basant sur l'approche multi-agents dont le principe est de distribuer l'expertise sur un ensemble d'agents cognitifs modélisant chacun une tâche du système, nous proposons l'architecture de notre système illustrée par la figure 18.

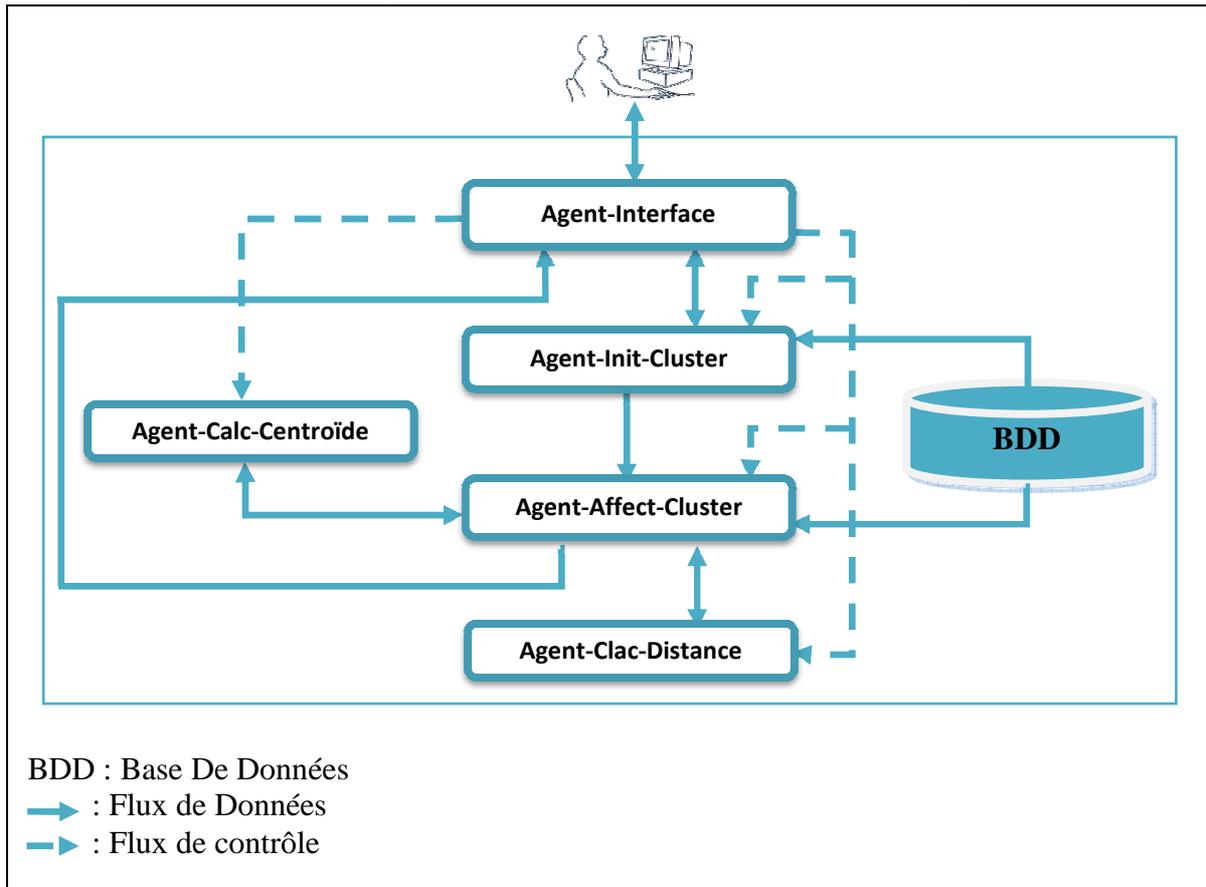


Figure 18 : Architecture générale du système

Notre système se compose de :

- Cinq agents cognitifs qui sont
  - **Agent Interface** : responsable de fournir le nombre de clusters  $k$  nécessaire pour la méthode  $k$ -means utilisée par le système.
  - **Agent Init-Cluster** : l'agent responsable de définir les classes initiales.
  - **Agent Calc-Centroïde** : calcul le centroïde de chaque cluster.
  - **Agent Affect-Cluster** : son rôle est de générer une nouvelle partition en assignant chaque enregistrement au groupe dont le centre de gravité est le plus proche en s'appuyant sur la valeur de la distance fournie par l'agent Calc-Distance.
  - **Agent Calc-Distance** : Calcul la distance en s'appuyant sur la distance euclidienne entre le centre du cluster (le centroïde données par l'agent Calc-Centroïde) et tous les enregistrements donnés.
- La base de données contenant les données préparées et transformées auparavant.

Le flux de contrôle indiqué dans le schéma assure la coopération entre les agents du système et garantit la sécurité de ce dernier.

### 3 Description des composants de l'architecture

#### 3.1 Agent interface

L'agent responsable de la communication avec l'utilisateur, il est considéré comme la fenêtre du système vers l'extérieur. C'est l'élément décisif du succès ou de l'échec du système dans la mesure où il fournit le nombre de clusters  $k$  nécessaire au système pour commencer la segmentation.

##### 3.1.1 Architecture de l'agent interface

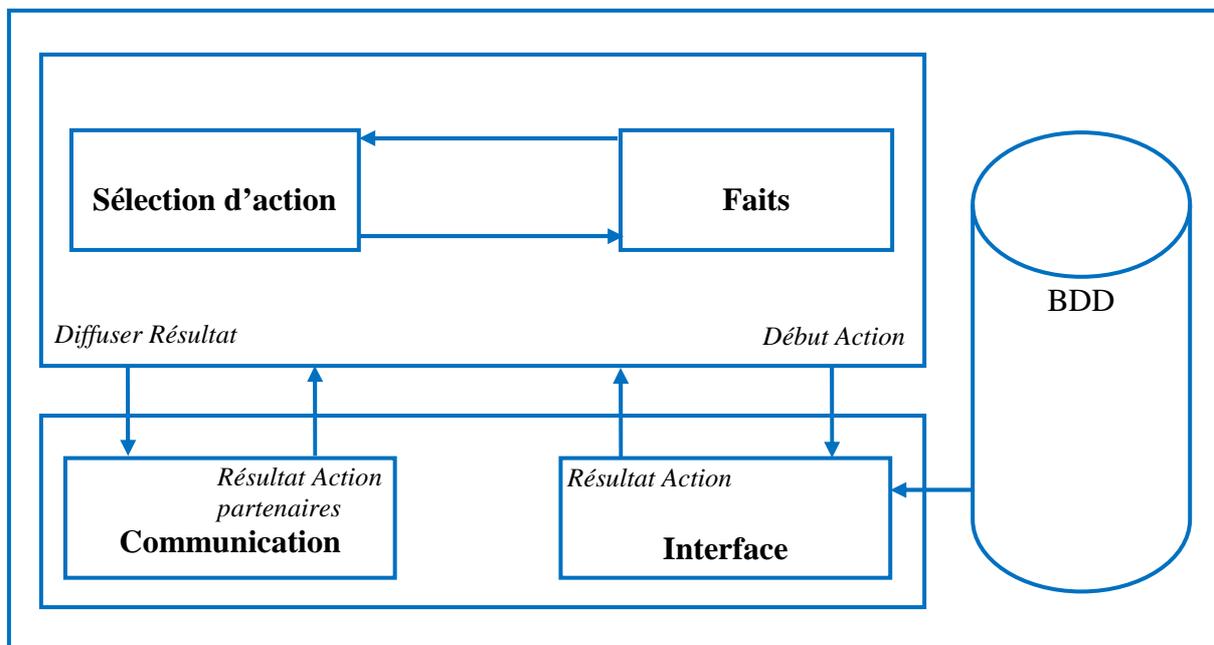


Figure 19 : Architecture de l'agent interface.

Les agents de notre système sont divisés en deux parties, une partie décisionnelle et une partie opérative. La partie décisionnelle est fondée sur un module de sélection d'action en utilisant la base de faits. Pour obtenir des informations ou agir sur l'environnement, ce module communique par messages avec la partie opérative de l'agent. La partie opérative dans le cas de l'agent interface est représentée par les modules suivants qui s'exécutent en parallèle :

- **Un module interface** : responsable de la liaison de l'utilisateur avec le système en présentant les fonctionnalités du système sous forme d'une interface graphique, il

collecte les données entrées, nécessaire au déclenchement du système, par l'utilisateur autorisé par test de droit d'accès. Ce module renseigne le module décisionnel sur le résultat de l'action (échec ou succès).

- **Un module de communication** : responsable de l'envoi des messages contenant les résultats (le nombre de clusters saisi par l'utilisateur et nécessaire pour le fonctionnement de l'agent Init-Cluster) et ceux émis lors du démarrage et de l'arrêt des actions, il reçoit les messages des autres agents (cluster terminaux envoyé par l'agent Affect-Cluster) et alimente la base de connaissances.

### 3.1.2 Fonctionnement de l'agent interface

Le fonctionnement de l'agent interface est le suivant :

a) L'authentification :

- Saisie du nom d'utilisateur et du mot de passe.
- Interrogation de la table utilisateurs qui contient les informations des comptes utilisateur.
- Comparaison des données saisies avec ceux de la base de données.

b) Saisie du nombre de clusters.

c) Envoi du nombre saisi aux agents Init-Cluster et Affect-Cluster.

### 3.1.3 Le savoir de l'agent interface

Les agents de notre système sont des agents cognitifs dont l'architecture est inspirée des systèmes experts, pour cela ils ont besoin de savoir pour pouvoir réaliser leurs tâches, le savoir de l'agent interface peut être décrit comme suit :

- **Si** l'utilisateur est identifié **Alors** lui permettre l'accès aux fonctions du système.
- **Si** l'utilisateur n'est pas identifié **Alors** afficher un message d'erreur
- **Si** l'utilisateur saisie le nombre k **Alors** lancer la requête d'activation des agents Init-Cluster et Affect-Cluster
- **Si** l'agent Affect-Cluster envoie message de fin du clustering **Alors** afficher résultat

### 3.2 Agent Init-Cluster

L'agent d'initialisation des clusters déclenché par l'agent interface, il se charge de déterminer les k premier enregistrements de la base (k le nombre de cluster saisi par l'utilisateur et envoyé par l'agent interface) commettant les clusters initiaux de notre système, et envoi leurs coordonnées (valeur des attributs) à l'agent Affect-Cluster.

