UNIVERSITE DE SAAD DAHLAB BLIDA

Faculté des sciences
Département d'Informatique

MEMOIRE DE MAGISTER

Spécialité système d'information et de la connaissance

GESTION DES CONNAISSANCES DANS LE DOMAINE MÉDICAL

Présenté par
ZERF Nadjet épouse BOUDJETTOU

Devant le jury composé de :

N.Benblidia	Maître de conférence, U. de Blida	Présidente.
R.Chalal	Maître de conférence, ESI	Examinateur
R.Ghomari	Maître de conférence, ESI	Examinateur
S.Oukid	Maître de conférence, U. de Blida	Rapporteur

Blida, Mars 2011

RESUME

La gestion des connaissances (knowledge management) est « un mode de gestion systématique des savoir- faire et des connaissances dans les organisations, dont la finalité est de leur permettre d'obtenir un avantage compétitif ». La gestion des connaissances a donc pour mission d'améliorer la performance de l'entreprise. En outre, elle permet d'obtenir une vision d'ensemble des compétences et des savoirs de l'entreprise.

La gestion des connaissances dans le domaine médical vise à améliorer les performances de l'organisation médicale en permettant aux individus de l'établissement de soins (médecins, infirmières, paramédicaux, etc.) de capturer, partager et appliquer des connaissances collectives pour prendre des décisions optimales en temps réel.

Dans ce travail, nous proposons une approche de gestion des connaissances basée sur les techniques d'intégration des données hétérogènes dans le domaine médical en réalisant un entrepôt de données, puis nous choisissons une technique du data mining pour l'extraction des connaissances à partir des données médicales, et enfin nous exploitons ces connaissances dans un système de raisonnement à base de cas pour l'orientation des patients vers les différents services.

Mots-clés. Entrepôt des données, Data mining, Extraction des connaissances à partir de données ECD, Gestion des connaissances médicales, règle d'association.

ملخص

إن تسيير المعارف هو نمط التسيير التنظيمي للمعلومات داخل المؤسسات بغية التمكن من الحصول على أسبقية تنافسية، إن تسيير المعارف مهمته تحسين الأداء المؤسساتي (داخل المؤسسات) إضافة إلى التمكين من الحصول على رؤية شاملة للامكانتيات و المعارف داخل المؤسسات .

إن تسيير المعارف في المجال الطبي يهدف إلى تحسين الأداء داخل المؤسسة الاستشفائية من خلال السماح للإفراد داخل هاته المؤسسات العلاجية (أطباء، ممرضون، شبه طبيون...) من تحصيل و تطبيق المعارف الجماعية بغية اتخاذ قرارات صحيحة.

في هذا العمل، نقترح مقاربة تسييريه للمعارف متمحورة على تقنيات اندماج المعلومات الغير المتجانسة في المجال الطبي من اجل خلق قاعدة معلوماتية و من ثمة اختيار المعلومات الصحيحة بدءا من المعلومات الطبية ، و في الأخير علينا استغلال و استخدام هذه المعارف في نظام التفكير المرتكز على الحالات لتوجيه المرضى إلى مختلف المصالح.

الكلمات الرئيسية: خزان المعلومات، استخراج المعارف من المعلومات، تسيير المعرفة الطبية، القواعد المشتركة.

Abstract

Knowledge management (KM) is "a systematic method of management know-how and knowledge in organizations, whose purpose is to enable them to obtain a competitive advantage." Knowledge management has the task of improving business performance. In addition, it provides an overview of the skills and knowledge of the company.

Knowledge management in the médical field aims to improve the performance of the médical organization by allowing individuals in the care facility (doctors, nurses, paramedics, etc...) to capture, share and apply collective knowledge in order to make optimal decisions in real time.

In this work we propose a knowledge management approach based on integration technique of hétérogèneous data in the médical field by creating a data warehouse, a technique of extracting knowledge from médical data by choosing a technique of data mining, and finally an exploitation technique of that knowledge in a case-based reasoning system.

Keywords. Data warehouse, Data Mining, Knowledge Discovery in Database (KDD), Médical Knowledge Management, Association Rules.

REMERCIEMENTS

Je tiens à remercier très sincèrement ma promotrice M^{me} S.OUKID pour ses précieux conseils, son indéfectible patience et surtout pour la disponibilité dont elle a fait preuve tout au long de ce travail de recherche. J'espère avoir acquis un peu de sa rigueur et de ses qualités scientifiques.

Je remercie Melle N.BENBLIDIA maître de conférences à l'université de Blida qui m'a fait l'honneur de présider le jury de soutenance.

J'aimerais par ailleurs exprimer toute ma gratitude aux vénérables membres du jury : Mr R.CHALAl maître de conférences à E.S.I et Mr R.GHOMARI maître de conférences à E.S.I, qui m'ont fait l'honneur de leur présence, et qui ont fait montrer d'un grand intérêt lors de ma soutenance.

De nombreuses personnes m'ont aidé ou encouragé pendant mes études. À mes collègues, amis et personnels de la faculté des sciences, j'aimerais également vous offrir mon amicale reconnaissance pour votre présence, gentillesse et appui qui ont su égayer ces années de travail.

Je voudrais plus particulièrement exprimer ma reconnaissance envers Ratiba et Brahim, avec qui une collaboration scientifique fructueuse s'est établie.

Il m'aurait été impossible de réaliser ce travail sans l'affection et le soutien de ma famille. Un gros merci à mes parents pour leur soutien et pour m'avoir transmis la confiance dans ma vie. Merci à mes frères Fouad et Mohamed, un très grand merci à toutes mes sœurs Nora, Samira, Ahlem, et les petits Ibtissem, Imed, Ayah et Latif. Ainsi qu'à toute ma belle famille.

Sans Hamza, mon cher mari, qui m'a poussée vers la recherche, cette thèse n'aurait pas vu le jour. Je le remercie infiniment pour son amour, sa présence et son encouragement de multiples façons dans les moments les plus difficiles. Je lui dédie et à mes filles Maria et Ikhlass ma thèse.

TABLE DES MATIERES

LISTE DES ILLUSTRATIONS, GRAPHIQUES ET TABLEAUX

LISTE DES EQUATIONS

INTRODUCTION

La gestion des connaissances (knowledge management) consiste à acquérir et représenter les connaissances utiles à un domaine, une tâche ou une organisation particulière dans le but d'en favoriser l'accès, la réutilisation et l'évolution. Ayant détecté les connaissances cruciales et les besoins d'un domaine, cela revient généralement à construire, maintenir et faire évoluer une représentation explicite de ces connaissances (appelée mémoire d'entreprise) **[1]**. Il s'agit ensuite de fournir un accès à ces connaissances, c'est-à-dire de les diffuser dans le but d'en permettre une utilisation efficace. La mémoire d'entreprise, support privilégié de la gestion, se situe en effet dans un contexte de très long terme puisqu'elle vise à capitaliser des connaissances dans le but de pouvoir les réutiliser en dehors de leur contexte d'origine et pour des objectifs différents.

La gestion des connaissances médicales vise à améliorer les performances de l'organisation médicale en permettant aux individus de l'établissement de soins (médecins, infirmiers, paramédicaux, etc.) de capturer, partager et appliquer des connaissances collectives pour faire prendre des décisions optimales en temps réel. La gestion des connaissances médicales a pour vocation de traiter les domaines entiers de la médecine et de la santé, elle s'occupe de :

- Augmenter la fiabilité des données (saisie, enregistrement, transmission);
- Sélectionner les données les plus pertinentes parmi la masse des informations disponibles;
- Régler le problème d'hétérogénéité des données et des sources de données médicales;
- Partager l'information et faciliter l'accès à la connaissance;
- Exploiter les connaissances médicales pour la prise de décision;
- Comprendre les mécanismes d'interprétation et de raisonnement médical;

- Rassembler le savoir et le savoir-faire des praticiens sur des supports facilement accessibles.

1 La problématique

Les systèmes d'informations hospitaliers sont au cœur du fonctionnement des structures de santé dont dépend en grande partie l'amélioration de la qualité de la prise en charge médicale des patients dans ces établissements.

L'orientation des patients vers les différents services et les différentes spécialités est un problème majeur dans les grands établissements de santé tels que les CHU (Centre Hospitalo-Universitaire). Si le patient ne peut pas reconnaître les premiers signes d'un problème médical ou d'une lésion, il peut prendre un rendez-vous de plus de deux ou trois mois chez un médecin spécialiste dans un service sans qu'il soit le bon médecin.

La prise de décision médicale pertinente tant au niveau diagnostic que thérapeutique nécessite une bonne connaissance du cas à traiter. Le médecin a besoin de connaître ses données cliniques, ce qui suppose une forte collaboration entre les différents professionnels de santé et une interopérabilité entre les systèmes utilisés. Etant donné la complexité du domaine médical, nous rencontrons plusieurs problèmes tels que:

✓ La diversité des sources d'information distribuées et leur hétérogénéité sont une des principales difficultés rencontrées dans le domaine médical. Cette hétérogénéité peut provenir du :

- Format ou de la structure des sources (sources structurées : bases de données relationnelles, sources semi-structurées : documents XML, ou non structurées : textes),
- Mode d'accès et de requête ou de l'hétérogénéité sémantique.

✓ Les problèmes d'accès aux informations pertinentes pour les soins sont liés à la dispersion des informations médicales sur différentes structures de santé dont les systèmes d'information sont souvent autonomes et hétérogènes.

✓ La formalisation de l'information ne permet pas l'extraction, le partage, la diffusion et l'exploitation de ces connaissances médicales.

✓ La difficulté de comprendre les mécanismes d'interprétation et de raisonnement médical.

2 L'objectif du mémoire

La prise en compte de ces problèmes est une des clés de la mise en place d'un système qui permet à la gestion des connaissances d'intervenir dans le domaine médical, nous proposons un système qui peut :

- Donner une solution d'intégration des données hétérogènes : elle cherche à résoudre tous ces types de conflits afin de combiner et fusionner correctement les informations issues de ces sources distantes. « L'entrepôt de données » (data warehouse) est l'approche proposée pour résoudre ce type de problème, elle consiste à extraire, à l'avance, les données pertinentes pour l'usage des utilisateurs, à les filtrer, les transformer et les stocker.

- Proposer une démarche d'extraction des connaissances pertinentes à partir des données. Cette démarche n'intègre aucune connaissance *a priori,* elle est basée sur des expériences passées et sans l'introduction de connaissances de l'expert. Le data mining présente des outils performants pour l'extraction et la structuration des connaissances médicales, à partir des données collectées par l'entrepôt de données.

- Proposer un système de raisonnement à base de cas de l'orientation des patients vers des services médicaux. La base de cas de ce système s'appuie sur les connaissances déjà extraites par le système du data mining précèdent.

Les premiers systèmes intelligents d'aide au diagnostic en médecine étaient des systèmes de raisonnement à base de cas. Puis, au fur et à mesure du temps, et suivant les besoins, des méthodes de data mining ont été intégrées à ces systèmes. Ces méthodes sont venues combler les lacunes des systèmes de Raisonnement à Base de Cas.

Notre système propose une solution de combinaison des technologies proposées: l'intégration des données hétérogènes, le data mining et le raisonnement à base de cas (CBR), pour permettre l'extraction automatique de nouvelles connaissances ou l'apprentissage automatique.

3 Structure du mémoire

Nous adoptons une approche pratique pour mener à bien nos recherches et valider nos propositions. Ainsi, pour présenter nos travaux, nous avons articulé ce mémoire en deux parties.

La première partie contient quatre chapitres, elle présente le contexte du problème, elle a pour but de montrer la synergie qui existe entre les technologies : la gestion des connaissances, les entrepôts de données, l'extraction des connaissances à partir de données et le raisonnement à base de cas.

Dans le premier chapitre, nous exposons le contexte de nos travaux, ou nous présentons la gestion des connaissances, les formes des connaissances et les processus de la gestion des connaissances, ensuite nous implémentons la gestion des connaissances dans notre domaine d'étude, c'est le domaine médical ou nous citons les caractéristiques de la connaissance médicale, les objets de base de la gestion des connaissances médicales, son cycle de vie, et enfin nous parlons sur le problème d'hétérogénéité des données médicales.

Dans le deuxième chapitre, nous proposons une plate forme pour l'intégration des données hétérogènes afin de combiner et fusionner correctement les informations et les connaissances issues des sources médicales. L'approche proposée dans ce cadre est l'entrepôt des données.

Dans le troisième chapitre nous traduisons les techniques d'extraction de connaissances à partir de données (ECD) stockées dans les entrepôts et le noyau de ce processus qui est le data mining, à travers ses différentes phases et ses méthodes. Ensuite nous citerons des exemples du data mining appliqués dans le domaine médical.

Enfin, le quatrième chapitre est consacré au raisonnement à base de cas (CBR), nous décrivons les modèles et techniques utilisés dans ce domaine, en introduisant notamment les étapes du raisonnement, ses composants et les conditions de son application, ensuite nous citerons les systèmes CBR dans le domaine médical.

La deuxième partie est consacrée à nos motivations et à la proposition de notre approche, elle contient deux chapitres, elle est organisée comme suit :

A partir de l'étude présentée dans la première partie, nous abordons dans le cinquième chapitre notre architecture pour permettre à résoudre notre problème : Dans la première partie de ce chapitre nous abordons la procédure d'intégration des données hétérogènes dans le domaine médical par la construction d'un entrepôt de données médicales; après, nous proposons une architecture sur l'extraction des connaissances à partir de données en passant par toute ses étapes et par le choix d'une technique du data mining. La troisième étape consiste à développer un système de raisonnement à base de cas en basant sur la base de connaissances extraites à partir de l'étape précédente.

Dans le dernier chapitre, nous essayons de valider l'approche adoptée dans le chapitre précédent et ceci en développant une application qui permettra de réaliser notre système.

La conclusion générale dresse un bilan global sur les recherches dans ce domaine et notre contribution dans ces recherches et enfin les perspectives attendues pour complémenter ce travail.

CHAPITRE 1

LA GESTION DES CONNAISSANCES MEDICALES

1.1 Introduction

Le nombre des connaissances disponibles dans les domaines de la médecine, de la biologie et de la santé publique s'accroît. En santé notamment, le développement de la connaissance clinique et physio-pathologique, la technicité croissante des examens complémentaires, la multiplication et la diversification des structures de prise en charge des malades, l'allongement de la durée de vie des individus et l'augmentation du nombre de paramètres nécessaires à la prise en charge des patients pose le problème de la maîtrise de l'information et de la connaissance. En santé, l'information pertinente concerne des groupes d'individus et comprend des données démographiques, épidémiologiques, sanitaires, sociales, économiques voire politiques. Face à l'augmentation des connaissances médicales et des paramètres de soins à prendre en compte, il apparaît nécessaire de recourir aux méthodes et techniques de la gestion de connaissances.

La gestion des connaissances dans le domaine médical pose un problème majeur. En effet :

* Il faut rester proche de la structure naturelle de l'information pour la dénaturer le moins possible la standardisation du langage médical.

* Il faut adopter la représentation informatique la plus efficace de la structure des données.

Cela soulève des problèmes au centre des réflexions de la gestion de connaissances médicales:

- Comment faut-il organiser les informations de façon à obtenir le système le plus efficace?
- Comment peut-on représenter les données présentées dans le langage en leur conservant le maximum de richesse?
- Comment extraire les connaissances à partir des données?
- Comment traiter ces connaissances, les partager et les exploiter?

Pour répondre à toutes ces questions nous allons introduire dans ce chapitre la gestion des connaissances et ses concepts, puis la gestion des connaissances dans le domaine médical.

1.2 La Gestion des Connaissances (GC)

La gestion des connaissances (GC) est un moyen permettant autant que possible, de valoriser les capacités et l'expérience de chacun à la place qui lui convient le mieux, de faire circuler l'information utile et d'aider à trouver au bon moment celle dont on a réellement besoin dans l'action **[2]**.

La gestion des connaissances (Knowledge Management) s'occupe ainsi d'identifier, collecter, capitaliser et diffuser les connaissances d'une entreprise **[3]**. Ce travail est loin d'être trivial, car au-delà des questions de base qu'il soulève (qu'est-ce qu'une connaissance, comment la représenter et comment structurer une collection de connaissances), il implique souvent d'organiser les connaissances selon une forme réutilisable dans des conditions inconnues au moment de cette organisation. En effet, l'un des problèmes de la capitalisation des connaissances est de définir des critères pour le choix des connaissances à conserver, dans le but de prévoir une réutilisation de ces connaissances.