#### 3.2.1 Architecture de l'agent Init-Cluster

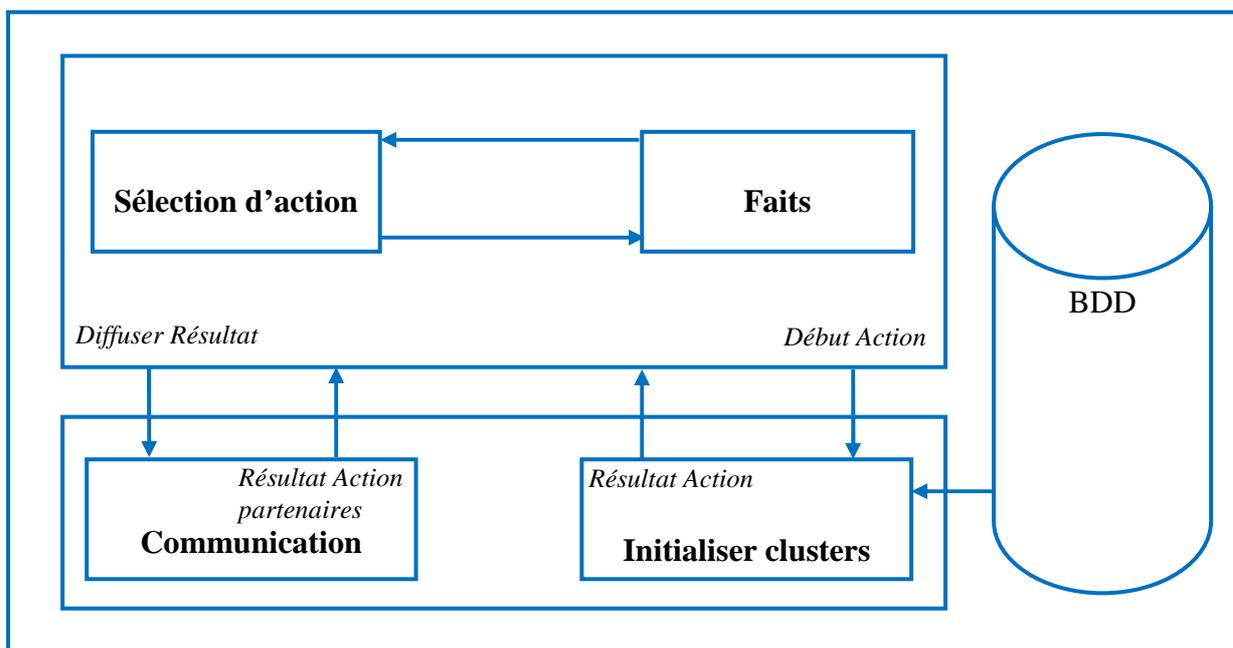


Figure 20 : Architecture de l'agent Init-Cluster.

Le fonctionnement de l'agent Init-Cluster se répartit sur deux module (voir figure 20), qui sont :

- **Un module de communication** : garanti la liaison de l'agent avec les autres agents du système, en envoyant les résultats de ses actions et en recevant les résultats des autres agents qui est dans notre cas l'envoi des enregistrements sélectionnés comme cluster initiaux à l'agent Affect-Cluster après avoir reçu le nombre de clusters k envoyé par l'agent interface, il garanti ainsi l'alimentation la base de connaissances.
- **Un module d'initialisation des clusters** : exécute l'action sélectionnée par la base de connaissance qui est le test de du nombre des clusters demandé par l'utilisateur par rapport au nombre d'objets (enregistrements) existants dans la base de données et la sélection des k premiers enregistrements pour former les clusters initiaux.

### 3.2.2 Fonctionnement de l'agent Init-Cluster

Pour sélectionner les clusters initiaux (les enregistrements qui forme chacune un segment), l'agent Init-Cluster suit les étapes suivantes :

- a) Recevoir le message de l'agent Interface contenant le nombre  $k$  de clusters.
- b) Test du nombre reçu par rapport au nombre d'enregistrements contenus dans la base de données.
- c) Accès à la base de données et sélection des  $k$  premiers enregistrements.
- d) Extraire les valeurs des attributs de ces enregistrements et les mettre sur des vecteurs.
- e) Envoi des vecteurs à l'agent Affect-Cluster en déclenchant ainsi ce dernier.

### 3.2.3 Le savoir de l'agent Init-Cluster

Le savoir qu'utilise l'agent Init-Cluster pour son raisonnement est le suivant :

- *Si* la valeur du nombre  $k$  n'est pas encore définie *Alors* se mettre en attente pour un message contenant cette dernière.
- *Si* le nombre  $k$  est défini *Alors* accéder à la base de données et extraire les  $k$  premier enregistrements.
- *Si* le nombre de clusters saisie par l'utilisateur est supérieur ou égal au nombre d'enregistrements existants dans la base de données *Alors* message d'erreur.
- *Si* les  $k$  enregistrements sont sélectionnés *Alors* envoyer un message à l'agent Affect-Cluster contenant les valeurs des attributs de ces derniers.

### 3.3 Agent Affect-Cluster

L'agent Affect-Cluster a pour rôle d'affecter les données extraites de la base de données au cluster adéquat en extrayant les données préparées précédemment et stockées dans la base de données, en acquérant les coordonnées des classes initiales envoyées par l'agent

Init-Cluster et en envoyant le tout à l'agent Clac-Distance en vue de calculer leurs distances. Ensuite, et après avoir reçu les distances calculées et envoyées par l'agent Clac-Distance, l'agent Affect-Cluster affecte chaque enregistrement à la classe la plus proche. Une fois l'affectation finie, un message est envoyé à l'agent Calc-Centroïde pour recalculer les centroïdes des classes nouvellement changées par le cycle d'affectation précédemment décrit.

### 3.3.1 Architecture de l'agent Affect-Cluster

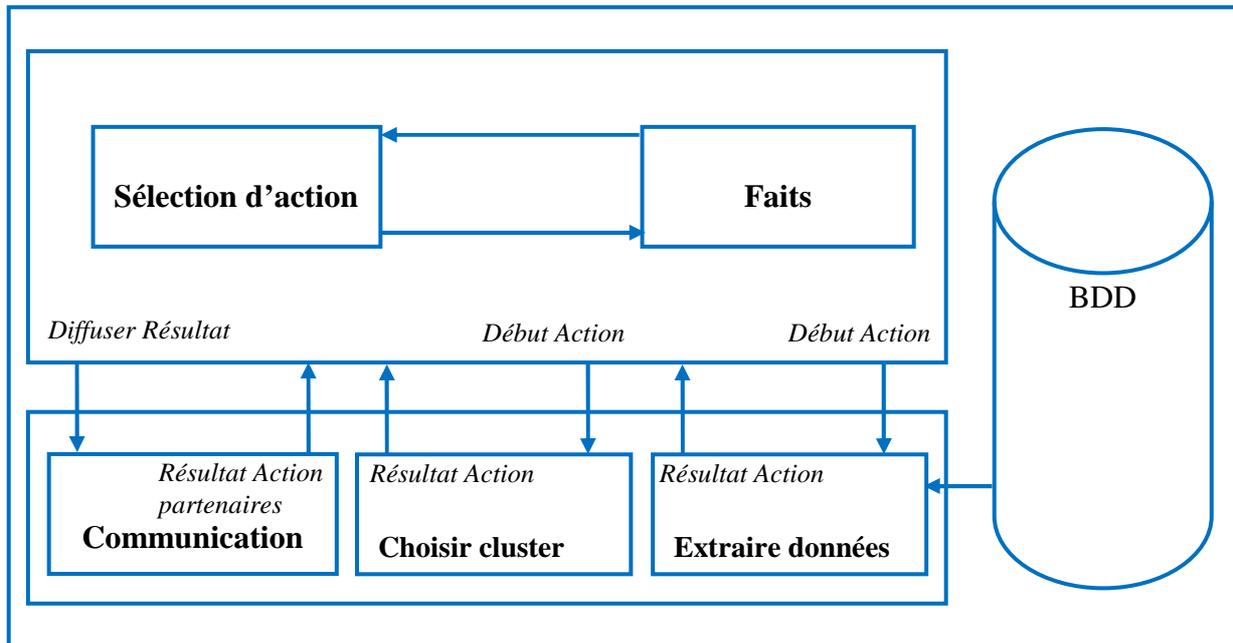


Figure 21 : Architecture de l'agent Affect-Cluster.

Les modules de l'agent Affect-Cluster qui lui permettent de fonctionner et d'accomplir ces tâches sont les suivants :

- **Un module de communication** : fournit une interface avec les autres agents, son rôle est de recevoir le nombre de clusters  $k$  envoyé par l'agent interface et les coordonnées (valeurs des attributs) des enregistrements choisies par l'agent Init-Cluster comme clusters initiaux et d'envoyer des messages aux agents Clac-Distance et Calc-Centroïde contenant les coordonnées des clusters initiaux et les données extraites de la base de données pour l'agent Clac-Distance et pour l'agent Calc-Centroïde les nouveaux clusters, et de recevoir des réponses à ces messages contenant les distances calculées par l'agent Clac-Distance et les centroïdes calculés par l'agent Calc-Centroïde.
- **Un module de choix de cluster** : il utilise les informations qui ont été envoyées par l'agent Clac-Distance et en fonction de la distance il choisit pour chaque

enregistrement le cluster qui lui est le plus proche (distance minimale). le choix des clusters s'arrête et par conséquent toute l'opération de clustering, une fois les distances récupérés de l'agent Clac-Distance ne changent pas ou le nombre d'itérations a été achevé.

- **Un module d'extraction de données** : son rôle consiste à l'extraction des données collectées, préparées, nettoyées et transformées stockées dans la base de données en vue d'être segmentées, cette action passe en parallèle avec la tâche de la définition des groupes initiaux et celle du calcul des distances dans la mesure où chaque fois qu'un nombre fixe d'enregistrements est extrait il sera envoyé à l'agent Calc-Distance et en même temps que le calcul de distance est activé, une autre quantité d'enregistrements sera extraite est envoyée, afin de réduire le temps nécessaire à la segmentation.