Nous nous intéressons plus particulièrement à la mise en oeuvre de connaissances utiles à la décision médicale afin d'en fournir un accès intelligent à travers d'outils d'aide à la décision.

1.2.1 Quelques définitions

Avant d'aborder la partie proprement liée à la gestion des connaissances, il convient de donner une brève explication des termes employés:

1.2.1.1 Donnée

Une donnée est un élément (signe ou un symbole) brut, qui n'a pas encore été interprété, mis en contexte. Une donnée est un fait discret et objectif se rapportant à un événement. Dans un contexte organisationnel, les données sont le plus souvent décrites comme des enregistrements structurés de transactions.

Les données décrivent seulement une partie de ce qui se passe lors d'une transaction. Elles ne fournissent pas un jugement ou une interprétation de la situation et ne

peuvent par conséquent constituer la base d'une prise de décision orientée vers une action à commencer.

1.2.1.2 Information

Une information est une donnée interprétée. En d'autres termes, la mise en contexte d'une donnée crée de la valeur ajoutée pour constituer une information et lui donner une signification. Considérons la donnée suivante: "34°C". Il s'agit bien d'un élément brut en dehors de tout contexte. Nous savons que cette donnée représente une température mais elle ne possède pas de réelle valeur. Est-ce la température de l'air ambiant? De l'eau du robinet? Ou d'un quelconque objet? Si nous remettons cette donnée dans son contexte, par exemple le bulletin des prévisions de météo pour la ville de Blida, nous créons de la valeur. Nous savons désormais à quoi correspond cette donnée initiale: la température sera de 34°C demain à Blida. Nous sommes passé d'un élément brut à un fait, d'une donnée à une information.

1.2.1.3 Connaissance

« La connaissance est un savoir incarné dans une personne physique. Autrement dit quelque chose qui est su par quelqu'un est une connaissance pour ce quelqu'un. La connaissance renvoie toujours à un possesseur humain à même de mettre en oeuvre, elle est à la base des compétences de tout individu » **[4]**.

Les connaissances représentent des informations organisées de manière à permettre d'agir, elles permettent à un individu de formuler un concept, à exploiter des informations à l'aide de mécanismes généraux de réflexion. Reprenons l'exemple précédent. Cette information nous permet de savoir qu'il va faire chaud demain à Blida, et nous allons nous habiller moins en conséquence. Imaginons qu'il ne fasse finalement que 16°C à Blida. La donnée relative à la température a changé, modifiant par la même occasion l'information: "la température sera de 16°C demain à Blida". La compréhension du modèle température-vêtement permet, au détenteur de cette connaissance, d'interpréter n'importe quel changement de la donnée température et d'agir en conséquence. Dans notre exemple, la connaissance de la relation entre la température et la manière de s'habiller nous permet de savoir qu'il faut mettre un pull.

On peut subdiviser la connaissance en connaissance factuelle, stratégique et de jugement.

- Le premier type de connaissance vérifie la capacité à décrire simplement des objets, des relations, des situations, des taxonomies, des structures, des topologies…

- Le second type (stratégique) permet à la personne qui possède une telle connaissance de définir la séquence de tâches dans une entreprise grâce à ses capacités à diriger et à contrôler.
- Finalement les connaissances de jugement permettent de décrire les façons de raisonner, de prescrire une action à entreprendre.

En résumé, la connaissance est donc un dérivé d'information qui est elle même un dérivé de donnée. Alors que l'on trouve les données dans des transactions ou dans des enregistrements, l'information dans des messages, la connaissance est obtenue à travers des individus ou des groupes de personnes possédant cette dernière, ou parfois encore dans des routines organisationnelles. Elle est livrée au moyen de supports structurés comme des livres ou des documents divers et à travers des contacts entre personnes qui peuvent aller de la simple conversation jusqu'à une relation contractuelle d'apprentissage.

1.2.1.4 Savoir

«Le savoir est quelque chose qui est su par un individu, une structure ou un collectif, voire une machine. Le savoir peut s'oublier, il peut aussi se stocker dans une mémoire humaine ou non humaine» **[4].**

1.2.2 Connaissances tacites et Connaissances explicites

Dans leur ouvrage de référence, "The knowledge-Creating Company", Nonaka et Takeuchi, deux experts japonais du knowledge management, mettent en évidence que la connaissance se présente sous deux formes différentes : une forme tacite et une forme explicite **[5].**

1.2.2.1 La connaissance tacite

Selon la définition de Polanyi dans **[6]** les connaissances tacites sont des connaissances non-verbalisables, intuitives et non-articulables.

La connaissance tacite est acquise par une expérience de collaboration et reste difficile à articuler, à formaliser et à communiquer. Elle est inconsciemment comprise et appliquée par la personne qui la possède, elle ne peut être communiquée directement d'une manière codifiée. Cette dernière concerne une expérience directe qui n'est pas codifiable par l'intermédiaire d'objets façonnés.

Ainsi Leonard et Sensiper **[7]** expliquent les connaissances tacites comme étant cachées, intangibles, subjectives et spontanées. Elles découlent en effet, de nos propres expériences et proviennent tout droit de notre inconscient ou de notre subconscient.

1.2.2.2 La connaissance explicite

Selon la définition de Hall et Andriani **[8],** la connaissance explicite est la connaissance ayant été capturée dans un code ou une langue qui facilite la communication. Elle implique le savoir-faire transmissible en langue formelle et systématique, elle exige une expérience directe de la connaissance. Elle peut être formellement articulée ou codée, plus facilement transférée ou partagée, objective et accessible.

La connaissance explicite joue un rôle toujours plus grand au sein des organisations, et de nombreuses personnes la considèrent de plus en plus comme un important facteur de production du savoir.

1.2.3 La création des connaissances

Selon Nonaka et al. une organisation crée de la connaissance grâce à l'interaction entre la connaissance explicite et de la connaissance tacite. La figure 1.1 ci-dessous illustre les différents modes de création des connaissances **[9].**

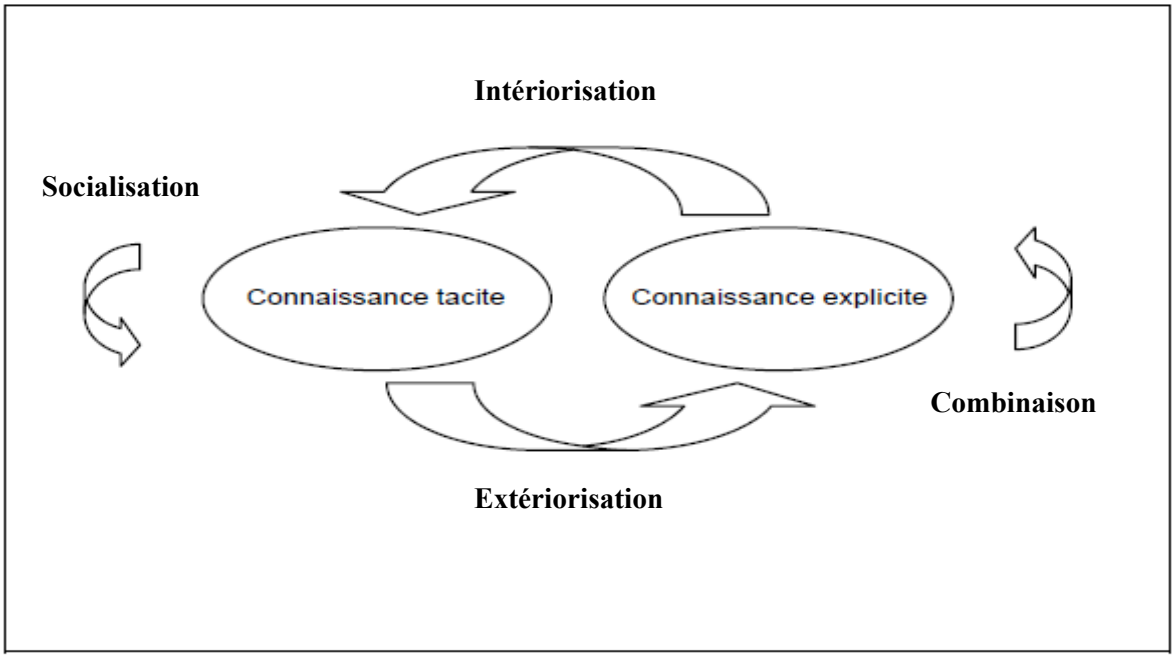

Figure 1.1 : Les modes de création de connaissances **[9].**

1.2.3.1 Socialisation

C’est un processus qui permet de convertir de la nouvelle connaissance tacite à travers des expériences communes. Comme la connaissance tacite est difficile à formuler et qu'elle est souvent spécifique, elle ne peut qu'être acquise à travers des expériences communes, par exemple en vivant ensemble ou en passant du temps ensemble (exemple : apprentissage).

1.2.3.2 Extériorisation

C’est le processus d'articuler la connaissance tacite en connaissance explicite. Quand la connaissance tacite est rendue explicite, elle est cristallisée, permettant ainsi d'être partagée par les autres et devient ainsi une base de nouvelle connaissance.

1.2.3.3 Combinaison

C’est le processus de convertir de la connaissance explicite en une série plus complexe et plus systématique de connaissance explicite. La connaissance explicite est collectée de l'intérieur et de l'extérieur de l'organisation, éditée ou traitée pour former de la nouvelle connaissance. Cette nouvelle connaissance explicite est ensuite divisée à travers les membres de l'organisation. L'utilisation créative de réseaux de communication informatisés et de bases de données facilite ce mode de conversion de connaissance.

1.2.3.4 Intériorisation

C’est le processus de convertir de la connaissance explicite en connaissance tacite. A travers l'intériorisation, la connaissance explicite crée est partagée à travers toute l'organisation et est convertie en connaissance tacite par les individus. Ce processus s'apparent beaucoup à du learning.

1.2.4 Les processus de gestion des connaissances

Afin d'illustrer les différents processus à prendre en compte pour une bonne gestion des connaissances, nous allons nous baser sur le modèle proposé par le Professeur Gilbert Probst et al. **[10].** Cette approche préconise de définir des taches afin de structurer la gestion des connaissances en une suite logique de processus, de suggérer des points d'intervention efficaces et finalement de fournir un modèle de base pour analyser et diagnostiquer rapidement les sources des problèmes de connaissances. Ce modèle permet de mettre en évidence l'interdépendance de ces taches. En effet, il est essentiel de ne pas

conduire des activités de gestion des connaissances de manière isolée les unes par rapport aux autres. La figure ci-dessous présente le modèle en question:

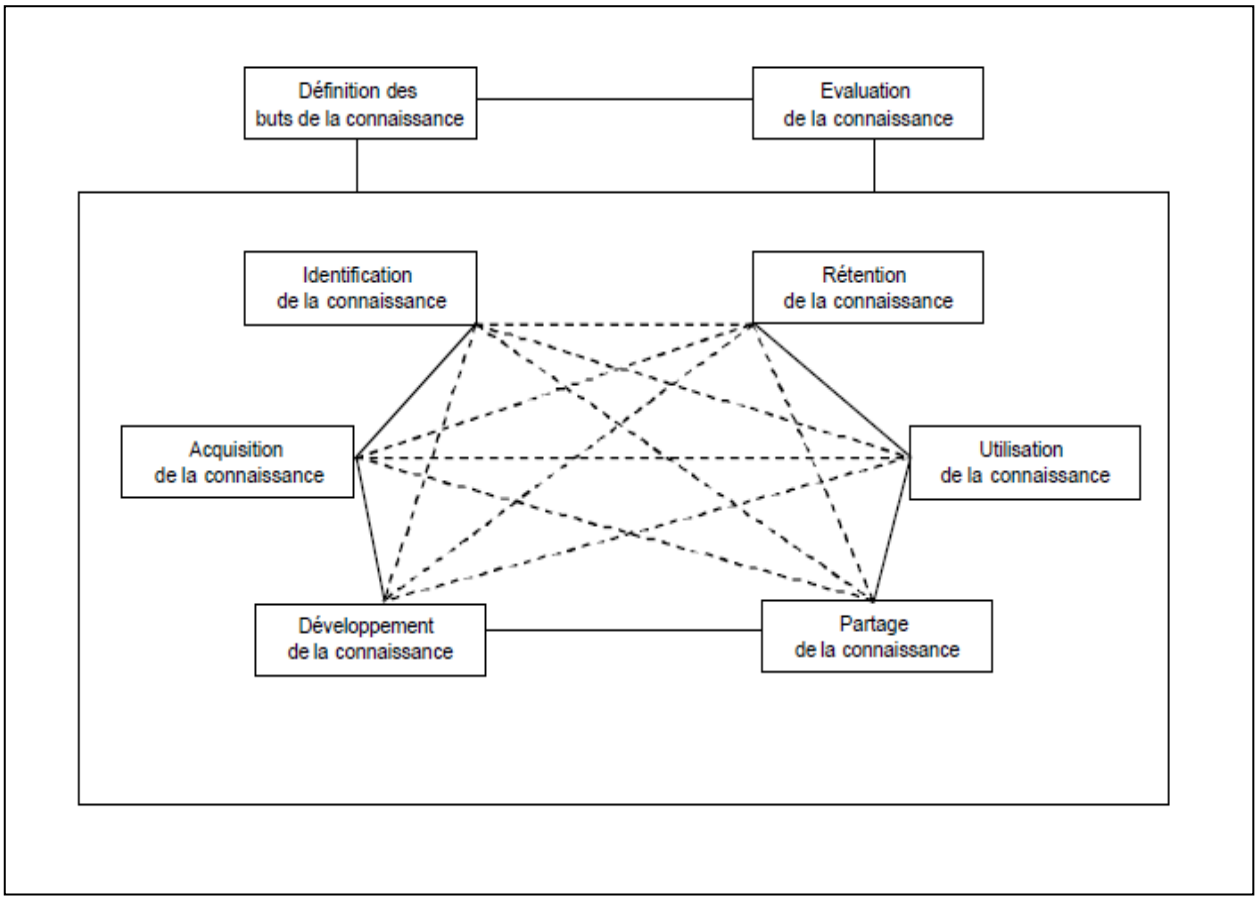

Figure 1.2 : Les processus de gestion des connaissances **[10].**

Nous allons maintenant voir en détail les différents processus de ce modèle:

1.2.4.1 Définition des buts de la connaissance

Les buts de la connaissance montrent le chemin à suivre aux activités relatives à la gestion des connaissances. Ils déterminent quelles compétences doivent être construites et à quel niveau.

• Les buts normatifs: contribuent à la création d'une culture d'entreprise sensible à la connaissance, dans laquelle les notions de partage et de développement de la connaissance créent des pré-conditions à la gestion efficiente des connaissances.

• Les buts stratégiques: définissent les capacités de base de l'entreprise et décrivent les futurs besoins en connaissances de cette dernière.

• Les buts opérationnels: s'assurent que les buts stratégiques et normatifs soient traduits en action concrète.

1.2.4.2 Identification de la connaissance

Avant d'investir dans le développement de nouvelles compétences, l'entreprise doit savoir quelles connaissances et expertise existent aussi bien à l'interne qu'à l'externe. On perd bien souvent la trace de données, d'informations et de capacités internes et externes. Ce manque de transparence conduit à de l'inefficacité, à des mauvaises décisions et à des activités redondantes. Une façon d'accroître ce niveau de transparence est la création de cartes de connaissances. Grâce aux technologies de l'information, de nombreux outils relativement performants, ont vu le jour et permettent de supporter un accès systématique aux différentes parties de la base de connaissance de l'entreprise.

1.2.4.3 Acquisition de la connaissance

La croissance explosive de la quantité d'informations et de connaissances de ces dernières années n'a pas rendu les choses plus faciles pour les entreprises. Pour pallier à ce problème, elles doivent, en général, acquérir cette connaissance auprès des experts (exemple : recruter des experts, des spécialistes).

1.2.4.4 Développement de la connaissance

Cela consiste à toute activité de gestion visant à produire de la nouvelle connaissance interne ou externe aussi bien sur le plan individuel que sur le plan collectif. Le développement de la connaissance, sur le plan individuel, repose sur la créativité et sur une résolution de problème systématique.

Le développement de la connaissance, sur le plan collectif, implique la dynamique de l'apprentissage en groupe. Il faut s'assurer que les groupes sont créés avec des gens possédant des compétences complémentaires et que chaque groupe définit des objectifs réalistes.

1.2.4.5 Partage de la connaissance

Les questions à se poser, lorsqu'on veut rendre la connaissance disponible et utilisable à travers l'organisation, sont: Qui doit savoir quoi, à quel niveau de détail, et comment l'organisation doit soutenir ces processus de distribution de la connaissance? Tout le monde n'a pas besoin de savoir tout! Du point de vue de l'infrastructure technique

de distribution, les technologies réseaux permettent de mettre en rapport des experts par le biais d'une plate-forme électronique (exemple : groupware, etc.).

Un bon partage et une bonne distribution des connaissances ne génèrent pas seulement des gains de temps et de qualité mais également un accroissement direct de la satisfaction du client. Sur cette base, des réactions et des décisions rapides peuvent être prises.

1.2.4.6 Utilisation de la connaissance

L'utilisation de la connaissance signifie le déploiement productif de la connaissance organisationnelle dans le processus de production. En effet, une identification ainsi qu'une utilisation réalisées avec succès ne garantissent pas encore son adoption de la part des individus. Il faudra donc particulièrement bien veiller à communiquer aux personnes l'avantage qu'elles peuvent retirer d'une orientation vers la gestion des connaissances.

1.2.4.7 Rétention de la connaissance

Une fois que la connaissance a été acquise ou développée, encore faut il pouvoir la garder au sein de l'entreprise. De nombreuses entreprises se plaignent en effet qu'elles ont perdu une partie de leur mémoire d'entreprise lors d'un processus de réorganisation. Pour éviter ce genre de lacune, on doit établir des processus performants de sélection de l'information pertinente, de stockage et ensuite d'incorporer ces données, ces informations dans la base de connaissance de l'entreprise **[11]**.

1.2.4.8 Evaluation de la connaissance

Ce dernier point représente le plus grand challenge dans le domaine de la gestion des connaissances. Les gestionnaires de connaissance ne possèdent aucun outil pour mesurer et évaluer la connaissance. Ils doivent se battre contre des problèmes difficilement modélisables et quantifiables **[12]**.

1.3 La Gestion des Connaissances Médicales

Dans le domaine médical, la gestion des connaissances est d'abord une méthode imposant la formalisation de la connaissance et permettant le partage et la diffusion des ces connaissances **[13].** Plusieurs bénéfices peuvent être espérés de l'application de ces principes:

- Sélectionner les données les plus pertinentes parmi la masse des informations disponibles pour avoir la connaissance.

- Partager et faciliter l'accès à la connaissance.
- Comprendre les mécanismes d'interprétation et de raisonnement médical,
- Exploiter ces connaissances pour la prise de décision.

Ainsi, la gestion des connaissances médicales s'attache à développer et à évaluer des méthodes et des systèmes pour l'acquisition**,** le traitement et l'interprétation des connaissances extraites à partir des données «patient» avec l'aide des connaissances issues de la recherche scientifique **[14]**. Pour réaliser ces objectifs, le domaine médical utilise des méthodes scientifiques qui héritent de l'informatique, l'intelligence artificielle, des mathématiques et des sciences de gestion **[15], [16].**

1.3.1 La définition de la gestion des connaissances médicales

Nous allons proposer deux définitions, la première de Audrey Baney **[17]** et la seconde rapportée par M.Stefanelli **[18] :**

1- La gestion des connaissances médicales vise à :

(1) Rassembler le savoir et le savoir-faire des praticiens sur des supports facilement accessibles;

(2) Faciliter leur transmission en temps réel à l'intérieur de l'établissement de soin et en différé à nos successeurs;

(3) Garder la trace de certaines activités ou actions sur lesquelles on peut devoir rendre des comptes à l'avenir. Chaque direction est chargée de définir, dans son domaine de responsabilité, ce qui doit être écrit et conservé et d'organiser cette conservation.

2- La gestion des connaissances médicales vise à améliorer les performances de l'organisation en permettant aux individus de l'établissement de soins (médecins, infirmiers, paramédicaux, etc.) de capturer, partager et appliquer des connaissances collectives pour prendre des décisions optimales en temps réel.

- Ces définitions proposent deux points de vue, en termes de moyens et en termes de buts, mais elles sont cohérentes et nous pouvons les instancier dans un milieu hospitalier **[19].**

La médecine est confrontée à des exigences de qualité et d'optimalité de soins qui obligent à gérer les connaissances médicales. Les outils proposés sont **[20]:**

1- la « traçabilité » des actes médicaux (exigence des tutelles);

2- la protocolisation des soins (exigences médicales);

3- le respect de références médicales opposable (RMO exigence médicale des tutelles).

1.3.2 Données et connaissances médicales

1.3.2.1 Caractéristiques des données médicales

Les données provenant d'un patient sont les données les plus difficiles à analyser. Plusieurs problèmes se posent **[21]:**

- L'anonymat des données;
- Les caractéristiques intrinsèques des données médicales;
- Haute dimensionnalité;
- Incertitude;
- Imprécision;
- Ambiguïté;

1.3.2.2 Caractéristiques des connaissances médicales

La difficulté de la formalisation de la connaissance médicale permet de qualifier la médecine comme une science « non exacte ». Elle est engendrée par l'incertitude des connaissances, ce qui implique un raisonnement difficilement reproductible **[22].**

* Le taux d'erreurs de diagnostic n'est pas négligeable.

* L'incertitude en médecine: l'interprétation de l'état du patient est sujette à l'appréciation du praticien.

* Des comportements du patient difficilement formalisables;

1.3.3 Les objets de base de la gestion des connaissances médicales

1.3.3.1 Le dossier médical

Le « lieu » de convergence des flux d'informations et de connaissances qui circulent dans l'hôpital au sujet d'un patient est le dossier médical, c'est le point de passage obligé par sa fonction d'archivage de toutes les données cliniques d'un patient, le contenant de toutes les informations nominatives au sujet d'un patient **[23].**

La capacité de lire un dossier médical est un savoir faire qui se transmet entre apprenants et apprentis. Le dossier médical hypertextuel peut être considéré comme un objet focal de la gestion des connaissances médicales.

1.3.3.2 Les Références Médicales Opposables (RMO)

RMO sont l'un des outils conventionnels de maîtrise médicalisée des connaissances. Elles visent à éviter le superflu, voire le dangereux dans la prise en charge de certaines maladies ou la prescription de certains examens ou médicaments. Parce qu'elles prennent en compte l'ensemble des connaissances médicales et leur évolution, les références médicales sont appelées à être réactualisées régulièrement et complétées. En tout état de cause, les médecins demeurent entièrement libres et responsables de leurs prescriptions. Exemple de référence médicale : Dans la plupart des cas il est inutile de répéter le dépistage du Cholestérol, s'il s'est révélé normal la première fois, avant 3 ou 5 ans selon le cas et en l'absence de pathologie, de traitement ou d'augmentation de poids **[24].**

1.3.3.3 Evidence-Based Medicine (EBM)

Appelée la médecine fondée sur des preuves ou médecine factuelle désignait, c'est une stratégie d'apprentissage des connaissances cliniques, elle fait maintenant partie intégrante de la pratique médicale et consiste à baser sur les décisions cliniques, non seulement sur les connaissances théoriques, le jugement et l'expérience qui sont les principales composantes de la médecine traditionnelle, mais également sur des « preuves » scientifiques, tout en tenant compte des préférences des patients. Par « preuves », on entend les connaissances qui sont déduites de recherches cliniques systématiques, réalisées principalement dans le domaine du prévision, du diagnostic et du traitement des maladies et qui se basent sur des résultats valides et applicables dans la pratique médicale courante **[25].**

1.3.3.4 Guide de Bonne Pratique (GBP)

Contient des protocoles de soins qui se développant dans de nombreuses disciplines sous la forme de guides à utiliser par les praticiens, ce guide nécessite de formuler clairement la question clinique, de rechercher les articles pertinents au sujet, d'évaluer la validité et l'intérêt des résultats et, enfin, d'intégrer ces preuves dans la pratique médicale courante afin de répondre à la question posée au départ **[22].**

En aucun cas, ces « preuves » ne peuvent remplacer le jugement et l'expérience du médecin, ce qui explique que la médecine factuelle complète la pratique médicale traditionnelle mais ne la remplace pas **[URL1].**

1.3.3.5 Nomenclature Générale des Actes Professionnels

Ce sont des Conventions nationales organisant les rapports entre les Médecins libéraux et l'Assurance maladie. Par exemple : la Majoration transitoire (MTC) pour des actes thérapeutiques non répétitifs réalisés en équipe sur un plateau technique lourd effectués dans le cadre de la chirurgie générale, digestive, orthopédique-traumatologique, chirurgie vasculaire, chirurgie cardio-thoracique et vasculaire, neurochirurgie, chirurgie urologique, chirurgie pédiatrique et chirurgie gynécologique, chirurgie cervicale et maxillo-facial nécessite 12,5 % de la cotation des actes, soit 2000 DA multiplié par le coefficient de cotation **[25].**

- Les acteurs de l'établissement de soins (médecins, infirmières, paramédicaux, biologiste, radiologiste…etc.) doivent pouvoir partager et appliquer des connaissances collectives et donc collectivement validées, en particulier les RMO ou les protocoles de soins se développant dans de nombreuses disciplines sous la forme de guides de bonne pratique (GBP), pour la partie de la médecine la plus protocolaire. Ils doivent aussi acquérir des savoirs et des savoir-faire qui leur permettront d'utiliser leur relation clinique avec le patient pour « instancier », de la meilleure façon qu'il soit, des connaissances générales au contexte du patient.

- Pour cela, il faut se donner les moyens de rassembler le savoir et le savoir-faire médical, les protocoles, les GBP, les références bibliographiques de la médecine factuelle sur des supports facilement accessibles, faciliter leurs transmission en temps réel à l'intérieur de l'établissement et en différé à toute personne habilitée à les demander (patient, médecin, infirmier, autre. . .). Enfin, Il est nécessaire de garder les traces de toutes les données médicales d'un patient durant toute sa vie.

1.3.4 Les types de connaissances médicales

Il est nécessaire de citer les types de connaissances mis en jeu dans le processus de gestion de connaissances médicales. Elles peuvent être ainsi partagées en catégories liées à la spécificité de la médecine et au secret médical **[26]:**

1.3.4.1 Les connaissances nominatives : dans cette catégorie on trouve:

- des connaissances nominatives administratives (le dossier administratif);
- des connaissances nominatives médicales (le dossier médical enregistrement de tous les événements ayant trait au patient).
- et le « dossier infirmier », permettant de piloter les traitements.

1.3.4.2 Les connaissances non nominatives : dans cette catégorie, on trouve les sources et les résultats d'études épidémiologiques, des conclusions bibliographiques sur les pratiques médicales (La médecine factuelle), des connaissances sur les conduites à tenir (les protocoles thérapeutiques, les GBP, les RMO).

1.3.5 Nature des connaissances à gérer

Les connaissances médicales mises en œuvre apparaissent statiques puis de plus en plus dynamiques. Le processus de gestion des connaissances médicales suit la même caractéristique, et les trois vues de la gestion des connaissances décrites sont adéquates à décrire ces vues de l'organisation à différentes distances **[26]:**

1.3.5.1 Les connaissances gérées par l'administration

Nous sommes ici dans le point de vue des connaissances de gestion vues par l'administration, à travers l'outil de gestion de la structure de santé. Dans ce cas, le modèle est très statique, où l'innovation n'apparaît et se continue que quand les incitations administratives le veulent bien. La dynamique des connaissances n'a pas besoin d'être modélisées, les innovations étant rares.

1.3.5.2 Les connaissances gérées par l'organisation médicale

Nous sommes ici au niveau de l'organisation de l'unité de soins, avec ses procédures liées à l'échange d'informations et de connaissances. C'est dans ce contexte que, par ses réflexions et outils, la gestion des connaissances peut apporter beaucoup : que ce soit dans son expérience de modélisation des objets du domaine, ou celle de la tâche ou encore dans

le fait de savoir recueillir la connaissance. Tout cela, nourri de nos réflexions sur les formes dans les organisations.

1.3.5.3 Les connaissances gérées par l'innovation médicale

Nous sommes ici dans le point de vue de l'innovation médicale étant donné qu'elle va mobiliser de nouveaux savoirs, de nouvelles connaissances et donc de nouvelles manières d'agir. Un nouvel examen ou un nouveau type de soins modifient les connaissances que la médecine peut avoir sur la personne soignée. Plus, ce nouvel examen, par exemple un nouveau geste, vient des connaissances des praticiens.

- Les trois points de vue de la gestion des connaissances ayant été ajustés dans le contexte médical, Notre étude porte sur le deuxième, celui où les besoins d'échange de connaissances sont évidents et où la gestion des connaissances peut et a commencé d'agir.

1.3.6 La transformation et le cycle de vie des connaissances médicales

À partir du dossier médical et de l'ensemble des connaissances que nous avons listées précédemment, on peut proposer de voir le flux des connaissances dans l'établissement de santé comme une boucle de transformations au sein de l'unité de soin avec des connexions vers l'extérieur comme le codage ou des transferts d'extraits de dossiers médicaux selon les modalités. On voit ainsi la figure 1.3, le dossier patient au départ d'une boucle qui va vers les données cliniques, (et cela demande de les extraire via un thésaurus appuyé sur une ontologie. *i.e.* un *thésaurus sémantique* et du traitement du langage médical) puis vers les résultats des études épidémiologiques (des données structurées). Ces études permettent de générer des RMO ou des GBP (que nous considérons comme du texte semi-structuré) ou de permettre au médecin de pratiquer la médecine factuelle (EBM). Enfin, le dossier médical permet d'effectuer le codage des patients et de communiquer vers l'extérieur **[27].**

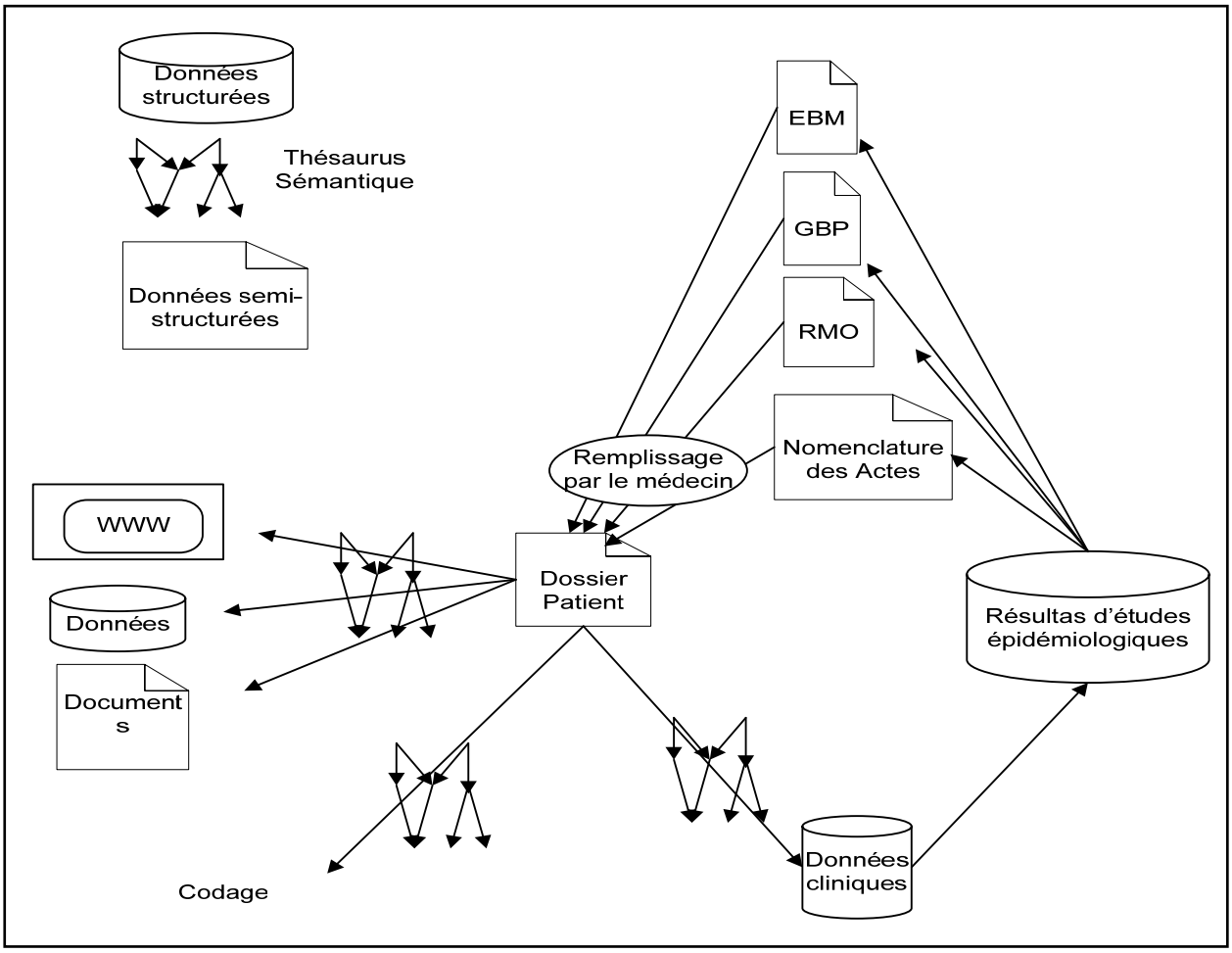

Figure 1.3 : Les transformations et le cycle de vie des connaissances médicales

1.3.7 Hétérogénéité des données médicales

La disponibilité croissante des sources de données variées et dispersées contenant des informations cruciales pour la prise de décision au sein des établissements de santé pose le problème de leur accès de façon efficace. Les systèmes d'information sont conçus pour fonctionner de façon autonome et non pour un besoin d'intégration. Dans la plupart des cas, les sources ont été développées indépendamment et sont par conséquent hétérogènes. Cependant, les établissements ont de nos jours le souci de faire inter-opérer ces différentes sources afin d'avoir une vision globale de leur système d'information. Pour réaliser cette interopérabilité, un certain nombre d'adaptations et de fonctionnalités sont nécessaires **[28].**

Un système d'information hospitalier est homogène si le logiciel qui gère les données est le même sur tous les sites et les services médicaux, si les données ont un même format et une même structure (même modèle de données) et appartiennent à un même univers de discours. A l'opposé, un système hospitalier est hétérogène s'il utilise des

langages de programmation ou d'interrogation ou des modèles ou des systèmes de gestion de données différents.