### 3.3.2 Fonctionnement de l'agent Affect-Cluster

Dans le but d'accomplir son rôle, l'agent Affect-Cluster suit le fonctionnement suivant :

- a) Recevoir le nombre de clusters  $k$  envoyé par l'agent interface.
- b) Extraire de la base de données les données à segmenter.
- c) Recevoir le message de l'agent Init-Cluster contenant les coordonnées des clusters initiaux.
- d) Envoyer les données extraites précédemment et les clusters initiaux à l'agent Calc-Distance en déclenchant ce dernier.
- e) Recevoir les distances calculés et envoyés par l'agent Calc-Distance.
- f) Affecter chaque enregistrement au cluster le plus proche (dont la distance est minimale).
- g) Envoyer les nouveaux clusters à l'agent Calc-Centroïde afin de calculer leurs nouveaux centres.
- h) Recevoir le message de l'agent Calc-Centroïde contenant les nouveaux centroïdes.

- i) Envoyer à l'agent Calc-Distance une demande de calcul de distance des données par rapport aux nouveaux centres.
- j) Revenir à l'étape 'd' et répéter toutes les étapes qui la suivent jusqu'à ce que les distances récupérés de l'agent Clac-Distance ne changent pas ou le nombre d'itérations est achevé.

### 3.3.3 Le savoir de l'agent Affect-Cluster

Le savoir de l'agent Affect-Cluster est composé des points suivants :

- *Si* le nombre de clusters  $k$  n'a pas encore été défini *Alors* se mettre en attente pour un message contenant ce dernier.
- *Si* la valeur du nombre  $k$  est définie *Alors* extraire les données stockées dans la base de données (travaille en parallèle avec l'agent Init-Cluster) et attendre les valeurs des segments initiaux.
- *Si* les coordonnées des groupes initiaux ont été définies *Alors* envoyer un message à l'agent Calc-Distance contenant les données extraites et les centroïdes des clusters initiaux.
- *Si* les distances entre les enregistrements et les centres des clusters sont acquises *Alors* affecter chaque enregistrement au cluster le plus proche.
- *Si* les nouveaux clusters ont été définis *Alors* les envoyés à l'agent Calc-Centroïde pour calculer leurs nouveaux centres.
- *Si* les nouveaux centres sont reçus *Alors* faire test d'arrêt.
- *Si* le résultat du test d'arrêt est faux *Alors* envoyer les nouveaux centroïdes et les données à l'agent Calc-Distance.
- *Si* le résultat du test d'arrêt est vrai *Alors* envoyer le résultat de segmentation à l'agent interface.

### 3.4 Agent Calc-Centroïde

L'agent de calcul des centroïdes, déclenché par l'agent Affect-Cluster dans le but de calculer les centres des groupes de données.

#### 3.4.1 Architecture de l'agent Calc-Centroïde

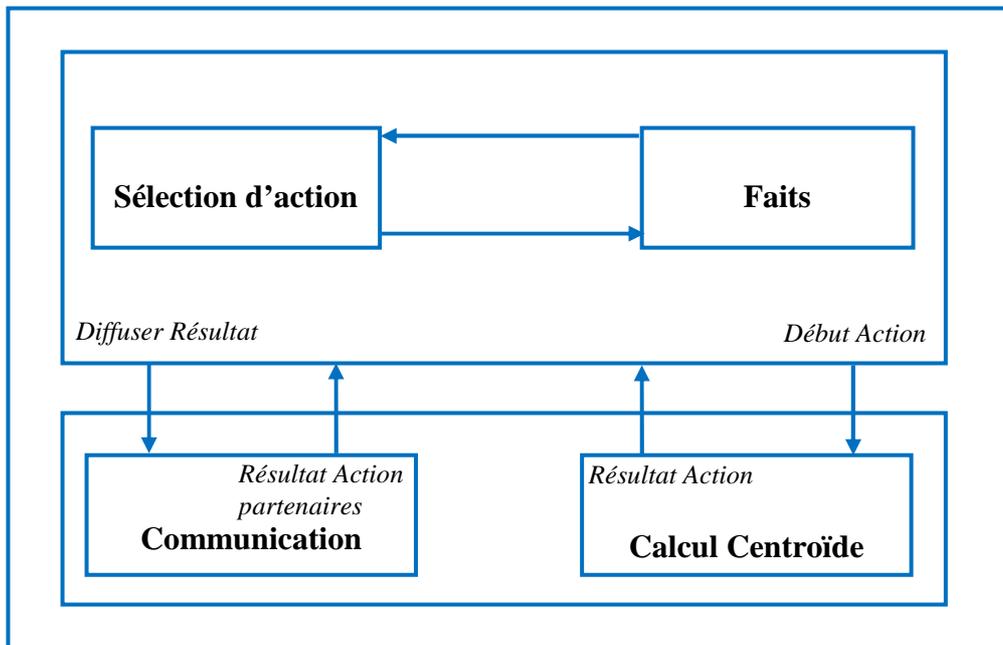


Figure 22 : Architecture de l'agent Calc-Centroïde.

Les modules fonctionnels propres à l'agent Calc-Centroïde sont les suivants :

- **Un module de communication** : c'est le module responsable de la réception et l'envoi des messages venants de et allant à l'agent Affect-Cluster, les messages entrants contiennent les segments récemment définis tandis que les messages sortants comportent les centroïdes calculés des segments reçus.
- **Un module de Calcul des Centroïdes** : responsable du calcul des centres des groupes envoyés par l'agent Affect-Cluster et reçus par le module de communication. Pour calculer les centroïdes nous utilisons la moyenne pondérée  $\pi_{kj}$  tel que :

$$\pi_{kj} = \frac{1}{|C_k|} \sum_{o_i \in C_k} P_{ij}$$

Où :

$\pi_{kj}$  : représente la valeur de la propriété  $j$  du centroïde du cluster  $k$

$C_k$  : représente les  $k$  classes ;

$O_i$  : représente les objets (enregistrements) à classer ;

$P_{ij}$  : représente la valeur de la propriété (attribut)  $P_j$  pour l'objet  $O_i$  ;

### 3.4.2 Fonctionnement de l'agent Calc-Centroïde

Pour calculer les centroïdes, l'agent Calc-Centroïde fonctionne comme suit :

- a) Recevoir le message de l'agent Affect-Cluster contenant les coordonnées des clusters récemment définis.
- b) Calculer pour chaque cluster son centroïde en utilisant la moyenne pondérée.
- c) Envoyer les centroïdes calculés à l'agent Affect-Cluster.

### 3.4.3 Le savoir de l'agent Calc-Centroïde

Le savoir de l'agent Calc-Centroïde est comme suit :

- **Si** le message contenant les coordonnées des clusters récemment définis n'est pas encore reçu **Alors** se mettre en attente pour un message contenant ces derniers.
- **Si** le message contenant les valeurs des propriétés des clusters récemment définis est reçu **Alors** calculer le centroïde de chaque clusters.
- **Si** les centroïdes ont été calculés **Alors** envoyer un message à l'agent Affect-Cluster contenant les valeurs des centroïdes des clusters envoyés par ce dernier.

### 3.5 Agent Calc-Distance

Pareil que l'agent Calc-Centroïde, l'agent Calc-Distance ne travaille qu'avec l'agent Affect-Cluster, il est déclenché par ce dernier dans le but de calculer les distances entre les centres des clusters calculés par l'agent Calc-Centroïde et les données à segmentées.

#### 3.5.1 Architecture de l'agent Calc-Distance

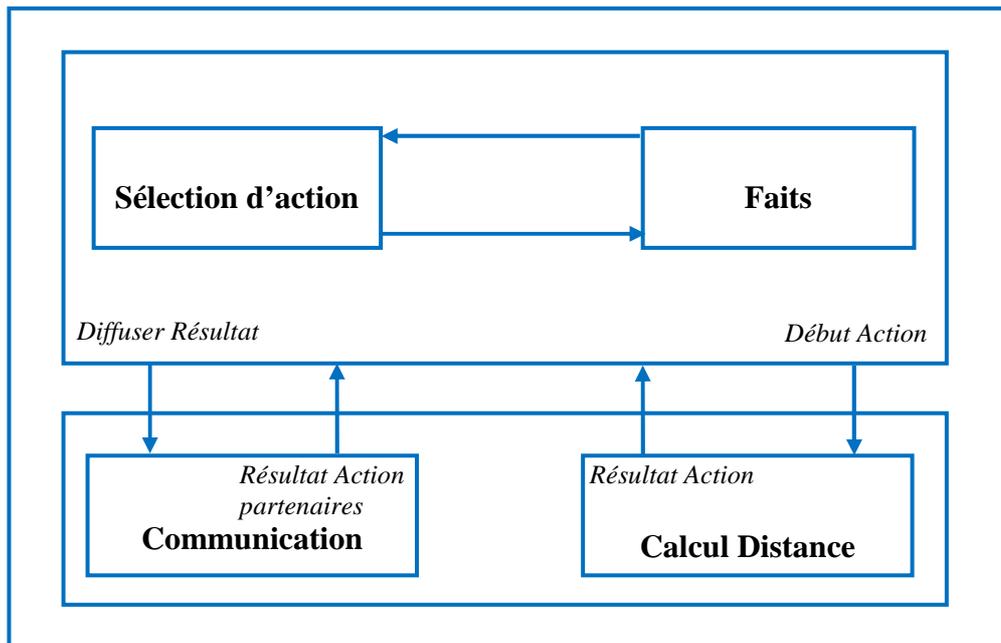


Figure 23 : Architecture de l'agent Calc-Distance.

Le fonctionnement de l'agent Calc-Distance s'effectue au niveau des deux modules suivants :

- **Un module de communication :** Comme tous les modules de communication des agents précédemment présentés, le module de communication de l'agent Calc-Distance permet l'échange de message entre l'agent Calc-Distance et son environnement, il reçoit les messages envoyés par l'agent Affect-Cluster et nécessaire au fonctionnement de l'agent à qui il appartient (l'agent Calc-Distance) et envoie les résultats des calculs réalisés par ce dernier.
- **Un module de calcul de distance :** il se charge de calculer les distances entre chaque enregistrement et les centroïdes de tous les clusters. La distance calculée est une distance euclidienne, tel que :

$$d(O_i, \pi_k) = \sqrt{\sum_{j=1}^C (P_{ij} - \pi_{kj})^2}$$

Où :

$O_i$  : représente l'objet (enregistrement)  $i$  à classer ;

$\pi_k$  : représente le centroïde du cluster  $k$  ;

$P_{ij}$  : représente la valeur de l'attribut  $j$  de l'enregistrement  $i$  ;

$\pi_{kj}$  : représente la valeur de l'attribut  $j$  du centroïde du cluster  $k$  ;

$C$  : représente le nombre d'attribut constant pour tous les enregistrements.