Les données médicales d'un patient sont hétérogènes, variées et dispersées, elles peuvent être stocké dans différents formats tels que : un fichier texte simple, une ordonnance, un compte rendu des résultats médicaux, un fichier image (résultant de documents ou d'images numériques (DCOM) d'un radio scanner ou IRM), un fichier commentaire vocal stocké dans un dictaphone, un fichier signal : Electro-Cardio-Gramme (ECG)…etc.

La figure 1.4 représente les différents formats de stockage de données médicales:

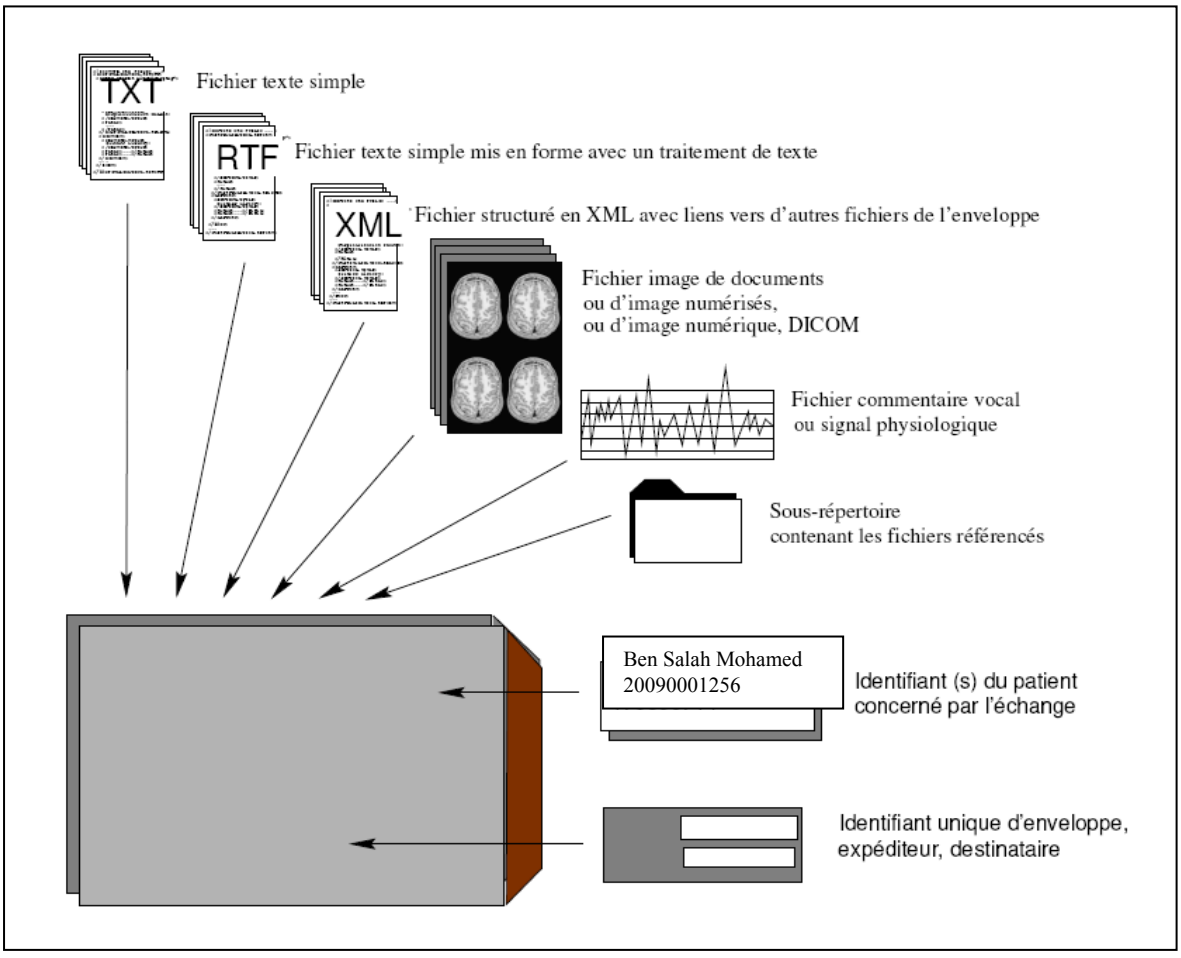

Figure 1.4 : Hétérogénéité des données médicales.

1.4 Conclusion

Dans ce chapitre nous avons parlé sur la gestion des connaissances, les formes des connaissances, comment créer les connaissances, et les processus de la gestion des connaissances, ensuite nous avons implémenter la gestion des connaissances dans notre domaine d'étude, c'est le domaine médical, ou nous avons cité les caractéristiques de la connaissance médicale, les objets de base de la gestion des connaissances médicales, son cycle de vie, et enfin nous avons parlé sur le problème d'hétérogénéité des données médicales.

Dans le chapitre suivant, et avant de se situer dans notre objectif d'étude, nous présentons une plate forme d'intégration des données hétérogènes pour résoudre le problème cité avant. Nous situons également la place des entrepôts de données dans le processus d'intégration comme une solution de notre problème et nous discutons les caractéristiques des entrepôts et le processus ETL (Extraction, Transformation, Loading) pour l'intégration des données.

CHAPITRE 2

LES ENTREPOTS DE DONNEES

2.1 Introduction

La diversité des sources d'information distribuées et leur hétérogénéité sont une des principales difficultés rencontrées par les utilisateurs aujourd'hui, notamment dans le domaine médical. Cette hétérogénéité peut provenir du format ou de la structure des sources. On distingue trois types de conflits de données: syntaxiques, schématiques et sémantiques.

- Les conflits syntaxiques sont relatifs aux différences conceptuelles des modèles utilisés pour la représentation de l'information (relationnel, orienté objet, etc.).
- Les conflits schématiques apparaissent dans les différentes manières de structurer et de classifier les données d'une source à l'autre pour la représentation d'une même information (conflits de types, de noms, etc.).
- Les conflits sémantiques ou cognitifs sont dus au fait que l'information est interprétée différemment en fonction du domaine d'application (conflit de valeurs, de taxonomie, de cognition).

L'interopérabilité cherche à résoudre tous ces types de conflits afin de combiner ou fusionner correctement les informations issues de ces sources distantes. Plusieurs approches ont été proposées dans ce cadre. Parmi eux l'entrepôt des données.

2.2 L'approche entrepôt de données (data warehouse)

Cette approche consiste à extraire, à l'avance, les données pertinentes pour l'usage des utilisateurs, à les filtrer, les transformer et les stocker. "Un entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision**" [29]**. Il offre aux utilisateurs (décideurs) un accès rapide aux données et informations essentielles afin d'optimiser la prise de décisions. Cet entrepôt peut être assimilé à un ensemble de vues matérialisées qui

suppose une certaine anticipation des besoins des utilisateurs. Les requêtes sont traitées non pas au niveau des sources d'information mais au niveau de l'entrepôt de données.

La conception d'un entrepôt de données pose plusieurs problèmes: d'abord la localisation des sources d'informations pertinentes, ensuite l'intégration des données qui demandent la connaissance des systèmes sources pour résoudre les conflits, et enfin l'extensibilité vers de nouvelles sources.

Une méthodologie de mise en œuvre d'entrepôts de données est basée sur trois éléments essentiels **[30]**: une base de métadonnées constituant l'interface entre l'équipe informatique et la communauté des utilisateurs, un ensemble de transformations appliquées aux données sources et la définition d'outils logiciels permettant de réaliser les traductions des données depuis les formats et modèles sources vers les formats et modèles cibles.

2.3 Les caractéristiques de l'entrepôt de données

Nous détaillons ces caractéristiques **[31], [32]** :

- **Orientées sujet:** les données des entrepôts sont organisées par sujet plutôt que par application. Cette orientation sujet va également permettre de développer son système décisionnel via une approche par itérations successives, sujet après sujet. L'intégration dans une structure unique est indispensable car les informations communes à plusieurs sujets ne doivent pas être dupliquées. Dans la pratique, une structure supplémentaire appelée Data Mart (magasin de données) peut être crée pour supporter l'orientation sujet.
- **Intégrées:** les données provenant des différentes sources doivent être intégrées, avant leur stockage dans l'entrepôt de données. Les données doivent être mises en forme et unifiées afin d'avoir un état cohérent.
- **Historisées:** la prise en compte de l'évolution des données est essentielle pour la prise de décision qui, par exemple, utilise des techniques de prédiction en s'appuyant sur les évolutions passées pour prévoir les évolutions futures.
- **Non volatiles:** à la différence des données opérationnelles, celles de l'entrepôt sont permanentes et ne peuvent pas être modifiées. Le rafraîchissement de l'entrepôt, consiste à ajouter de nouvelles données, sans modifier ou perdre celles qui existent. La non volatilité des données est en quelque sorte une conséquence de l'historisation. Une même requête effectuée à quelques mois d'intervalle en

précisant la date de référence de l'information recherchée donnera le même résultat.

2.4 La structure de l'entrepôt de données

Un entrepôt de données se structure en quatre classes de données, organisées selon un axe historique et un axe synthétique **[31].**

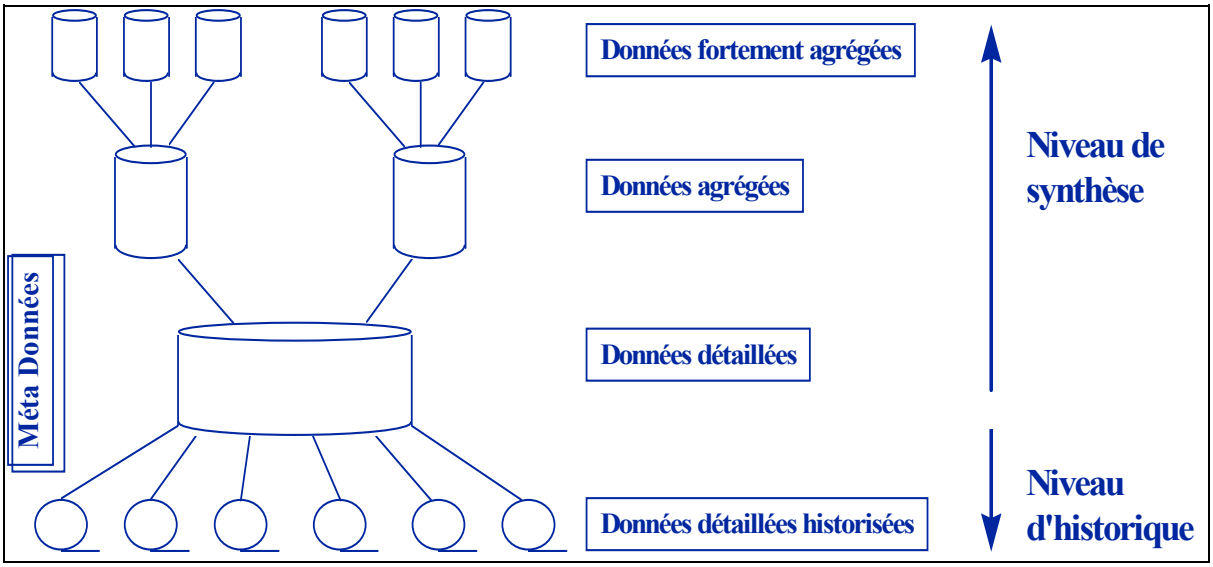

Figure 2.1: Structure d'un entrepôt de données.

- **Les méta-données**

Elles regroupent l'ensemble des informations concernant l'entrepôt de données et les processus associés. Elles constituent une véritable aide permettant de connaître l'information contenue dans l'entrepôt de données. Par exemple la sémantique (signification), l'origine, les règle d'agrégation, le format de stockage.

- **Les données agrégées**

Elles correspondent à des éléments d'analyse représentatifs des besoins utilisateurs, elles constituent déjà un résultat d'analyse et une synthèse de l'information contenue dans le système décisionnel, et doivent être facilement accessibles et compréhensibles. La facilité d'accès est apportée par des structures multidimensionnelles qui permettent aux utilisateurs de naviguer dans les données suivant une logique intuitive, avec des performances optimales.

- **Les données détaillées**

Elles reflètent les événements les plus récents. Les intégrations régulières des données issues des systèmes de production vont habituellement être réalisées à ce niveau.

- **Les données historisées**

Un des objectifs de l'entrepôt de données est de conserver en ligne les données historisées. Chaque nouvelle insertion de données provenant du système de production ne détruit pas les anciennes valeurs, mais crée une nouvelle occurrence de la donnée.

2.5 L'architecture de l'entrepôt de données

Pour implémenter un entrepôt de données, trois types d'architectures sont possibles **[33]** :

2.5.1 L'architecture réelle

Elle est généralement retenue pour les systèmes décisionnels. Le stockage des données est réalisé dans un SGBD séparé du système de production. Le SGBD est alimenté par des extractions périodiques. Avant le chargement, les données subissent d'importants processus d'intégration, de nettoyage, de transformation **[34]**.

L'avantage est de disposer de données préparées pour les besoins de la décision et répondant aux objectifs de l'entrepôt de données. Les inconvénients sont le coût de stockage supplémentaire et le manque d'accès en temps réel.

Nous avons adapté dans notre travail l'architecture réelle.

2.5.2 L'architecture virtuelle

Dans cette architecture les données résident dans le système de production. Elles sont rendues visibles par des produits middleware ou par des passerelles. Il en résulte deux avantages : pas de coût de stockage supplémentaire et l'accès se fait en temps réel. L'inconvénient est que les données ne sont pas préparées.

2.5.3 L'architecture hybride

C'est une combinaison de l'architecture réelle et de l'architecture virtuelle. Elle est rarement utilisée. L'objectif est d'implémenter physiquement les niveaux agrégés afin d'en faciliter l'accès et de garder le niveau de détail dans le système de production en y donnant l'accès par le biais de middleware ou de passerelle.

2.6 Le Data Mart

Le Data Mart est une base de données moins coûteuse que l'entrepôt de données, et plus légère puisqu'il est destiné à quelques utilisateurs d'un département.

C'est une petite structure très ciblée et pilotée par les besoins utilisateurs. Il a la même vocation que l'entrepôt de données (fournir une architecture décisionnelle), mais vise une problématique précise avec un nombre d'utilisateurs plus restreint (orienté sujet). En général, c'est une petite base de données (SQL ou multidimensionnelle) avec quelques outils, et alimentée par un nombre assez restreint de sources de données. Son coût est moins que le coût de l'entrepôt de données.

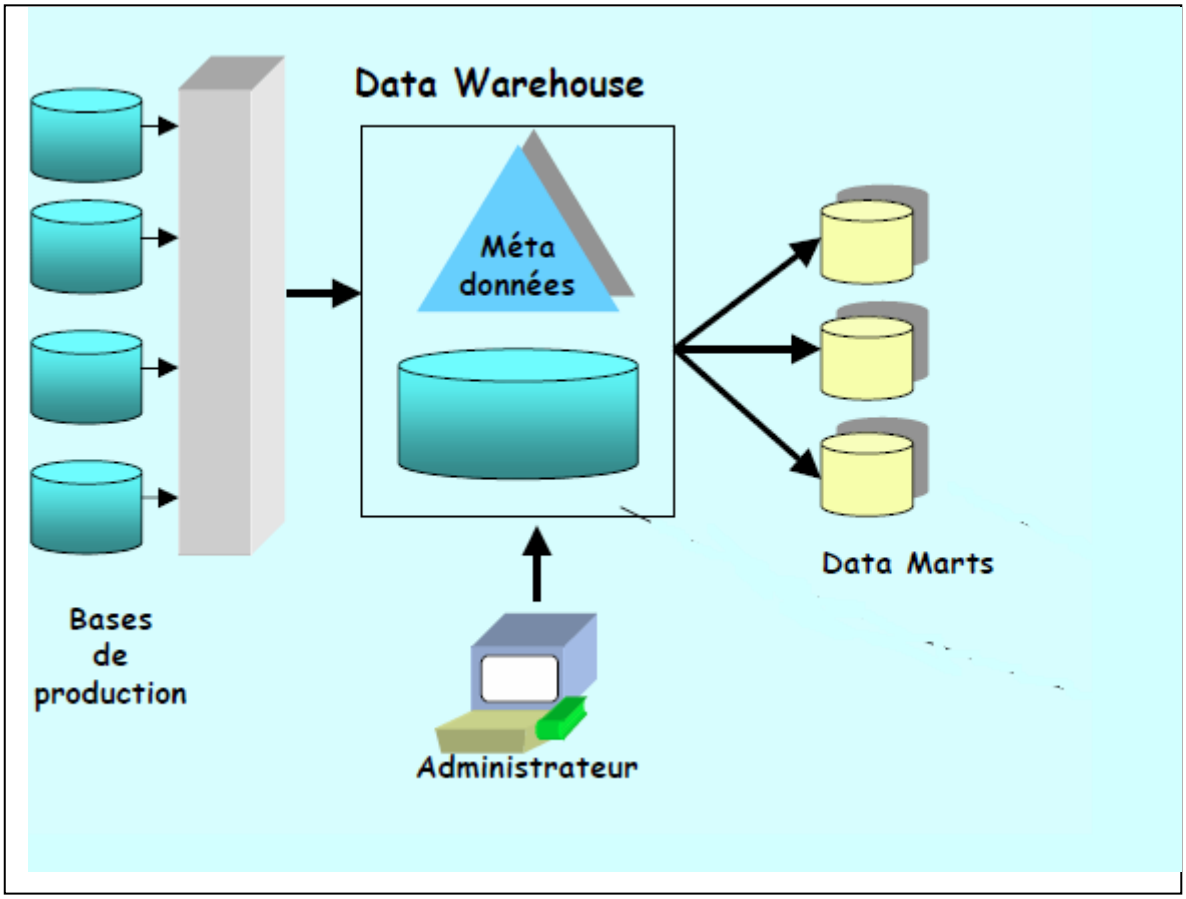

Figure 2.2 : Construction du Data Mart

2.7 Les étapes de construction d'un entrepôt de données

Les données intégrées de l'entrepôt ont une grande valeur pour l'entreprise mais leur intégration n'est pas une tache simple, car le passage des données sources vers l'entrepôt nécessite un processus complexe d'extraction, de transformation et de chargement, de plus les changements dans les sources doivent être régulièrement propagés

vers l'entrepôt pour le mettre à jour. La figure suivante illustre le processus ETL (Extraction, Transformation et Chargement) pour l'intégration des données.

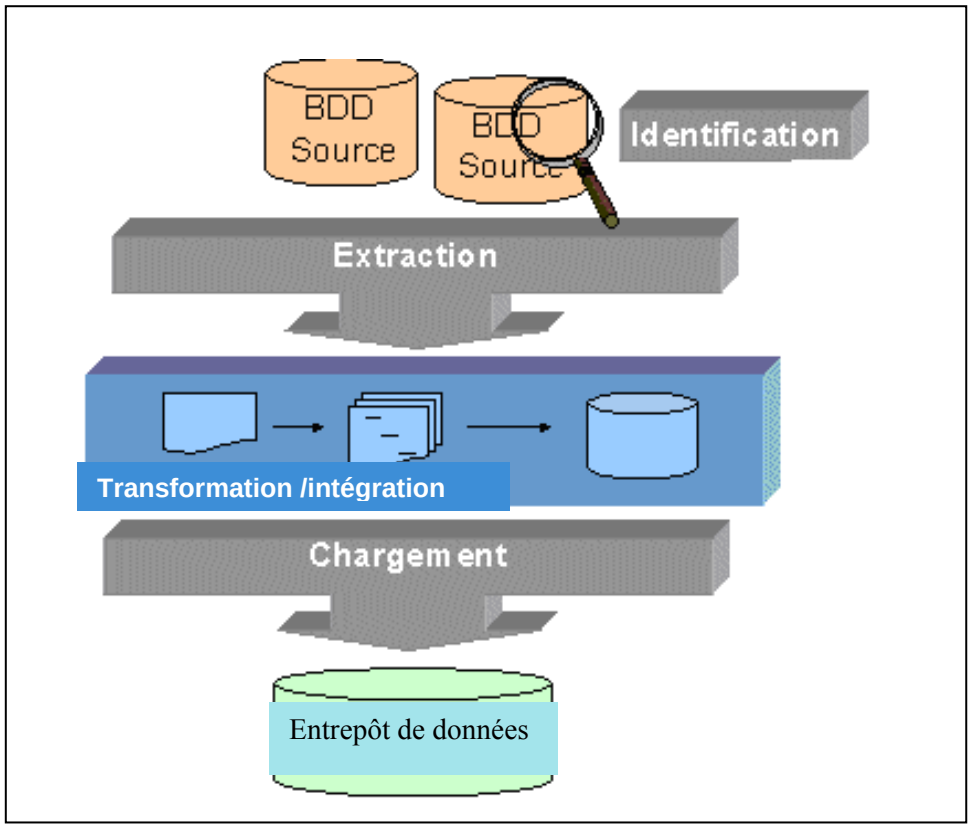

Figure 2.3 : Le processus ETL pour l'intégration des données

2.7.1 L'extraction des données

Cette phase collecte les données utiles des sources de données opérationnelles à partir des différentes sources hétérogènes, elles peuvent être structurées, semi-structurées ou non structurées. Elles sont :

-Hétérogènes : différents SGBD et différentes méthodes d'accès;

-Diffusées : différents environnements matériels et différents réseaux interconnectés ou non;

-Complexes : différents modèles logiques et physiques principalement orientés vers les traitements transactionnels.

2.7.2 La transformation et l'intégration des données

L'objectif étant de ne perdre aucune donnée provenant des sources d'origine et également d'aboutir à des données représentées dans un même format homogène **[35]**.

Au niveau de transformation les sources d'informations ainsi que les modules d'interfaces locales sont regroupées. A chaque source on trouve un adaptateur /moniteur qui a deux fonctions :

1- Transformer le schéma source et ses instances en une représentation intermédiaire.

2- Détecter automatiquement les changements dans la source pour les propager vers l'intégrateur.

- L'*adaptateur* transforme les données à partir de la représentation locale vers le format objet. Ce module est activé pour l'alimentation d'un entrepôt de données nouvellement créé ou pour remonter les mises à jour au niveau intermédiaire ou de mise en œuvre.
- *Le moniteur* détecte automatiquement les changements de sa source et les propage vers l'intégrateur. Par rapport à la détection automatique de changement, il est possible de classifier les sources **[36]:**

- Sources coopératives : celles qui supportent des mécanismes de règles actives ou de notification **[37],** pour notifier les changements de manière automatique.

- Sources permettant des requêtes à la demande : elles autorisent le moniteur à poser des questions de manière périodique pour pouvoir détecter les changements pertinents.

L'intégration entraîne plusieurs activités de « nettoyage » liées à la transformation des données pour les rendre conformes au schéma de l'entrepôt et aux critères de qualités choisis **[38], [39].** Une problématique similaire est abordée dans le domaine des systèmes multibases de données, ou l'intégration des schémas est décomposée en trois activités :

- **Identification des entités ayant une sémantique similaire** : elle est compliquée à cause des problèmes d'hétérogénéité structurelle que présentent les sources de données médicales, une même entité peut être représentée de différentes manières dans plusieurs schémas (conflit structurel), donc des entités distinctes peuvent aussi être représentées de façon similaire.

- **Identification et résolution des éventuels conflits entre les entités** : une fois identifiées les entités similaires, le problème est alors de résoudre les conflits qui peuvent apparaître. Ces conflits comprennent des problèmes de domaine, d'hétérogénéité ou encore d'incohérence. Les problèmes de domaine sont résolus en utilisant des techniques ou il s'agit de trouver la valeur correcte d'un attribut dans une liste, par exemple de trouver la bonne adresse d'un patient, ou de corriger un nom mal écrit.
- **Intégration des entités** : la consolidation des entités vise à les unifier au sein d'un même schéma celui de l'entrepôt, la fusion relationnel avec des sémantiques similaires est possible grâce à un ensemble d'opérateurs qui visent à faire l'union, la jointure ou la généralisation de deux relations, portant à la fois sur les schémas et les données qu'elles stockent **[40]**.

La figure suivante illustre les composants logiciels utilisés pour la construction de l'entrepôt de données :

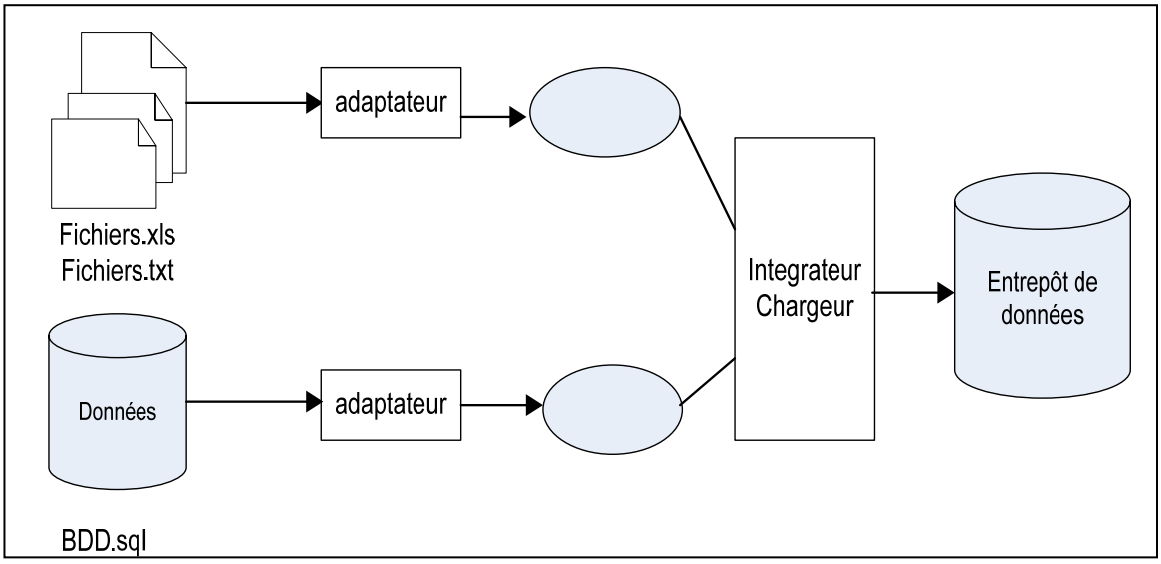

Figure 2.4 : Les composants utilisés pour la construction de l'entrepôt.

2.7.3 Le chargement des données

Une fois que l'intégrateur a fusionné les données en provenance des adaptateurs, les données intégrées peuvent alors être chargées dans l'entrepôt. Le chargement est fait soit lors de la création de l'entrepôt, soit lors du rafraîchissement. Le processus de détection de changements dans les sources et leur propagation vers l'entrepôt est connu sous le nom de rafraîchissement et il en existe deux types : reconstruction ou incrémental **[41].**

1- Dans le premier, il s'agit d'éliminer l'ancien contenu de l'entrepôt et de le remplir à nouveau périodiquement, avec une période qui dépend des besoins des utilisateurs et de la charge d'accès à l'entrepôt.

2- Dans le rafraîchissement incrémental, l'entrepôt est mis à jour en utilisant les changements dans les sources détectés par le moniteur et traités par l'intégrateur à chaque fois qu'une source change ou bien de manière périodique.

2.8 Interrogation de l'entrepôt

Dans cette section nous abordons la problématique d'interrogation d'entrepôt de données, les requêtes entraînent en général des opérations coûteuses d'agrégation sur de gros volumes de données, mettent en évidence la limitation de la technologie actuelle pour l'expression et le traitement des requêtes, nous discutons d'abord les aspects qui concernent l'expression et le traitement des requêtes dans les systèmes relationnels et multidimensionnels. Ensuite, nous expliquons le précalcul d'agrégats. Et, finalement nous abordons le traitement parallèle des requêtes.

2.8.1 Expression et le traitement des requêtes

Il est connu que SQL, le langage d'interrogation standard des SGBD relationnel n'est pas bien adapté à l'interrogation des entrepôts **[42].** En effet l'expression des requêtes d'analyse devient très complexe parce qu'il est souvent nécessaire d'utiliser de multiples jointures et unions. De plus les fonctions d'agrégat offertes par SQL, telles que COUNT, SUM et AVG sont souvent insuffisantes pour exprimer des analyses plus complexes.

Plusieurs extensions de SQL ont été proposées pour faciliter l'expression des requêtes d'analyse, jagadish et al **[43]** ont proposé une extension de SQL pour rendre plus facile l'écriture, la lecture et l'optimisation de requêtes intégrant généralement la notion de dimension et les agrégations nécessaires. L'opérateur CUBE a été proposé pour faciliter l'expression des requêtes d'agrégation entraînant plusieurs GROUP BY. Concéderons par exemple la requête suivante de la relation vente du schéma en étoile de la Figure 2.4 :

```
Select          Produit, Magasin, Temps, SUM(Quantite)
From            Ventes
Group by CUBE   Produit, Magasin, temps
```

Le résultat de cette requête est l'union GROUP BY pour tous les sous ensembles possible de {Produit, Magasin, Temps}. A partir de cette représentation, les opérations communes à plusieurs entre eux sont combinées et les calculs effectués sont utilisés pour en calculer d'autres.