### 3.5.2 Fonctionnement de l'agent Calc-Distance

Les étapes que suit l'agent Calc-Distance pour son fonctionnement sont les suivantes :

- a) Recevoir le message de l'agent Affect-Cluster contenant les coordonnées des centres des clusters et celles des données à classer.
- b) Calculer en utilisant la distance euclidienne, la distance de chaque enregistrement par rapport à tous les centres des clusters.
- c) Envoyer le résultat du calcul à l'agent Affect-Cluster.

### 3.5.3 Le savoir de l'agent Calc-Distance

L'agent Calc-Distance comporte le savoir suivant :

- *Si* le message contenant les coordonnées des centroïdes des clusters et les coordonnées des enregistrements n'est pas encore reçu *Alors* se mettre en attente pour un message contenant ces derniers.
- *Si* le message est reçu *Alors* calculer la distance entre chaque enregistrement et tous les centres des clusters.
- *Si* le calcul a été achevé *Alors* envoyer les distances résultantes à l'agent Affect-Cluster.

### 4 La communication inter-agents

La communication inter-agents permet aux agents de bénéficier des informations et du savoir-faire des autres agents et par conséquent augmente les capacités perceptives des agents, elle constitue l'un des moyens principaux assurant la répartition des tâches et la coordination des actions [38].

Dans les systèmes cognitifs (ce qui est le cas de notre système), les communications s'effectuent par envois de messages. Le langage de communication que nous avons choisi pour l'envoi de messages entre nos différents agents est le langage ACL (Agent Communication Language) dont la spécification permet une interopérabilité maximale entre les agents. L'ACL a été spécifié par la FIPA (Foundation for Intelligent Physical Agents) en 1996 [48].

La syntaxe d'un message FIPA-ACL est illustrée comme suit :

```
(request
  :sender Agent_A
  :receiver Agent_B
  :content
  (...)
  :in-reply-to action
  :replay-with reponse
  :language FIPA-SL0
)
```

Figure 24 : Structure d'un message FIPA-ACL.

## 5 Fonctionnement du système

Cette partie consiste à présenter notre application en se basant sur le langage AUML.

### 5.1 Diagramme de classes

La figure 25 consiste en la représentation du diagramme des classes existantes dans notre système et les relations qui les lient.

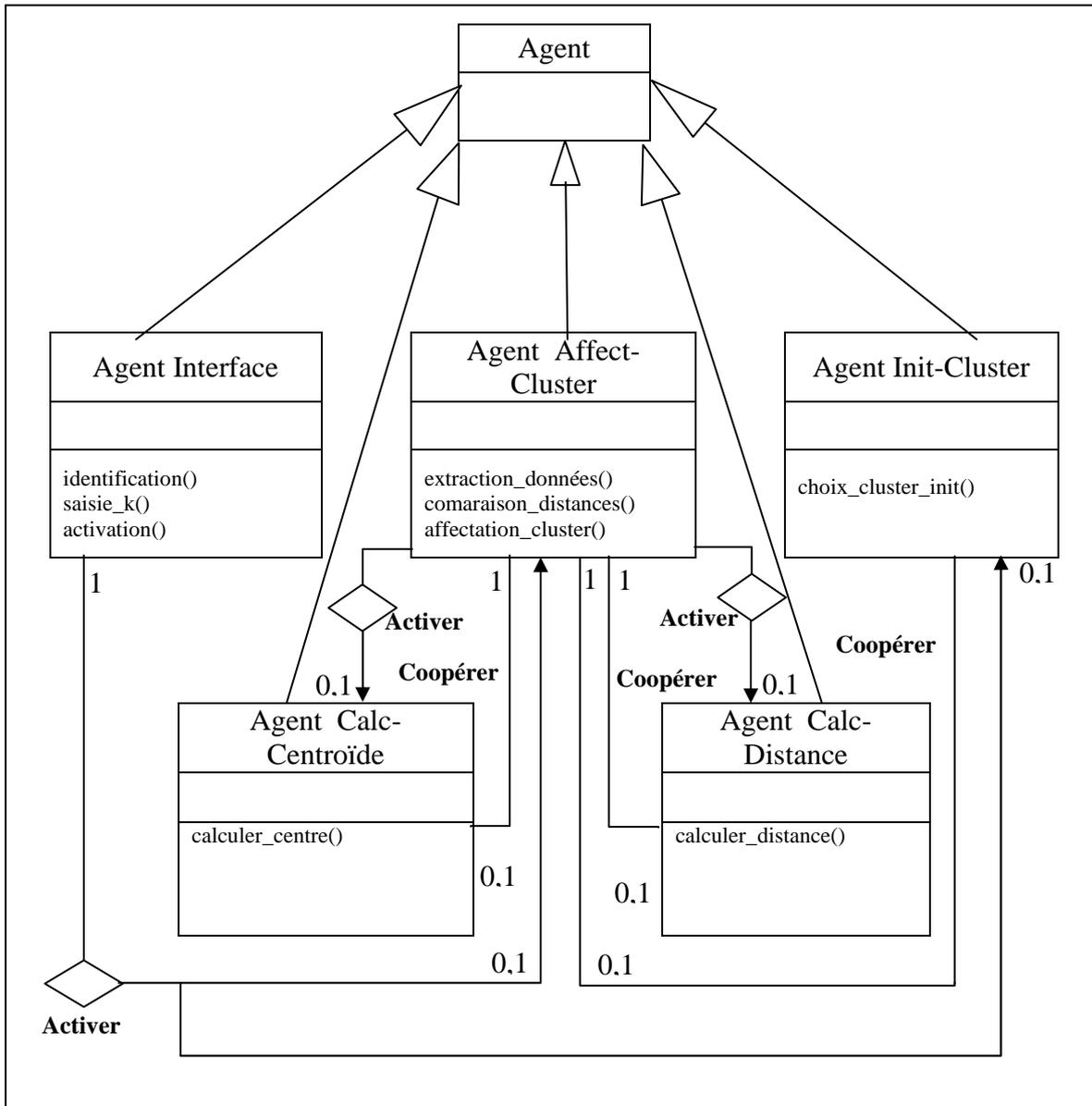


Figure 25 : diagramme de classe AUML du modèle proposé.

### 5.2 Diagramme de séquence

Le diagramme de séquence suivant décrit le processus de fonctionnement général de notre système.

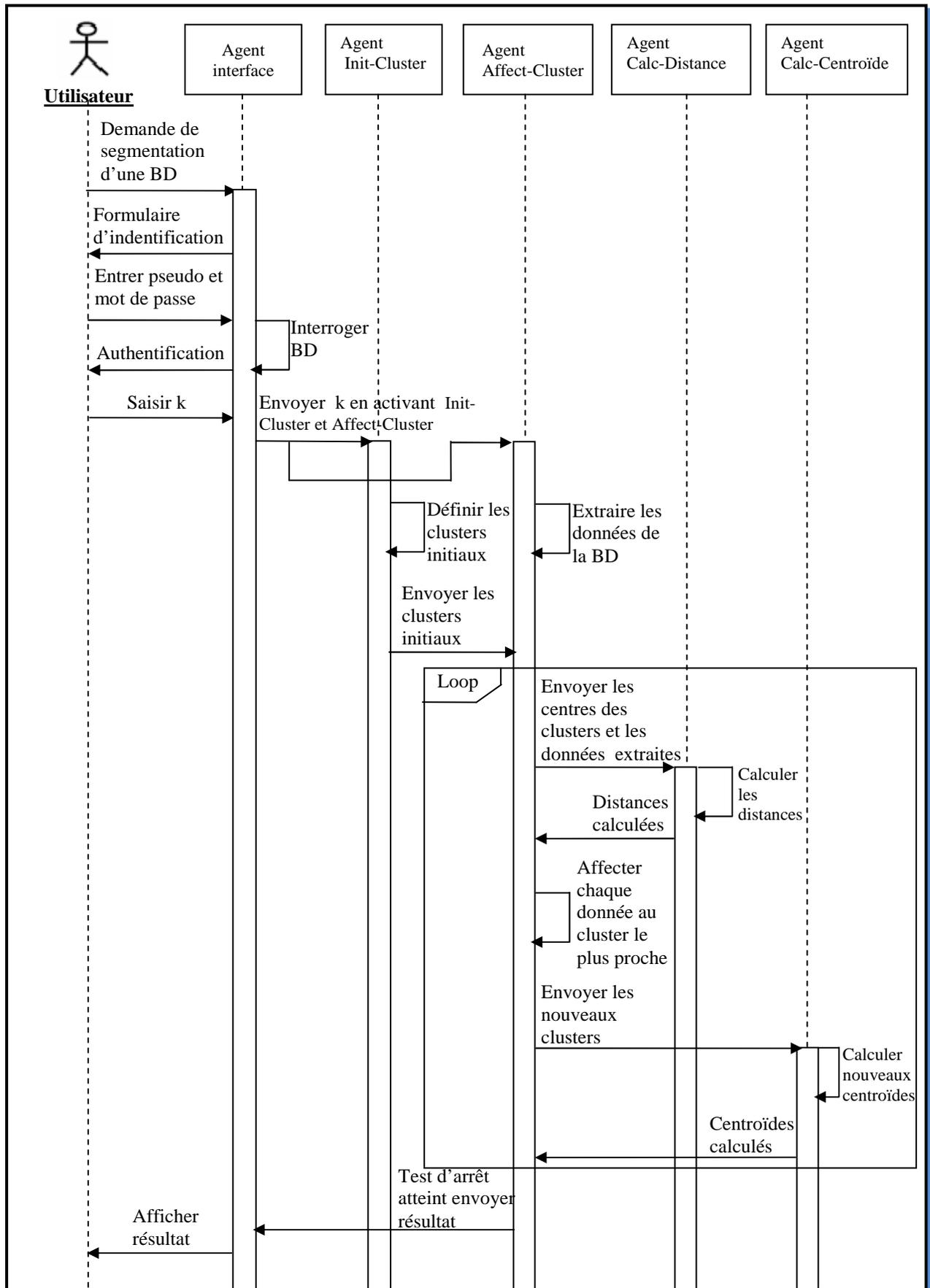


Figure 26 : diagramme de séquence AUML.

### 6 Conclusion

Nous avons présenté dans ce chapitre notre architecture d'un système de data mining dont le tâche est la segmentation d'un grand ensemble de données en utilisant la méthode de k\_means et en se basant sur les systèmes multi-agents.

Nous avons utilisée une architecture à base d'agent coopérants en bénéficiant de leur autonomie, modularité, distribution et intelligence pour minimiser le temps immense de calcul nécessaire à la segmentation d'un grand nombre de données.

Le chapitre suivant, présente les résultats de la phase d'implémentation de notre système.

## **Chapitre IV**

---

***ETUDE DE CAS.***

### 1 Introduction

Dans ce chapitre nous allons montrer la validité et fiabilité du modèle proposé précédemment est dédié à la segmentation des données dans une plate forme multi agents. Afin d'aboutir à ce but, nous allons faire une étude de cas, d'où nous allons appliquer notre modèle sur un entrepôt de données qui doit passer par un processus de préparation de données.

### 2 Les étapes du processus adopté

Nous avons vu dans le chapitre 1 que le processus de fouille de données passe par plusieurs étapes nécessaires à l'extraction de connaissance. Le processus adopté par notre système est composé des étapes majeures suivantes:

#### 2.1 Préparation des données

La préparation des données (collecte des données) consiste à obtenir des données en accord avec les objectifs visés. Ces données proviennent généralement des bases de production ou d'entrepôts. Les données sont structurées en champs typés (dans un domaine de définition). L'obtention des données est souvent réalisée à l'aide d'outils d'extraction de données par requête (OLAP, SQL, etc.)

Les données collectées sont tout d'abord copiées sur une machine adéquate, pour des raisons de performance, mais surtout parce qu'elles seront modifiées [49].

Pour notre projet nous avons choisi de faire notre étude de cas sur la base de données du recensement de l'année 1990 des états unies d'Amérique. Les données collectées représentent un échantillon d'un pour cent des échantillons du PUMS (Public Use Microdata Samples) tiré du département de commerce des états unies (U.S. Department of Commerce). La figure 27 représente une partie des attributs de la base de données extraite.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AAGE	C	X	1	0 1	Age Allocation Flag No Yes
AANCSTR1	C	X	1	0 1	First Ancestry Allocation Flag No Yes
AANCSTR2	C	X	1	0	Second Ancestry Allocation Flag No
...					
ADISABL1	C	X	1	0 1	Work Limitation Stat. Allocation Flag No Yes
ADISABL2	C	X	1	0 1	Work Prevention Stat. Allocation Flag No Yes
AENGLISH	C	X	1	0 1	Ability to Speak English Allocation Flag No Yes
...					
ANCSTRY1	C	X	3	999	Ancestry First Entry See Appendix I Ance Not Reported
ANCSTRY2	C	X	3	000 999	Ancestry Second Entry See Appendix I Anc No Secondary Ancestry Not Reported
AOCCUP	C	X	1	0 1	Occupation Allocation Flag No Yes
ARACE	C	X	1	0 1	Detailed Race Allocation Flag No Yes

VAR: = Variable Name  
TYP: = Variable Type (C = Categorical, N = Numeric Continuous)  
DES: = Designation (P = Primary Variable, X = Non-Primary)  
LEN: = Length (of the Variable in Characters)  
CAT: = Category (of the Variable)

**Figure 27 : quelques attributs de la base de données extraite**

### 2.2 Nettoyage et transformation des données

Un grand nombre d'attributs moins utiles dans les données initiales ont été supprimées, les variables continues qui sont de nombre limité ont été discrétisées et le petit nombre de variables discrètes qui ont un grand nombre de valeurs possibles ont été regroupées à avoir moins de valeurs possibles. Cela nous a donné 68 attributs catégoriques.

Old Variable	New Variable
Age	dAge
Ancstry1	dAncstry1
Ancstry2	dAncstry2
Avail	iAvail
Citizen	iCitizen
Class	iClass
Depart	dDepart
Disabl1	iDisabl1
Disabl2	iDisabl2
English	iEnglish
Feb55	iFeb55

**Figure 28 : Une partie des attributs préfixés.**

Variable	Procedure
dAge	discAge
dAncstry1	discAncstry1
dAncstry2	discAncstry2
dHispanic	discHispanic
dHour89	discHour89
dHours	discHours
dIncome1	discIncome1
dIncome2	discIncome2to8
dIncome3	discIncome2to8
dIncome4	discIncome2to8
dIncome5	discIncome2to8
dIncome6	discIncome2to8
dIncome7	discIncome2to8
dIncome8	discIncome2to8
dIndustry	discIndustry
dOccup	discOccup
dPOB	discPOB
dPoverty	discPoverty
dPwgt1	discPwgt1
dRearning	discRearning
dRpincome	discRpincome
dTravtime	discTravtime
dWeek89	discWeek89
dYrsserv	discYrsserv

**Figure 29 : La liste des attributs transformés.**

La transformation des données du format original vers un formalisme adéquat pour l'extraction des connaissances (la segmentation de ces données) était faite par la séquence d'opérations suivantes :

- **Randomisation** : les enregistrements des données originales ont été permutés aléatoirement.
- **Sélection des attributs** : un préfixe a été ajouté aux attributs originaux, le 'i' pour les attributs dont les valeurs ont été utilisées tel qu'elle était, et le 'd' pour les attributs dont les valeurs ont été transformées (voir figure 28).
- **Transformation** : dans cette étape les valeurs des variables préfixés par le 'd' sont transformés en de nouvelles valeurs (voir figure 29). La transformation des variables dAncestry1, dAncestry2, dHispanic, dIndustry, dOccup, dPOB a été faite par un déraffinement des valeurs originales de ces dernières (voir figure 30). Les variables restantes et celles résultantes du déraffinement sont des variables continues, la transformation de ces variables était de les discrétisées, les deux fonctions suivantes représentées en T-SQL (Transact-SQL) permettent la discrétisation des attributs dAge et dAncestry1 (après avoir été déraffiné) respectivement.

---

```
create function discAge( @arg varchar(255) )
```

---

```
RETURNS int
```

```
AS
```

```
BEGIN
```

```
  DECLARE @value bigint
```

```
  DECLARE @ret int
```

```
  SET @value = @arg
```

```
  IF @value = 0
```

```
    SET @ret = 0
```

```
  ELSE IF @value <13
```

```
    SET @ret = 1
```

```
  ELSE IF @value <20
```

```
    SET @ret = 2
```

```
  ELSE IF @value <30
```

```
    SET @ret = 3
```

```
  ELSE IF @value <40
```

```
    SET @ret = 4
```

```
  ELSE IF @value <50
```

```
    SET @ret = 5
```

```
  ELSE IF @value <65
```

```
    SET @ret = 6
```

```
  ELSE
```

```
    SET @ret = 7
```

```
  RETURN(@ret)
```

```
END
```

---

## Chapitre IV : Étude De Cas

---

---

```
create function discAncstry1( @arg varchar(255) )
```

---

```
RETURNS int
```

```
AS
```

```
BEGIN
```

```
  DECLARE @value bigint
```

```
  DECLARE @ret int
```

```
  SET @value = @arg
```

```
  IF @value = 999
```

```
    SET @ret = 0
```

```
  ELSE IF @value <100
```

```
    SET @ret = 1
```

```
  ELSE IF @value <200
```

```
    SET @ret = 2
```

```
  ELSE IF @value <300
```

```
    SET @ret = 3
```

```
  ELSE IF @value <360
```

```
    SET @ret = 4
```

```
  ELSE IF @value <400
```

```
    SET @ret = 5
```

```
  ELSE IF @value <500
```

```
    SET @ret = 6
```

```
  ELSE IF @value <600
```

```
    SET @ret = 7
```

```
  ELSE IF @value <700
```

```
    SET @ret = 8
```

```
  ELSE IF @value <800
```

```
    SET @ret = 9
```

```
  ELSE IF @value <900
```

```
    SET @ret = 10
```

```
  ELSE
```

```
    SET @ret = 11
```

```
  RETURN(@ret)
```

```
END
```

---

000-099	WESTERN EUROPE (EXCEPT SPAIN)
000-001	ALSATIAN
000-001	Alsace Lorraine
002	ANDORRAN
002	Andorra
003-004	AUSTRIAN
003	AUSTRIAN
003	Austria
004	TIROL
004	Tirol
005-007	BASQUE
005	BASQUE
005	Euskalduna
005	Euzkadi
006	FRENCH BASQUE
007	SPANISH BASQUE
007	Vasco
...	
400-499	NORTH AFRICA AND SOUTHWEST ASIA
400-401	ALGERIAN
400-401	Algeria
402-403	EGYPTIAN
402	Copt
402	Egypt
402-403	Fellahin
404-405	LIBYAN
404-405	Libya
406-407	MOROCCAN
406	MOROCCAN
406	Moor
407	IFNI
408-410	TUNISIAN
408-410	Tunisia
411	NORTH AFRICAN
412	ALHUCEMAS
412	Ceuta
412	Chafarinas
412	Melilla
413	BERBER
414	RIO DE ORO
414	Sagua El Hamra