2.8.2 Le précalcul d'agrégats

Comme nous l'avons déjà mentionné, les opérations typiques sur un entrepôt sont les agrégats. Pour calculer les résultats d'une requête il est possible de calculer par avance des agrégats et de les stocker. Ces données pourront alors être utilisées par le système lors du traitement d'autres requêtes **[44].**

Il existe deux approches principales pour le précalcul d'agrégats **[45]** : complet et sélectif :

- Le précalcul complet vise à matérialiser les agrégats correspondant à toutes les combinaisons de valeurs de toutes les dimensions.
- Le précalcul sélectif se base sur l'observation qu'il y a des agrégats qui peuvent servir à en calculer d'autres **[46].**

Dans un SGBD relationnel, les agrégats précalculés sont représentés par des vues matérialisées associées à une relation de base **[47].**

2.8.3 Traitement parallèle de requêtes

Le volume de données d'un entrepôt peut atteindre un ordre très important, et dans ce cas, les méthodes traditionnelles de stockage et de traitement de requêtes doivent être parallélisées. Les travaux portent sur l'application de techniques parallèles pour le traitement de requêtes dans les systèmes relationnels sont nombreux et, à l'heure actuelle, plusieurs SGBDR les utilisent. Il s'agit de tirer profit de la structure sous forme d'*étoile* des schémas relationnels pour répondre aux requêtes de types *jointure* en *étoile*, par exemple, considérons la requête du type jointure en étoile suivante :

```
Select      M.ville, SUM(v.quantite)
From        Ventes V, Magasin M, Produit P, Temps T
Where       M.magasin = V.magasin   AND
              P.produit = V.produit    AND
```

T.temps = V.temps

Group by M.ville

Le résultat de cette requête est la quantité totale de vente par ville, étant donné la grande taille de l'entrepôt, l'exécution d'une telle requête devient une opération coûteuse, la solution de ce problème propose un algorithme qui exploite le schéma en étoile en partitionnant les données entre plusieurs processeurs et en exécutant la jointure en parallèle.

2.9 Modélisation multidimensionnelle

La modélisation multidimensionnelle consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions. Les données sont organisées de manière à mettre en évidence le sujet (le fait) et les différentes perspectives de l'analyse (les dimensions). En partant de cette définition, nous remarquons les concepts de fait et de dimension. Le fait représente le sujet d'analyse. Il est composé d'un ensemble de mesures qui représentent les différentes valeurs de l'activité analysée. Par exemple, dans le fait Ventes (voire Figure 2.5), nous pouvons avoir la mesure "Quantité de produits vendus par magasin". Les mesures doivent être valorisées de manière continue **[48], [49].**

Pour arriver à construire un modèle approprié pour un entrepôt de données, nous pouvons choisir, soit un schéma relationnel (le schéma en étoile, en flocon de neige ou en constellation) ; soit un schéma multidimensionnel.

2.9.1 Schémas relationnels

Dans les schémas relationnels, nous trouvons trois types de schémas conçus pour les systèmes décisionnels :

2.9.1.1 Le schéma en étoile

Il se compose du fait central et de leurs dimensions. Dans ce schéma, il existe une relation pour les faits et plusieurs pour les différentes dimensions autour de la relation centrale. La relation de faits contient les différentes mesures et une clé étrangère pour faire référence à chacune de leurs dimensions **[50].**

La figure 2.5 montre le schéma en étoile en décrivant les ventes réalisées dans les différents magasins d'une entreprise au cours d'un jour. Dans ce cas, nous avons une étoile centrale avec une table de faits appelée « Ventes » et autour leurs diverses dimensions : « Temps », » Produit » et « Magasin ».

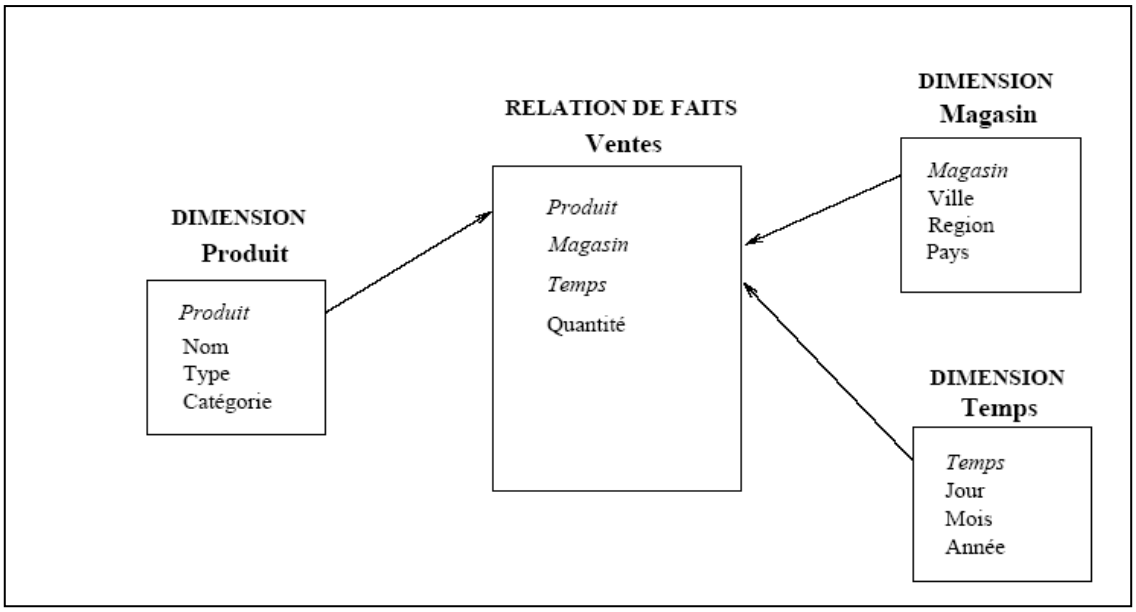

Figure 2.5 : Exemple de modélisation en étoile.

2.9.1.2 Le schéma en flocon de neige

Il dérive du schéma précédent avec une relation centrale et autour d'elle les différentes dimensions, qui sont éclatées ou décomposées en sous hiérarchies. L'avantage du schéma en flocon de neige est de formaliser une hiérarchie au sein d'une dimension **[51], [52],** ce qui peut faciliter l'analyse. Un autre avantage est représenté par la normalisation des dimensions, car nous réduisons leur taille, bien que cette normalisation rende plus complexe la lisibilité et la gestion dans ce type de schéma. En effet, ce type de schéma augmente le nombre de jointures à réaliser dans l'exécution d'une requête.

Les hiérarchies pour le schéma en flocon de neige de l'exemple de la figure 2.6 sont :

Dimension Temps = Jour, Mois, Année

Dimension Magasin = Commune, ville, Région, Pays

La figure 2.6 montre le schéma en flocon de neige avec les dimensions « Temps » et « Magasin » éclatées en sous hiérarchies.

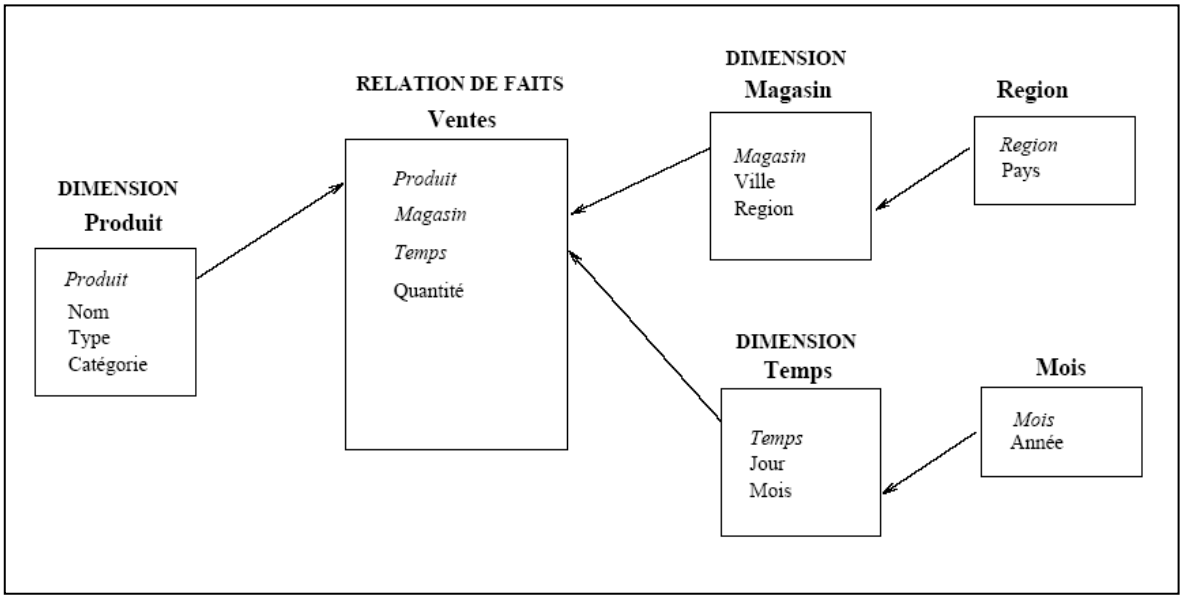

Figure 2.6 : Exemple de modélisation en flocon de neige.

2.9.1.3 Le schéma en constellation

Le schéma en constellation représente plusieurs relations de faits qui partagent des dimensions communes. Ces différentes relations de faits composent une famille qui partage les dimensions mais où chaque relation de faits a ses propres dimensions **[53].**

La figure 2.7 montre le schéma en constellation qui est composé de deux relations de faits. La première s'appelle « Ventes » et enregistre les quantités de produits qui ont été vendus dans les différents magasins pendant un certain temps. La deuxième relation « Achats » gère les différents produits achetés aux fournisseurs pendant un certain temps.

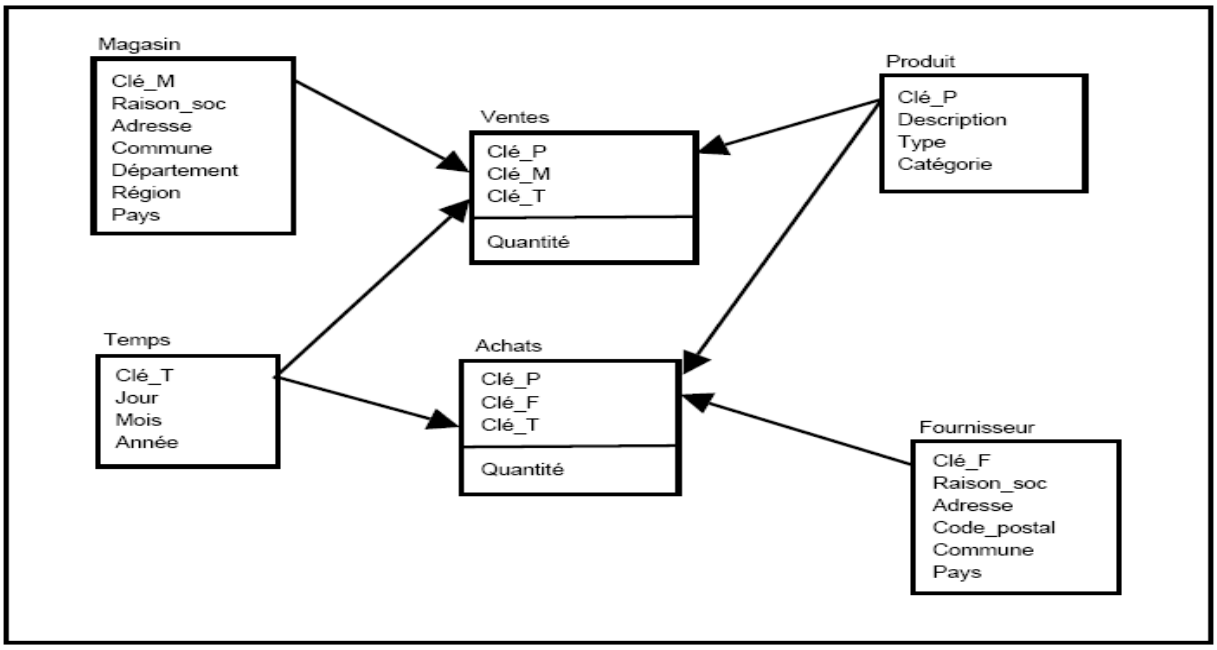

Figure 2.7 : Exemple de modélisation en constellation.

2.9.2 Schéma multidimensionnel (Cube)

Dans le modèle multidimensionnel, le concept central est le cube, lequel est constitué des éléments appelés cellules qui peuvent contenir une ou plusieurs mesures. La localisation de la cellule est faite à travers les axes, qui correspondent chacun à une dimension. La dimension est composée de membres qui représentent les différentes valeurs **[54], [55].**

En reprenant une partie du schéma en étoile de la figure 2.5, nous pouvons construire le schéma multidimensionnel suivant :

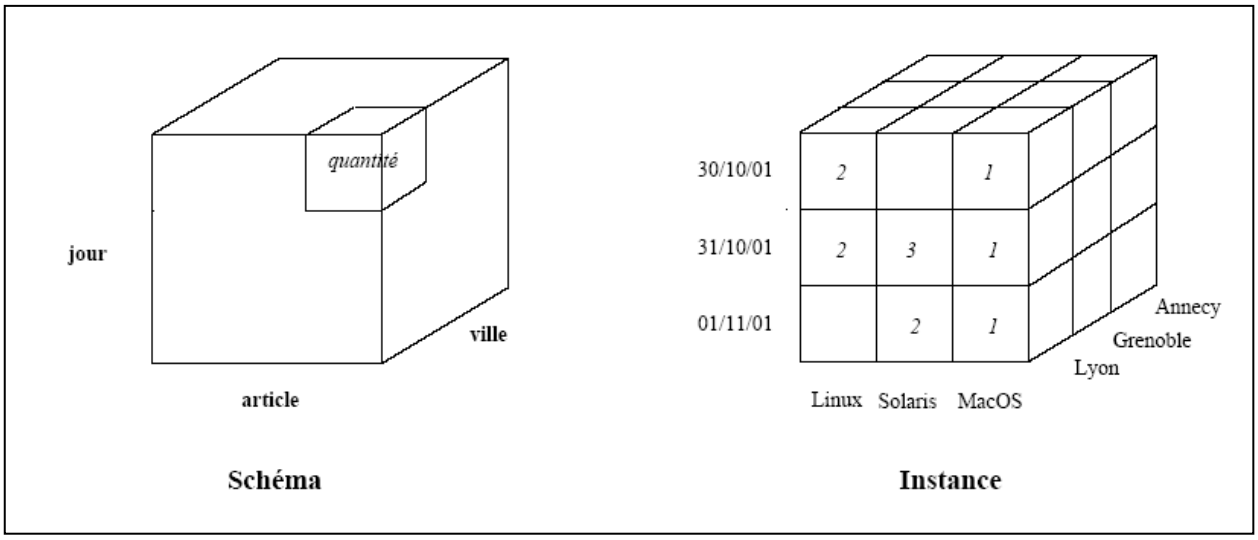

Figure 2.8 : Exemple de schéma multidimensionnel (CUBE).

2.10 Conclusion

Dans ce chapitre nous avons parlé du problème d'hétérogénéité des données due à la diversité des sources de données, notamment dans le domaine médical, puis nous avons proposé une solution (l'approche des entrepôts de données) pour l'intégration de ces données. Au long de ce chapitre, nous avons basé sur les caractéristiques des entrepôts de données, ses étapes de construction, après nous avons cité les méthodes d'interrogations et de modélisation des entrepôts de données.

Dans le chapitre qui suit, nous allons parler des techniques d'extraction de connaissances à partir de données stockées dans les entrepôts.

CHAPITRE 3

EXTRACTION DES CONNAISSANCES À PARTIR DES DONNEES

3.1 Introduction

Avec l'émergence de l'informatique, de nombreux Systèmes d'Information Hospitaliers (SIH) ont commencé à faire leur apparition. Ces SIH sont à l'origine du stockage d'énormes quantités de données médicales de tous types. Et comme pour les autres domaines, le data mining présente des outils performants pour l'extraction et la structuration des connaissances médicales, à partir des données collectées par les SIH. Les premiers systèmes intelligents ou d'aide au diagnostic en médecine étaient des systèmes de raisonnement à base de cas. Puis, au fur et à mesure du temps, et suivant les besoins, des méthodes de data mining ont été intégrées à ces systèmes.

Dans ce chapitre, nous présentons le processus d'extraction des connaissances à partir des données (ECD) et ses différentes étapes, puis nous insistons sur son noyau qui est le Data Mining, à travers ses différentes phases et ses méthodes. En dernier nous allons parler du data mining dans le domaine médical et nous allons citer quelques exemples.

3.2 L'Extraction des Connaissances à partir des Données

Fayyad et al. **[56]** a défini le processus ECD (Knowledge Discovery in database KDD) par : *"le processus ECD est un processus non trivial d'identification de modèles valides, nouveaux, potentiellement utiles, compréhensibles à partir d'une base de données"*. En analysant les termes utilisés dans cette définition, nous obtenons une définition plus pertinente :

– Modèle : c'est une représentation d'un phénomène réel, le plus souvent constitués d'objets mathématiques comme par exemple des ensembles de données, des tables, des matrices, des fonctions, des relations, des listes de règles, des systèmes d'équations, des arbres, des graphes, des hypergraphes, des réseaux, des opérateurs fonctionnels linéaires et non-linéaires, etc.

– Processus : c'est l'exécution d'un ensemble de plusieurs tâches qui peut être exécuté d'une façon itérative.

– Processus Non-trivial : c'est un processus qui exécute des tâches dans un ordre spécifique

– valide : un modèle valide est un modèle appliqué à une base de test doit renvoyer un degré de certitude.

Dans les processus ECD, il existe plusieurs entités qui interagissent lors des différentes phases du processus. Les entités humaines qui ressortent sont : l'utilisateur, l'expert en data mining, l'analyste de données, l'analyste du domaine d'application **[57].**

– L'utilisateur est la personne à qui est destiné le système ECD. Le plus souvent l'utilisateur apparaît comme étant un expert du domaine et qui possède certaines notions en data mining lui permettant d'interagir dans les différentes étapes du processus.

– L'expert en data mining est la personne qui met en place le processus ECD et le plus souvent, il est assisté par l'utilisateur et l'analyste de données **[57].**

– L'analyste de données s'occupe de sélectionner et de transformer les données pour les préparer au processus.

– L'analyste du domaine est un expert du domaine qui peut analyser les résultats afin de les valider.

3.3 Caractéristiques des systèmes d'extraction des connaissances à partir des données

– La maniabilité : un système maniable est un système qui permet à l'utilisateur d'interagir le plus aisément possible dans les différentes phases du processus

– La réutilisation : cela permet au système de s'adapter à la résolution de nouveaux problèmes ou d'être appliqué à différents domaines sans que la structure du système ne soit changée. Le système Health-Mining **[58]** présente cette caractéristique. Il peut se spécialiser pour le traitement d'un type de maladie et par la suite il peut facilement se réadapter pour le traitement d'une autre maladie.

– L'interprétabilité : elle concerne la modélisation des connaissances et les résultats obtenus d'une façon qu'ils puissent être assimilés et interprétés par les praticiens. Le

système développé par Wang et al **[59]** présente plusieurs types de visualisation cela dépend du niveau d'analyse à effectuer.

– Le multi-modale : il permet au système de faire appel à des méthodes concourantes dans la résolution d'un même problème. Ceci afin d'augmenter la fiabilité du système. Les systèmes mis en place par l'équipe Jiawei et al **[60],** utilisent trois méthodes qui sont de même type ; elles permettent à l'utilisateur de présenter différents résultats et c'est à l'utilisateur de choisir celui qui convient.

– La fusion : ceci permet au système de combiner des connaissances ou des données provenant des sources différentes (données : différentes bases médicales, bases réparties, connaissances : différentes expertises médicales). Le système PADMA **[61]** donne la possibilité aux praticiens d'exploiter des données distribuées.

3.4 Disciplines d'extraction des connaissances à partir des données

Grâce aux techniques d'extraction des connaissances, les bases de données volumineuses sont devenues des sources riches et fiables pour la génération et la validation de connaissances. L'extraction de connaissances à partir des données se situe à l'intersection de nombreuses discipline **[62]**, comme l'apprentissage automatique, la reconnaissance de formes, les bases de données, les statistiques, la représentation des connaissances, l'intelligence artificielle, les systèmes experts, etc.

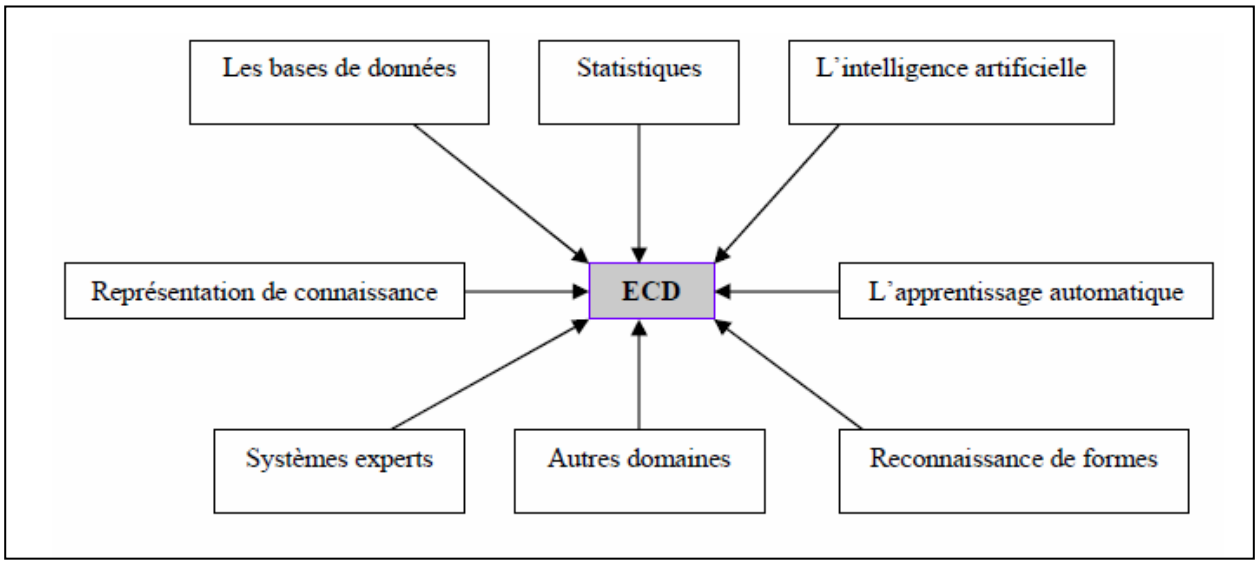

Figure 3.1 : L'Extraction des connaissances à partir des données à la confluence de nombreux domaines **[62].**

3.5 Processus d'extraction de connaissances à partir de données

Le processus d'ECD se décompose en plusieurs étapes. Il y a plusieurs définitions du processus d'ECD. Le nombre d'étapes défini dans un processus d'ECD varie de 4 à 9 suivant la définition donnée. Dans le tableau 3.1, défini par Eva Andràssyovà **[63],** nous présentons une liste représentative des processus d'ECD avec leurs différentes phases. Ce tableau est organisé de façon à retrouver sur une même ligne des tâches similaires.

Simoudis **[57]**	Mannila **[64]**	Fayyad et al.**[56]**	Brachman & Anand **[65]**
	Compréhension du domaine	Apprentissage du Domaine d'application	Découverte des actions à réaliser
Sélection des données		Création d'un ensemble de données cibles	Découverte des données
Transformation des données	Préparation des données	Nettoyage et prétraitement des données	Nettoyage des données
		Réduction et projection de données	
		Sélection des fonctions de data mining	Développement de modèles
Data Mining (Fouille de données)	Découverte de modèles	Sélection des algorithmes de data mining	Analyse des données
		Data minig	
Interprétation des résultats	post-traitement des modèles	Interprétation	Génération de rendement
	Utilisation des résultats	Exploitation des connaissances découvertes	

Tableau 3.1 : Liste des différentes étapes de processus d'ECD

Le modèle du processus d'extraction des connaissances que nous avons retenu se décompose en plusieurs phases : la sélection des données, la préparation des données, le data mining et enfin l'interprétation et l'évaluation des résultats (voir Figure 3.2). Le choix de ces quatre phases provient de la synthèse des processus ECD du tableau 3.1 :

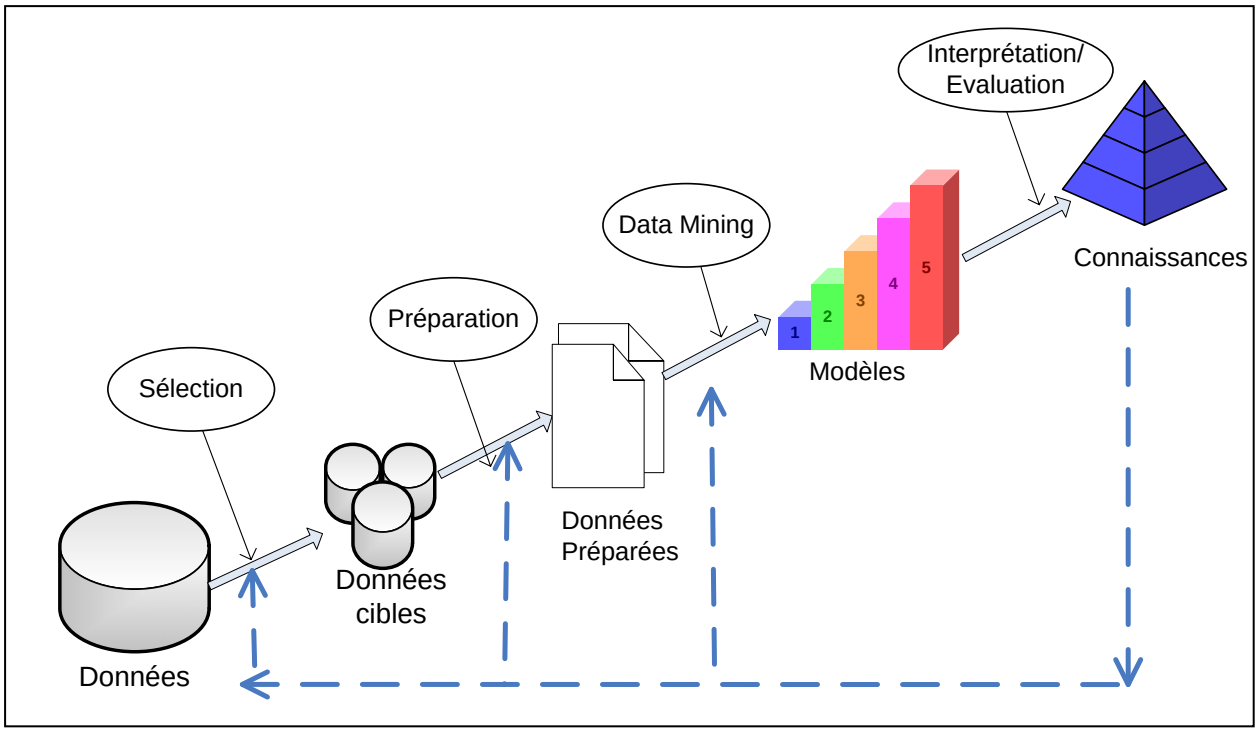

Figure 3.2 : Processus d'ECD

3.5.1 Sélection des données

Cette phase ne se limite pas à la seule sélection des données qui vont être exploitées par le système ECD **[60]**. Elle comprend également l'analyse du problème à résoudre **[57], [56],** ce qui permet d'en déduire le ou les types de données qui sont exploitées, ainsi que les méthodes qui pourraient être utilisées pour résoudre ce problème.

Un système ECD idéal est un système qui nécessite l'intervention d'aucune entité, c'est-à-dire un système automatisé qui va extraire de nouvelles connaissances à partir de grandes bases de données mises à sa disposition sans l'intervention de l'utilisateur. Actuellement, ce type de système présente de nombreux inconvénients. Le premier de ceux-ci est la perte de temps et de ressources nécessaires à l'exploitation de l'ensemble des données disponible au système.

3.5.2 Préparation des données

Les données à analyser par les méthodes de data mining sont parfois incomplètes, inconsistantes, erronées, incompatibles entre elles, inadaptées ou encombrantes **[60].** Ces types de données sont courants et se retrouvent régulièrement dans les bases de données et d'entrepôts de données.

Dans cette phase, plusieurs procédures sont nécessaires et chacune d'entre-elles a des tâches bien précises dans le traitement des données.

- La procédure de nettoyage des données permet de compléter les données manquantes et de régulariser les données erronées et inconsistantes.
- La procédure de transformation permet de modeler les données sous une forme exploitable par les méthodes de data mining.
- La procédure de réduction des données permet de réduire la taille des données tout en gardant leur intégrité.

3.5.3 Le data mining

C'est le coeur du processus d'ECD. Il s'agit à ce niveau de trouver des connaissances à partir des données **[66]**. Tout le travail consiste à appliquer des méthodes intelligentes dans le but d'extraire cette connaissance. Il est possible de définir la qualité d'un modèle en fonction de critères comme les performances obtenus, la fiabilité, la compréhensibilité, la rapidité de construction et d'utilisation et enfin l'évolutivité.

Tout le problème du data mining réside dans le choix de la méthode adéquate à un problème donné. Il est possible de combiner plusieurs méthodes pour essayer d'obtenir une solution optimale globale.

Nous ne détaillerons pas d'avantage le data mining dans ce paragraphe car elle fera l'objet d'une section complète (*cf*3.6).

3.5.4 Evaluation et présentation des résultats

La validation complète d'un système à base de connaissances consisterait à s'assurer de l'adéquation entre la connaissance modélisée dans la base de connaissances et la connaissance de l'expert. Deux types de techniques de validation peuvent être envisagés :

– La validation par l'examen des résultats obtenus à partir de la base de connaissances.

– La validation par l'étude de la cohérence de la base de connaissances.

Pour certains domaines d'application (le diagnostic médical, par exemple), le modèle présenté doit être compréhensible. Une première validation doit être effectuée par un expert qui juge la compréhensibilité du modèle. Cette validation peut être, éventuellement, accompagnée par une technique statistique.

Dans notre système les résultats du data mining seront exploités par le sous-système de raisonnement à base de cas.

3.6 Le data mining

Le data mining est le noyau du processus ECD il couvre plusieurs domaines, l'analyse de données, les bases de données, l'apprentissage, les statistiques, les systèmes à base de règles. Il dispose d'outils performants afin de structurer et d'extraire des connaissances : la classification, la segmentation, la recherche de règle d'association, …etc.

Les techniques de data mining ont été employées avec beaucoup de succès dans de nombreux secteurs d'application comme par exemple : la gestion de la relation client (GRC), la gestion des connaissances ou l'indexation de documents, etc. Comme tous les domaines ces techniques ont été intégrées à des applications médicales, et ont connu un franc succès du fait de leur variété et de leur capacité à traiter des données complexes. En outre, le data mining s'applique sur des données hétérogènes comme par exemple : des données structurées de type alphanumérique, semi-structurées et de type multimédia (image, son, vidéo).

3.6.1 Historique

L'analyse des données s'est développée et son intérêt grandissait en même temps que la taille des bases de données. Vers la fin des années 80, des chercheurs en base de données, tel que Rakesh Agrawal **[67]**, ont commencé à travailler sur l'exploitation du contenu des bases de données volumineuses comme par exemple celles des tickets de caisses de grandes surfaces, convaincus de pouvoir valoriser ces masses de données dormantes. Ils utilisaient l'expression "*database mining*" mais, celle-ci étant déjà déposée par une entreprise (Database mining workstation), ce fut "*data mining*" qui s'imposa. En mars 1989, Shapiro Piatetski **[68]** proposa le terme "*knowledge discovery*" à l'occasion d'un atelier sur la découverte des connaissances dans les bases de données. Actuellement,

les termes data mining et knowledge discovery in data bases (*KDD*, ou *ECD*) sont utilisés plus ou moins indifféremment **[69]**.

La communauté de "*data mining* " a initié sa première conférence en 1995 à la suite de nombreux atelier (workshops) sur le *KDD* entre 1989 et 1994. La première revue du domaine " *Data mining and knowledge discovery journal* " publiée par "Kluwers " a été lancée en 1997.

3.6.2 Définition

«Le data mining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données informatiques (souvent grandes), de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données» **[70]**.

La définition la plus communément admise de Data Mining est celle de **[72]** : *«Le Data mining est un processus non trivial qui consiste à identifier, dans des données, des schémas nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables».*

D'après H, Haddad **[71]**, *«le data mining est l'art d'extraire des informations (ou même des connaissances) à partir des données »*.

3.6.3 Principales tâches de data mining

On dispose de données structurées. Les objets sont représentés par des enregistrements (ou descriptions) qui sont constitués d'un ensemble de champs (ou attributs) prenant leurs valeurs dans un domaine. De nombreuses tâches peuvent être associées au Data Mining, parmi elles nous pouvons citer **[73]**:

3.6.3.1 La classification

" La classification consiste à examiner des caractéristiques d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini."

Dans le cadre informatique, les éléments sont représentés par un enregistrement et le résultat de la classification viendra alimenter un champ supplémentaire. Elle permet de créer des classes d'individus.