**Figure 30 : Quelques valeurs résultantes du déraffinement de variable dAncstry1.**

10000	,5,0,1,0,0,5,3,2,2,1,0,1,0,4,3,0,2,0,0,1,0,0,0,0,10,0,1,0,1,0,1,
10001	,6,1,1,0,0,7,5,2,2,0,0,3,0,1,1,0,1,0,0,0,0,1,0,0,4,0,2,0,0,0,1,4
10002	,3,1,2,0,0,7,4,2,2,0,0,1,0,4,4,0,1,0,1,0,0,0,0,0,1,0,2,0,4,0,10,
10003	,4,1,2,0,0,1,3,2,2,0,0,3,0,3,3,0,1,0,0,0,0,0,0,1,4,0,2,0,2,0,1,4
10004	,7,1,1,0,0,0,0,2,2,0,0,3,0,0,0,0,0,0,0,0,1,0,0,0,0,0,2,2,0,0,0,4
10005	,1,1,2,0,2,0,4,0,0,0
...	
14277	,6,11,1,0,0,1,3,2,2,0,0,0,0,3,3,0,2,1,0,0,0,0,1,0,6,1,2,0,0,0,1,
14278	,7,1,2,0,0,0,0,2,2,0,0,4,0,0,0,0,0,0,0,0,1,0,0,0,0,0,2,2,0,0,0,4
14279	,2,1,1,0,0,2,0,2,2,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,9,0,2,2,4,0,0,4
14280	,2,11,1,0,0,0,0,2,2,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,2,4,0,0,
14281	,5,1,1,0,0,1,2,2,2,0,0,0,0,5,5,0,3,0,0,1,0,0,0,0,9,0,2,0,4,0,1,4
...	
21110	,3,1,1,0,0,1,5,2,2,0,0,3,0,3,3,0,1,0,0,0,0,1,0,0,7,0,2,0,2,0,10,
21111	,3,0,1,0,0,0,0,1,1,0,0,0,0,0,0,0,1,0,0,1,0,1,0,0,0,0,2,2,4,0,0,4
21112	,3,11,1,0,0,2,0,2,2,0,0,0,0,2,0,0,1,0,0,0,0,0,0,0,9,0,2,2,4,0,0,
21113	,4,2,1,0,0,4,2,2,2,0,0,3,0,1,3,0,1,0,0,0,0,1,0,1,8,0,2,0,3,0,2,4
...	
21114	,7,1,1,0,0,0,0,2,2,0,0,3,0,0,0,0,0,0,0,0,1,0,0,0,0,0,2,2,1,0,0,4
22212	,7,1,2,0,0,0,0,2,2,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,2,2,1,0,0,2
22213	,1,0,1,0,2,0,4,0,0,0
22214	,3,11,1,0,0,5,3,2,2,0,0,0,0,1,3,0,1,0,0,0,0,0,0,0,11,0,2,0,4,0,1
...	
23072	,6,1,1,0,0,6,1,2,2,0,0,2,0,5,5,0,0,1,0,0,0,0,0,0,7,0,2,0,0,0,1,4
23073	,7,1,1,0,0,0,0,2,2,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,2,2,4,0,0,2
23074	,4,1,1,0,0,1,2,2,2,0,0,0,0,3,3,0,2,0,0,0,0,0,0,0,3,0,2,0,0,0,1,4
23075	,1,1,2,0,2,0,4,0,0,0
...	
24045	,4,11,1,0,0,4,5,2,2,0,0,4,0,3,3,0,1,0,0,0,0,0,0,0,9,0,2,0,4,0,1,
24046	,4,11,1,0,0,1,5,1,2,0,0,0,0,4,3,0,2,0,0,0,0,0,0,0,4,0,2,0,0,0,1,
24047	,5,1,3,0,0,3,4,2,2,0,0,0,0,2,2,0,3,0,0,0,0,0,0,0,10,0,2,0,0,0,1,
24048	,6,1,1,0,0,5,3,1,2,0,0,0,0,3,3,0,3,0,0,1,0,0,0,0,5,0,2,0,0,0,1,4
24049	,4,3,1,0,0,5,2,2,2,0,0,0,0,3,3,0,2,0,0,0,0,0,0,1,5,0,2,0,0,0,1,2
24050	,6,1,1,0,0,1,4,1,2,0,0,0,0,3,3,0,4,0,0,1,0,0,0,0,8,1,2,0,0,0,1,
...	
28626	,6,1,1,0,0,2,1,2,2,0,0,0,0,3,5,0,1,0,0,0,0,0,0,0,5,1,2,0,1,0,1,2
28627	,1,11,1,0,4,0,0,
28628	,3,1,1,0,0,0,0,2,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,4,0,0,4
28629	,6,1,1,0,0,3,2,2,2,0,0,0,0,3,3,0,2,0,0,0,0,0,0,0,9,0,2,0,0,0,1,3
28630	,6,11,1,0,0,1,3,2,2,0,0,0,0,0,3,0,0,0,0,0,0,0,0,1,3,0,2,0,0,0,1,
28631	,6,1,1,0,0,1,3,2,2,0,0,0,0,3,3,0,2,0,0,1,0,0,0,0,4,0,2,0,0,0,1,4
28632	,2,1,1,0,0,1,5,2,2,0,0,1,0,1,1,0,1,0,0,1,0,0,0,0,7,0,2,0,4,0,1,4
...	
30080	,3,11,1,0,0,1,4,2,2,0,0,0,0,3,3,0,2,0,0,0,0,0,0,0,8,0,2,0,4,0,1,
30081	,6,11,1,0,0,0,0,2,2,0,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,
30082	,2,1,2,0,0,0,0,2,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,4,0,0,4
30083	,5,1,2,0,0,1,4,2,2,0,0,0,0,3,5,0,2,0,0,0,0,0,0,1,8,0,2,0,0,0,1,4
30084	,1,11,1,0,2,0,4,0,0,
30085	,4,0,1,0,0,1,3,2,2,0,0,0,0,4,4,0,3,0,0,0,0,0,0,0,8,0,2,0,0,0,1,4
30086	,7,1,1,0,0,4,0,2,2,0,0,5,0,3,0,0,0,0,1,1,1,0,0,0,10,0,2,0,1,0,0,

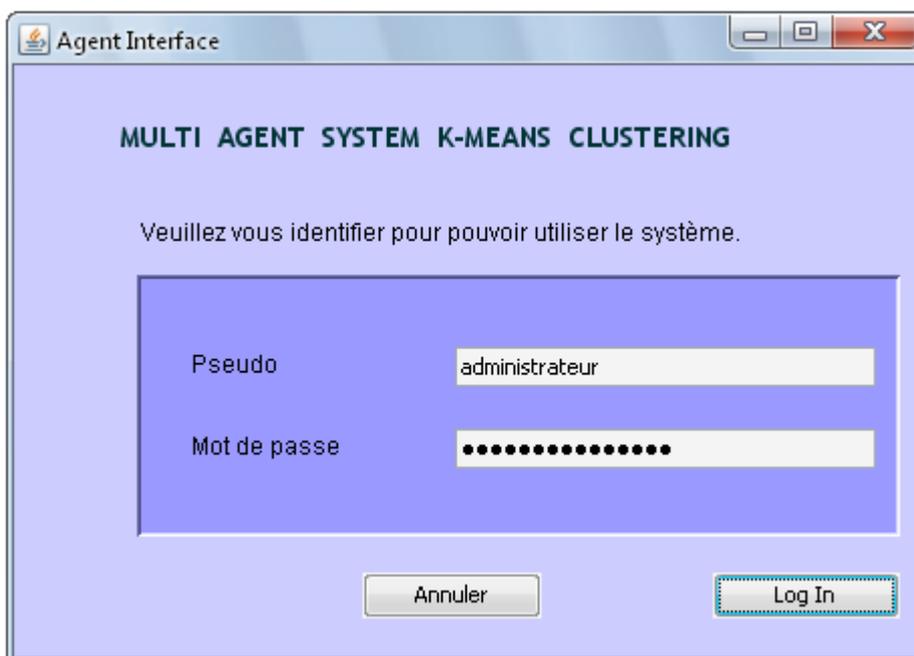
Figure 31 : Extrait de la base de données après les opérations de préparation et de transformation.

La figure 31 représente un extrait de la base de données après les opérations de préparation et de transformation des données. Ces données concernent respectivement les attributs suivant : caseid, dAge, dAncestry1, dAncestry2, iAvail, iCitizen, iClass, dDepart, iDisabl1, iDisabl2, iEnglish, iFeb55, iFertil, dHispanic, dHour89, dHours, iImmigr, dIncome1, dIncome2, dIncome3, dIncome4, dIncome5, dIncome6, dIncome7, dIncome8, dIndustry, iKorean, iLang1, iLooking, iMarital, iMay75880, iMeans de la base de données.

### 2.3 Data mining

Nous désignons par l'étape du data mining l'application de notre système sur le jeu de données précédemment préparé et dont le résultat sera la segmentation de ces données en un nombre de classes relativement homogènes.

La figure 32 montre la première fenêtre qui apparaît lors du lancement du système et qui est la fenêtre principale de l'agent Interface, elle désigne que l'agent est activé. Cette fenêtre permet d'identifier les utilisateurs ayant le droit d'utiliser le système.



**Figure 32 : fenêtre d'authentification.**

Après avoir été identifié, l'utilisateur est autorisé de poursuivre l'utilisation du système et cela en lui envoyant un message demandant de saisir le nombre de clusters qu'il veut avoir pour ses données, comme l'illustre la figure 33.



**Figure 33 : demande de saisie du nombre de clusters.**

Dès que le nombre de clusters est saisi, l'agent interface l'envoi aux deux agents Init-Cluster et Affect-Cluster en les activant, l'agent Init-Cluster commence à initialiser les premiers clusters et les envois par la suite à l'agent Affect-Cluster, ce dernier qui a déjà commencé à extraire les données de la base, reçoit les clusters initiaux et commence à affecter les autres données à ces clusters en bénéficiant les services offerts par les agents Calc-Centroïde et Calc-Distance (voir chapitre 4).

Après plusieurs itérations et dès que le test d'arrêt est validé, l'agent Affect-Cluster envoi le résultat de la segmentation à l'agent Interface, ce dernier se charge de l'afficher sur l'écran en retournant par conséquent le résultat à l'utilisateur.

La figure 34 représente le résultat de la segmentation de 1935 enregistrements de la base de données avec pour nombre de clusters  $k$  la valeur 11.



Une partie du code de notre système est présenté dans ce qui suit:

```
package k-means;

import java.util.ArrayList;
import java.util.Random;
import java.awt.geom.Point2D;
import jade.core.Agent;
import jade.core.Runtime;
import jade.core.ProfileImpl;
import jade.core.behaviours.CyclicBehaviour;
import jade.domain.DFService;
import jade.domain.FIPAException;
import jade.domain.FIPAAgentManagement.DFAgentDescription;
import jade.lang.acl.ACLMessage;
import jade.lang.acl.MessageTemplate;
import jade.lang.acl.UnreadableException;
import javax.swing.JFrame;
import javax.swing.JPanel;
import java.awt.BorderLayout;
import java.awt.Toolkit;
import javax.swing.JButton;
import javax.swing.JLabel;

public class InterfaceAgent extends Agent {
    private JFrame jFrame = null;
    ...
    // Initialisation de l'agent
    protected void setup() {
        getJFrame().setVisible(true);
        try {
            // Création de la description de l'agent
            DFAgentDescription dfd = new DFAgentDescription();
            dfd.setName(getAID());
            DFService.register(this, dfd);
        }
        catch (FIPAException e) {
            e.printStackTrace();
        }

        protected void takeDown() {
            try {
                // Suppression de l'agent du DF
                DFService.deregister(this);
            }
        }
    }
}
```

```
    }
    catch (FIPAException e) {
        e.printStackTrace();
    }
}
...
// Lecture de la valeur de k
    int NombreClust;
        NombreClust = 0; test = 0;
    do {
        BufferedReader stdin = new BufferedReader(new
InputStreamReader(System.in));
        System.out.println("Donnez le nombre de Clusters");
        String line = stdin.readLine();
        try{
            NombreClust = Integer.parseInt(line);
            test = 1;
        }
        catch(NumberFormatException NFE1) {
            System.out.println("Erreur, pas un Nombre");
            test = 0;
        }
    }
...
// Initialisation des clusters
// @param points
// @param clusters

void init(ArrayList<Point> points, ArrayList<Centroide> clusters) {
m_points = points;
m_clusters = clusters;

m_init= true;
}
...
// Calcul de la distance minimale de chaque point
//avec les centres des clusters
for (Point p : m_points) {
double distance = Double.MAX_VALUE;
Centroide min_centre = null;

for (Centroide c : m_clusters) {
```

## Chapitre IV : Étude De Cas

---

```
double d = p.getP().distance(c.getP());
if (d < distance) {
distance = d;
min_centre = c;}
}
}
...
// Calcul du centroïde de chaque cluster
for (Centroide c : m_clusters) {
c.addTrack((Point2D) c.getP().clone());
int num_points = c.getPoints().taille();
if (num_points > 0) {
double mean_x = 0.0;
double mean_y = 0.0;
for (Point p : c.getPoints()) {
mean_x += p.getP().getX();
mean_y += p.getP().getY();
}
mean_x = mean_x / num_points;
mean_y = mean_y / num_points;
c.getP().coordoné(mean_x, mean_y);
} else {

// Si un centroïde n'a aucun points reliés à lui, le placer à coté d'un
autre

//centroïde qui a plusieurs points
Centroide max = c;
int cont = 0;
for (Centroide autre_centre : m_clusters) {
int num = autre_centre.getPoints().taille();
if (num > cont) {
cont = num;
max = autre_centre;
}
}

Random r = new Random();
int devX = r.nextInt(20) - 10;
int devY = r.nextInt(20) - 10;
c.getP().coordoné(max.getP().getX() + devX, max.getP().getY() + devY);
```

```
}  
}
```

### 3 Outils de programmation

Pour implémenter un système tel celui que nous avons proposé nous avons besoin d'un langage tel JAVA.

#### 3.1 Pourquoi JAVA?

Le langage java, né en 1995, est un langage orienté objets, il permet d'écrire de façon simple et claire des programmes portables sur la majorité des plates-formes. De plus, il bénéficie d'une très grande bibliothèque de classes avec lesquelles l'utilisateur pourra composer des interfaces graphiques, créer des applications multithreads, animer une page HTML par des applets ou encore communiquer en réseau [50].

Dans notre cas nous avons besoin d'implémenter un système multi-agents. Pour cela nous devons utiliser une plateforme multi-agent qui est un ensemble d'outils nécessaire à la construction et à la mise en service d'agents au sein d'un environnement spécifique. Ces outils peuvent servir également à l'analyse et au test des SMA ainsi créés. Parmi les plateformes libres, il y en a quelques unes très connues pour avoir été utilisées dans le développement de plusieurs applications : JADE, Zeus, MadKit, AgentBuilder, Jack, JAFMAS, AgentTool, DECAF, RMIT, etc. Mais la plus connue des plateformes est JADE (Java Agent Development Framework).

#### 3.2 La plateforme JADE

JADE est une plateforme créée par le laboratoire TILAB, elle est implémentée entièrement en JAVA et répond aux spécifications FIPA (Foundation for Intelligent Physical Agents) qui est une organisation dont l'objectif est la production des standards pour l'interopération d'agents logiciels hétérogènes. Ainsi la plateforme JADE fournit un grand nombre de classes qui implémentent le comportement des agents qu'elle crée. JADE possède trois modules principaux (nécessaire aux normes FIPA) qui sont :

- **DF** "Directory Facilitator" fournit un service de "pages jaunes" à la plate-forme.
- **ACC** "Agent Communication Channel" gère la communication entre les agents.

- AMS “Agent Management System” supervise l'enregistrement des agents, leur authentification, leur accès et l'utilisation du système.

Ces trois modules sont activés à chaque démarrage de la plate-forme. Par ailleurs, la plate-forme possède une architecture très précise qui permet la construction normalisés d'agents. Pour cela, elle se décompose en plusieurs classes dont structure est illustrée par la figure suivante [51].

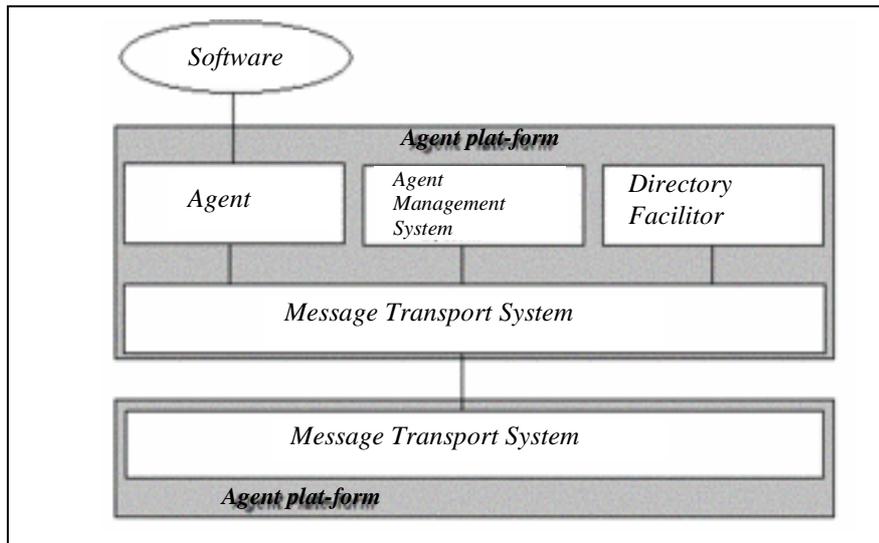


Figure 35 : structure de la plate-forme JADE [51].

À coté des modules décrits précédemment, la plateforme jade possède son propre nombre d'agents garantissant chacun un service différent, mais dont l'ensemble forme l'outil graphique de la plateforme, qui sont:

- **Le Remote Management Agent (RMA)** : il agit en tant que console graphique pour la gestion et le contrôle de la plateforme.
- **L'agent Dummy** : il permet le débogage et le suivi des agents et il permet aussi l'envoi et reçoit des messages.
- **L'agent Sniffer** : son rôle est de visualiser les interactions des agents en gardant un historique sur ces interactions, en d'autres termes il surveille les échanges de messages dans une plateforme.
- **L'agent Introspector** : cet agent prend en charge la surveillance de l'état des agents durant leurs cycles de vie, mais aussi les messages émis et reçus.
- **L'agent DF GUI** : il a pour but d'inspecter le service des pages jaunes [52].

### 3.2.1 L'environnement d'exécution JADE

Un avantage majeur de la plateforme JADE est de garantir la distributivité des agents et ce s'accomplit par la notion de conteneur.

L'environnement d'exécution s'appelle conteneur, chaque conteneur peut englober plusieurs agents. L'ensemble des conteneurs actifs constitue une plateforme. Cette plateforme peut s'exécuter sur plusieurs machines, elle contient toujours un conteneur spécial qui est dit principal (*Main Container*) et qui regroupe les agents fournissant les services de base de la plateforme.

La plateforme d'agents peut être distribuée sur plusieurs machines même dans le cas où ces derniers possèdent des systèmes d'exploitation différents.

L'architecture d'une plateforme JADE est basée sur la coexistence de plusieurs Machines Virtuelles (VM) Java et la communication entre elles se fait par la méthode RMI (Remote Method Invocation) de Java, qui est une API Java permettant de manipuler des objets distants instanciés sur une autre machine virtuelle, se situant sur une autre machine du réseau comme si l'objet était sur la machine virtuelle de la machine locale.

### 3.2.2 La communication entre les agents JADE

La plateforme JADE offre à ses agents la possibilité de communiquer via le langage de communication FIPA-ACL.

Chaque agent possède une file d'attente où il peut stocker les messages qu'il reçoit, la plateforme garantit le dépôt d'un message dans la file d'attente, dès lors l'agent récepteur sera prévenu de son arrivée et par la suite il peut prendre ce message au moment fixé par le programmeur.

Il existe plusieurs types de communications, selon la position des agents dans le système, nous pouvons citer :

- ***Même conteneur*** : dans ce cas où nous n'avons besoin d'invocation distante.
- ***Plusieurs conteneurs dans la même plateforme avec cache*** : dans ce cas l'RMI est invoqué une seule fois afin de sérialiser et désérialiser le message ACL.
- ***Plusieurs conteneur dans la même plateforme sans cache*** : deux appels RMI, le premier pour mettre à jour le cache depuis la table descriptive global des agents, et le second pour envoyer le message, l'objet qui décrit l'agent et qui est

retourné du premier appel est sérialisé et désérialisé, puis le message ACL suit le même chemin.

- **Plusieurs plateformes JADE** : un appel direct à distance est effectué vers le ACC, et par conséquent une invocation va -t- être provoquée à distance par la méthode CORBA de l'OMG (Object Management Group) qui est une méthode de communication permettant de manipuler des objets à distance avec n'importe quel langage contrairement à la méthode RMI qui est une solution tout Java, CORBA est beaucoup plus compliqué à mettre en œuvre, c'est la raison pour laquelle de nombreux développeurs se tournent généralement vers la RMI.

Suite à l'invocation CORBA citée ci-dessus, nous passons à une transformation d'un objet java vers une chaîne de caractères Java, et par la suite vers un flux de byte IIOP sur le côté expéditeur et une transformation inverse du coté récepteur.

- **Plusieurs plateformes non JADE**: la même chose que pour le cas précédent, mais ce qui se passe à l'autre bout de la liaison dépend de la nature et l'implémentation de l'autre plate-forme, celle qui reçoit le message [53].

L'utilisation de cette plateforme, facilite la programmation d'un système multi agents, de plus elle garantie la distributivité et ainsi la communication sur divers postes de travail sur le réseau, sans oublier le fait qu'elle fournit toutes les caractéristique dont nous avons besoin pour, modéliser le concept d'agents.

## 4 Conclusion

Dans ce chapitre nous avons présenté un exemple concret de l'application de notre système sur une base de données qui est celle du recensement de l'année 1990 des états unies d'Amérique. Le grand nombre de données que contient cette base ainsi que l'importance des informations que peuvent en être déduites était la principale cause de l'avoir choisi pour notre étude de cas.

Le résultat de la segmentation de cette base de données était représenté sous forme de diagramme montrant pour chaque point le segment à qui il appartient, ce résultat pourra avoir plusieurs traductions qui dépendront du besoin de la personne qui utilisera le système.

## **Conclusion générale**

## ***Conclusion générale***

Dans le cadre de cette thèse, nous avons essayé d'apporter des solutions au problème de complexité des systèmes de data mining. Pour y parvenir, nous nous sommes basés sur le couplage entre les systèmes multi agents et la fouille de données. Nous avons montré que ces deux domaines sont complémentaires et peuvent évoluer dans le cadre d'un processus unique. Leur association est capable de faciliter le processus d'extraction de connaissance.

Pour cela nous avons présenté la notion de data mining ou la fouille de données qui représente un ensemble de techniques d'exploration de données permettant d'extraire d'une base de données des connaissances sous la forme de modèles de description, les différentes tâches que peut effectuer un système de data mining et les différentes techniques qui peuvent être appliquées pour accomplir ces tâches.

Nous avons vu aussi la notion de systèmes multi-agents commettant un paradigme de modélisation des systèmes dits complexes dont le principe est de distribuer la complexité sur un ensemble d'entités communicantes, autonomes, réactives et dotées de compétences appelées agents ainsi que leurs contributions dans les systèmes de data mining qui représentent une certaine complexité dans la mesure où il faut fouiller des bases de données extrêmement géantes pour en extraire des informations utiles.

Par la suite nous avons présenté un modèle qui incarne une approche d'intégration d'agents dans le data mining, tout en argumentant notre opinion.

L'étude du cas sur la base de données du recensement de l'année 1990 des états unis d'Amérique nous a permis l'exploitation de notre approche et sa validation à travers un système de clustering composé de cinq agents, qui remplissent des fonctions complémentaires pour diminuer la complexité de la tâche de clustering. Ce système est réalisé dans un environnement assez performant pour la modélisation de systèmes multi agent qui est la plateforme JADE.

Le travail réalisé dans cette thèse ouvre diverses perspectives de recherche :

- Tout d'abord, essayer d'introduire la notion des algorithmes génétiques pour optimiser les résultats du clustering en utilisant des différentes valeurs de  $k$ .
- Rendre notre système applicable sur des données distribuées géographiquement à travers plusieurs sites.
- essayer d'augmenter le niveau de raisonnement de nos agents en utilisant par exemples un raisonnement à base de cas.

D'après les points que nous avons cités, nous pouvons juger le domaine data mining comme un domaine apte à l'évolution et riche en sujets de recherche qui attend, celui qui vient pour relever ses défis.

# *Références*

- [1] G. PIATETSKY-SHAPIRO, Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from «university» to «business» and «analytics», Data mining and Knowledge Discovery, 15(1), 99-105.
- [2] The Gartner Group, [www.gartner.com](http://www.gartner.com).
- [3] D. HAND, H. MANNILA et P. SMYTH, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.
- [4] P. CABENA, P. HADJINIAN, R. STADLER, J. VERHEES et A. ZANASI, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.
- [5] E-G. TALBI, Fouille de données (Data Mining) : Un tour d’horizon, Laboratoire d’Informatique Fondamentale de Lille.
- [6] J. HAN, M. KAMBER, Data Mining: Concepts and Techniques, Simon Fraser University, 2000.
- [7] J. HAN, M. KAMBER, Data Mining: Concepts and Techniques, Second Edition, University of Illinois at Urbana-Champaign, 2006.
- [8] M. J. BERRY, G. S. LINOFF, Data Mining Techniques For Marketing, Sales, and Customer Relationship, Management, Second Edition, 2004.
- [9] M. J. BERRY, G. S. LINOFF, Mastering Data Mining: The Art and Science of Customer Relationship Management, 2000.
- [10] D.T. LAROSE, Discovering Knowledge In Data: An Introduction to Data Mining, Central Connecticut State University, 2005.
- [11] B. AGARD, A. KUSIAK, Exploration Des Bases De Données Industrielles À L’aide Du Data Mining – Perspectives, 9ème Colloque National AIP PRIMECA, avril 2005.
- [12] Rapporté de <http://www.nasdaq.com/>
- [13] O. R. ZAÏANE, Principles of Knowledge Discovery in Databases, CMPUT690, University of Alberta, 1999.
- [14] Outil de reporting décisionnel : DATA WAREHOUSE - Cube OLAP : Présentation conviviale des données chiffrées, Rapporté de

<http://www.siom.fr/sycube.pdf>

- [15] Initiation au décisionnel (Business Intelligence, DataWarehouse, OLAP) Rapporté de <http://taslimanka.developpez.com/tutoriels/bi/>
- [16] Rapporté de [http://fr.wikipedia.org/wiki/Base\\_de\\_données\\_multimédia](http://fr.wikipedia.org/wiki/Base_de_données_multimédia)
- [17] G. HEBRAIL, Transformation de longues séries temporelles en descriptions symboliques, ENST Paris, LTCI-UMR 5141 CNRS, Département Informatique et Réseaux, France.
- [18] L'encyclopédie en ligne WIKEPEDIA 2009.
- [19] Ph. PREUX, Fouille de données : Notes de cours, Université de Lille 3, 9 octobre 2008.
- [20] G. DONG, J. PEI, Sequence Data Mining, Springer Edition, 2007.
- [21] S. PRABHU, N. VENKATESAN, Data Mining and Warehousing, New Age International (P) Ltd., Publishers, New Delhi, 2007.
- [22] Rapporté de [http://www.e-marketing.fr/Glossaire/ConsultGlossaire.asp?ID\\_Glossaire=5779](http://www.e-marketing.fr/Glossaire/ConsultGlossaire.asp?ID_Glossaire=5779).
- [23] Rapporté de <http://www.mercator-publicitor.fr/lexique-publicite-definition-marketing-direct>.
- [24] Introduction Data Mining, rapporté de [www-lmgm.biotoul.fr/enseignements/M2Pro\\_Bioinfo/intro.pdf](http://www-lmgm.biotoul.fr/enseignements/M2Pro_Bioinfo/intro.pdf)
- [25] G. CALAS, Études des principaux algorithmes de data mining, Spécialisation Sciences Cognitives et Informatique Avancée, France.
- [26] B. LAVOIE, Arbres de décisions, Synthèse de lectures, Séminaire sur l'apprentissage automatique, Programme de Doctorat en Informatique Cognitive, Université du Québec à Montréal, 15 mars 2006.
- [27] S. BOURAZZA, Variantes d'algorithmes génétiques appliquées aux problèmes d'ordonnancement, 2006.
- [28] R.GILLERON, M. TOMMASI, Découverte de connaissances à partir de données, 2000.
- [29] K. TEKNOMO, What is K Nearest Neighbors Algorithm?, <http://people.revoledu.com/kardi/tutorial/KNN/What-is-K-NearestNeighbor-Algorithm.html>, 2006.
- [30] C. SCHARFF, Méthode des k plus proches voisins, IFI, 2004.
- [31] Rapporté de [http://interstices.info/encart.jsp?id=c\\_41867&encart=3&size=](http://interstices.info/encart.jsp?id=c_41867&encart=3&size=)

600,500.

- [32] C. GROUIN, Les techniques de la fouille de données, INaLCO, 2009/2010.
- [33] La détection automatique de clusters, Rapporté de <http://www.eisti.fr/~lassi/PFE/ID/DataMining/SITE/detection.htm>
- [34] O. EL GANAOUI, M. Perrot, Segmentation par régions: une méthode qui utilise la classification par nuées dynamiques et le principe d'hystéresis, 31 decembre 2004.
- [35] K. P. SYCARA, Multiagent Systems, American Association for Artificial Intelligence, 1998.
- [36] S. G. CHEHBI, Intelligence Artificielle Distribuée, <http://www.maef-software.com/IADIR.html>, 02 avril 2009.
- [37] J. TISSEAU, G. PRIGENT, Simulation de la disponibilité des systèmes d'information -Etude & Réalisation d'une plateforme de simulation multi-agents -, Centre Européen de Réalité Virtuelle, France, 16 juin 1999.
- [38] J. FERBER, Les Systèmes Multi Agents: vers une intelligence collective, 1995.
- [39] F. SANDAKLY, Contribution à la mise en œuvre d'une architecture à base de connaissances pour l'interprétation de scènes 2D et 3D, thèse doctorat, université de Nice-Sophia Antipolis, 1995.
- [40] J. P. BRIOT et Y. DEMAZEAU, Introduction aux agents, 2001.
- [41] J-P. SANSONNET, A. BOULARIAS, La plateforme MadKit : Systèmes Multi-Agents, Master Recherche Informatique LRI Université Paris Sud XI, Février 2005.
- [42] Y. ZHANG, M. BRADY, and S. SMITH, Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximisation algorithm, IEEE Transaction on Medical Imaging, vol. 20, no. 1, pp. 45–47, 2001.
- [43] B. SCHERRER, M. DOJAT, F. FORBES, C. GARBAY, Une Approche SMA pour la Segmentation Markovienne des Tissus et Structures Présents dans les IRM Cérébrales, INRIA, Laboratoire Jean Kuntzmann, Université de Grenoble (MISTIS).
- [44] A. SAIDANE, H. AKDAG et I. TRUCK, Une Approche SMA de l'Agrégation et de la Coopération des Classifieurs, 3rd International

Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 27-31, 2005 – TUNISIA.

- [45] K. KOPERSKI, J. HAN, Discovery of spatial association rules in geographic information databases, 4th International Symposium Advances in Spatial Databases, SSD, Springer-Verlag, 1995, vol.951, pp. 47-66.
- [46] O. SHEHORY, K. SYCARA, P. CHALASAMI, et S. JHA, Agent cloning: an approach to agent mobility and resource allocation. IEEE Communications, Vol. 36, No. 7, Juillet, 1998, pp. 58-67.
- [47] R. BEN HAMED, H. BAAZAOUI, S. FAIZ, Vers un algorithme RASMA : RAS basé multi-agent, Extraction et Gestion des Connaissances EGC, Lille, 17 janvier 2006.
- [48] F. Y. VILLEMEN, Agent Communication Language, Systèmes Intelligents NFP212, Année 2009-2010.
- [49] <http://www.grappa.univ-lille3.fr/polys/fouille/sortie004.html#fig:prep>
- [50] I. CHARON, Le langage Java : concepts et pratique -le JDK 5.0, janvier 2006.
- [51] 20 notes sur java pour le web, JADE (Java Agent DEvelopment Framework), <http://20-notes-sur-java-pour-le-web.over-blog.com/article-3279214.html>, juillet 2006.
- [52] T. YUAN, Software Agents, Introduction to JADE, 2008.
- [53] F. BELLIFEMINE, A. POGGI, G. RIMASSA, JADE – A FIPA compliant agent framework.