Les techniques les plus appropriées à la classification sont :

- les arbres de décision,
- le raisonnement basé sur la mémoire,
- l'analyse des liens.

3.6.3.2 L'estimation

Elle consiste à estimer la valeur d'un champ à partir des caractéristiques d'un objet. Le champ à estimer est un champ à valeurs continues. L'estimation peut être utilisée dans un but de classification. Il suffit d'attribuer une classe particulière pour un intervalle de valeurs du champ estimé.

Un des intérêts de l'estimation est de pouvoir ordonner les résultats pour ne retenir si on le désire, que les «n » meilleures valeurs. Cette technique sera souvent utilisée en marketing, combinée à d'autres, pour proposer des offres aux meilleurs clients potentiels. Enfin, il est facile de mesurer la position d'un élément dans sa classe si celui ci a été estimé, ce qui peut être particulièrement important pour les cas voisins.

La technique la plus appropriée à l'estimation est:

- les réseaux de neurones.

3.6.3.3 La prédiction

Cela consiste à estimer une valeur future. En général, les valeurs connues sont historisées. On cherche à prédire la valeur future d'un champ. Cette tâche est proche des précédentes. Les méthodes de classification et d'estimation peuvent être utilisées en prédiction.

Les techniques les plus appropriées à la prédiction sont :

- L'analyse du panier de la ménagère;
- Le raisonnement basé sur la mémoire;
- Les arbres de décision;
- les réseaux de neurones;

3.6.3.4 La segmentation

Consiste à former des groupes (clusters) homogènes à l'intérieur d'une population. Pour cette tâche, il n'y a pas de classe à expliquer ou de valeur à prédire définie *a priori*, il s'agit de créer des groupes homogènes dans la population (l'ensemble des enregistrements). Il appartient ensuite à un expert du domaine de déterminer l'intérêt et la signification des groupes ainsi constitués. Cette tâche est souvent effectuée avant les précédentes pour construire des groupes sur lesquels on applique des tâches de classification ou d'estimation.

La technique la plus appropriée à cette tache est :

- L'analyse des clusters.

3.6.3.5 Les règles d'association

Les règles d'association sont traditionnellement liées au secteur de la distribution car leur principale application est "*l'analyse du panier de la ménagère (market basket analysis*)" qui consiste en la recherche 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 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.

Dans notre travail, nous avons adopté cette technique pour la détection des règles d'associations de type « ensemble de signes cliniques => service ». La recherche des règles d'association est une méthode non supervisée car on ne dispose en entrée que de la description des signes cliniques.

Quelques définitions

Une règle d'association est une règle de la forme

Si <condition> Alors <résultats>.

Par exemple, une règle à trois faits sera de la forme :

Si X et Y alors Z.

Le choix d'une règle d'association se fait à partir des quantités numériques qui ont été définies.

Algorithme Apriori [67]

Cet algorithme a pour principe la recherche de règles intéressantes parmi toutes les règles. Pour ce faire, il s'effectue le calcul de support de tous les ensembles des enregistrements qui existent dans la base de données. Ainsi, un enregistrement ayant un Support supérieur à *minsup* (support minimum donné) est qualifié de fréquent et il est retenu dans la base des connaissances.

Par convention, on s'accorde à nommer *k_ensemble* un ensemble de *k* éléments. Pour générer un (*k* + 1) *ensemble* fréquent, il faut unir deux *k_ensembles* fréquents, qui diffèrent d'un seul élément. Ainsi, par union successive de *k_ensembles*, l'algorithme va établir la liste exhaustive des *1_ensembles* fréquents, puis des *2_ensembles* fréquents, puis des *3_ensembles* fréquents, et ainsi de suite jusqu'à ce qu'il n'y ait plus d'ensembles fréquents de cardinalité supérieure.

Détail de l'Algorithme Apriori

// Entrée : *minisup*, L_k ensembles d'items fréquents.

// Sortie : ensemble de règles d'associations.

$L1 = \{1\_ensembles\}$ faire

Tant que $L_k \neq \emptyset$;

$C_{k+1} = apriori - gen(L_k)$

// génère des $k + 1$ *ensembles* à partir de *k_ensembles*

Pour chaque transaction $t \in BD$(*BasedeDonnes*) faire

C_t = sous-ensemble de C_{k+1} inclus dans *t*.

Pour chaque $c \in C_t$ faire c.count++

$L_k = c \in C_k | c.count \geq minisup$ // retenir les candidats fréquents

Réponce = $L_1, L_2, \cdots, L_{k-1}$

1 La fréquence

A chaque règle, on associe une mesure de fréquence, c'est le nombre d'apparition d'une condition dans la base de données:

Fréquence = freq(condition).

2 Le Support

A chaque règle, on associe une mesure de support, c'est le nombre d'apparition d'une règle (condition et résultat) dans la base de données :

Support = freq(condition et résultat).

3 Le taux de confiance

C'est le rapport entre le support, où tous les champs figurant dans la règle apparaissent, et la fréquence, où les champs de la partie condition apparaissent, soit :

Confiance = freq(condition et résultat)/ freq(condition).

3.7 Le data mining dans le domaine médical

L'intégration d'un module de data mining dans le domaine médical fut d'un grand apport : Le Data Mining offre au praticien des outils d'analyse spécifiques aux données médicales à traiter et aux problèmes cliniques à résoudre.

Avec la numérisation des données des différents services hospitaliers, les praticiens ressentent le besoin de croiser les informations issues de ces différents services **[61]** afin d'en déduire de nouvelles informations qui leur permettront de diagnostiquer des cas qui présentent certaines complexités. Le data mining est basée sur cette optique de croisement de l'information et d'extraction de nouvelles connaissances. Cela consiste à combiner plusieurs méthodes des différentes phases du processus de data mining.

Dans les hôpitaux universitaires, pour parfaire les connaissances des étudiants, il leur est nécessaire d'être assistés sur une longue période. Pour raccourcir cette période, il leur est nécessaire de disposer d'un outil d'apprentissage **[74], [75]** qui leur permet de rencontrer divers cas et d'évaluer leurs diagnostics pour ces cas. Le data mining présente plusieurs méthodes pour mettre des systèmes d'aide à l'apprentissage comme le raisonnement à base de cas **[76].**

Les praticiens sont souvent confrontés à étudier des résultats d'analyses, et le plus souvent ces résultats sont présentés dans une structure qui ne facilite pas leur interprétation. Le data mining présente des outils de visualisation assez divers qui permettent de visualiser des résultats de façon telle qu'ils deviennent faciles à interpréter.

3.8 Exemples des systèmes de data mining dans le domaine médical

Dans cette partie, nous allons présenter certains systèmes médicaux qui font appel au data mining. Nous présenterons des systèmes dédiés aux prédictions, d'autres à la classification, d'autre à la découverte de modèle, à la planification et à la fusion.

3.8.1 Systèmes de modélisation

- Le groupe Lanner et l'Université de Anglia-Est **[77]** ont mis en place un système pour analyser la base des enregistrements des patients diabétiques. Ce système avait pour but de retrouver les règles qui décrivent d'éventuelles associations entre les observations faites sur les patients diabétiques et l'incidence de mort subite chez ces mêmes patients. La méthode mise en œuvre fait appel aux règles d'inductions associées au recueil simulé.

- Dans l'étude de Yiwen et al. **[78],** deux méthodes sont mises en œuvre pour extraire des connaissances sous forme de règles à partir de bases de données dédiées au diabète. Les deux méthodes dont il s'agit sont les règles d'association et les arbres de décision.

- Le TEMPADIS **[79]** est un système qui se base sur une approche d'ensemble des séquences d'évènements afin de découvrir des modèles séquentiels dans une base médicale. Ce système a été appliqué sur une base de données résultant de quatre années de surveillance de patients atteints du virus de l'immunodéficience humaine (VIH).

- Un projet de système ECD initié par le Centre Médical Universitaire de Duke, ambitionne d'utiliser des techniques de data mining afin de découvrir l'ensemble des facteurs de risque de naissance prématurée. La base de données utilisée regroupe 3902 cas de patients d'obstétrique. Par ailleurs, ce système fait appel aux techniques d'entrepôts de données, à l'interrogation des données, à leur traitement et à leur analyse.

- Freitas **[61]** propose un système ECD qui se base sur les méthodes de règles d'associations. Ce système a également été appliqué à une base de données concernant des patients diabétiques. Toutefois, il se distingue des autres systèmes par la sollicitation de l'expertise du médecin lors du processus de data mining, et ce en vue d'améliorer l'extraction des connaissances. Ainsi, dans ce processus semi-automatique, le médecin intervient en spécifiant les mises en correspondance des attributs des différents formats et de leur codage.

3.8.2 Systèmes de diagnostic

- L'équipe de H.Douglas et J. Cios **[74]** a mis en place un système de diagnostic d'hypoperfusion myocardique à partir d'images scintigraphiques de type SPECT (Single Proton Emission Computed Tomography), d'informations cliniques et des interprétations des médecins. L'étude menée par cette équipe est en soi très intéressante, malheureusement elle se base sur un échantillon réduit de cas (276 dossiers). Les résultats obtenus manquent donc de fiabilité.

- L'étude de Wang et al **[59]** est partie du constat qu'il y a un grand intérêt pour l'exploitation de données de série chronologique mais étonnamment il y a peu d'applications dans le monde médical. Cette étude présente un outil qui permet à des utilisateurs de diriger efficacement et aisément de grandes collections de séries chronologiques. L'approche consiste à extraire des dispositifs à partir d'une série chronologique de longueur arbitraire, et emploie la fréquence relative de ces dispositifs pour colorer une carte binaire. En visualisant les similitudes et les différences parmi les cartes binaires, un utilisateur peut rapidement découvrir des faisceaux, des anomalies, et d'autres régularités dans la collecte de données.

- Le système décrit par Fisher D. Gennari J dans **[75]** est constitué d'une interface qui permet à l'utilisateur d'exprimer une requête pour retrouver l'image ou les images qui correspondent à sa description. Le système est constitué d'un module de data mining qui apporte des informations supplémentaires telles que le pourcentage de tumeurs malignes parmi les tumeurs correspondant à la description donnée. Ces méthodes de Data Mining font pour l'essentiel appel à des règles d'association.

3.8.3 Systèmes de classification

- Jiawei et al **[60]** ont développé un système automatisé pour la classification des articles dans la structure des dépôts médicaux de document. Le système consiste à classifier les articles du plus grand dépôt médical, MEDLINE, en utilisant les méthodes de data mining. Le système fait appel à trois méthodes qui sont des variantes de la méthode classification associative. Ces méthodes prennent en considération les items fréquents et les multi-labels issus de la base de MEDLINE.

3.8.4 Systèmes de planification

➢ Le projet Health-mining **[58]** est un système générique de gestion de la maladie, établissant un processus de déroulement des opérations et les outils nécessaires qui peuvent être réutilisés directement quand le type de maladie à analyser change. Health-mining est un cadre généralisé qui implique non seulement des éléments de logiciel mais les praticiens aussi.

3.8.5 Systèmes de fusion

➢ F. Azuaje et *al.* **[80]** présentent une technique de fusion de l'information basée sur un modèle d'ECD. En utilisant deux types de bases : une base contenant des données d'électrocardiogrammes et des données d'une base des dossiers patients spécialisés dans les maladies du coeur. Le système utilise trois méthodes de fusion de données. Deux de ces méthodes combinent l'information au niveau de récupération des données à partir des deux types de bases, et fusionnent les données au niveau de l'entrée du système. Les résultats de ces trois modèles sont comparés et évalués contre l'exécution d'un système contenant une seule source.

➢ Le PADMA, acronyme de PArallel Data Minig Agent **[61]** est un système qui permet la détection des modèles dans les textes non structurés des rapports et les analyses de laboratoire de patients atteints d'hépatite C. Ce système est basé sur la succession distribuée des données, l'analyse distribuée des données et la visualisation interactive des données.

➢ Le Laboratoire de l'intelligence Artificielle de l'Université de l'Arizona a implémenté un prototype d'un système d'information de connaissances médicales **[81].** Ce système se base sur les méthodes de data mining. L'architecture de ce système propose de croiser des données de différentes sources et de faciliter l'accès aux grandes bases de données de l'institut National du cancer (NCI).

3.9 Conclusion

Dans ce chapitre, nous avons vu une présentation du processus d'extraction des connaissances à partir des données (ECD) et le noyau de ce processus qui est le Data Mining, à travers ses différentes phases et ses méthodes. Ce processus est appliqué à divers domaines tels que le génie biomédical (la génomique, l'imagerie médicale, la recherche bioclinique ...), ensuite nous avons cité des exemples du data mining appliqués dans le domaine médical.

Nous constatons que les concepts ECD et data mining recouvrent plusieurs dimensions dans le domaine médical et peuvent s'adapter à différents problèmes à des échelles différentes. Cela peut aller à l'extraction d'une information sur une facette d'un objet à la découverte de nouvelles connaissances à partir de grands entrepôts de données.

Dans cette thèse, nous ciblons un domaine d'application qui est le monde médical, comme nous avons déjà mentionné. Les connaissances extraites par notre système de ECD sont exploitées d'une façon explicite pour la construction de la base de connaissance du système de raisonnement à base de cas.

CHAPITRE 4

LE RAISONNEMENT À BASE DE CAS

4.1 Introduction

Le plus grand objectif de l'intelligence artificielle consiste à permettre à un ordinateur de reproduire le raisonnement humain. Avec l'accroissement de la puissance de calcul et de la mémoire des machines modernes, il est devenu clair que ces ressources ne suffiraient pas, à elles seules, pour conférer l'intelligence à cet assemblage de puces électroniques, aussi sophistiqué soit-il. Plusieurs techniques ont vu le jour pour tenter de rendre l'ordinateur plus intelligent, notamment les systèmes experts à base de règles, les moteurs d'inférence. Malheureusement, toutes ces approches nécessitent un expert connaisseur du domaine et voire même initié à l'intelligence artificielle pour programmer le système, pour lui fournir un ensemble exhaustif de règles ou de prédicats. Ce transfert de connaissance nécessite beaucoup de temps et d'efforts, ainsi qu'une connaissance approfondie du domaine traité. Un élément qui n'est pas couvert par le système expert ou le moteur d'inférence peut exiger la récriture de plusieurs règles afin d'apporter une correction, pour permettre l'apprentissage.

Il nous faut une méthode de raisonnement plus proche de celle de l'être humain afin de reproduire l'intelligence humaine au sein d'une machine. Le raisonnement à base de cas (case-based reasoning) représente une alternative aux systèmes experts qui se rapproche davantage du raisonnement humain. Plutôt que de dériver directement une solution à partir d'un problème, le raisonnement à base de cas recherche un problème similaire précédemment résolu et adapte la solution précédente à la situation actuelle.

Notre étude s'attachera dans un premier temps à décrire les aspects importants du cycle de raisonnement à base de cas dont le processus de raisonnement, la représentation des cas, leur recherche ainsi que le stockage dans une base de cas. Puis, dans un second temps, nous allons nous attarder sur la description des systèmes de CBR dans le domaine médical.

4.2 Systèmes CBR

4.2.1 Description d'un système CBR

Le raisonnement à base de cas constitue une technique d'apprentissage et de raisonnement par l'exemple permettant de contourner certaines complexités d'implantation d'un moteur d'inférence. Cette approche tire parti de la régularité du monde réel afin de résoudre des problèmes en recherchant la solution d'un cas semblable rencontré et résolu dans le passé. A partir d'un problème, un système à base de cas effectue une recherche dans sa base de connaissances, retourne un cas similaire au problème, en extrait la solution recherchée, l'adapte et stocke le nouveau cas engendré **[82]** .

4.2.2 Les principes du CBR

Nous considérons un *cas*, noté (pb, sol(pb)), comme la description d'un problème pb associé à sa solution sol(pb) **[83].** L'objectif principal du processus de CBR est d'établir une solution sol(cible) d'un problème cible, noté cible, en réutilisant la solution sol(source) contenue dans un cas source (source, sol(source))connu. Les deux étapes fondamentales pour cela sont la remémoration et l'adaptation (voir la Figure 4.1) La *remémoration* consiste à retrouver, parmi un ensemble de cas sources stockés dans une *base de cas*, un problème source jugé similaire au problème cible à résoudre. La solution du problème sol(source) est ensuite construite par *adaptation* de la solution du problème source remémoré.

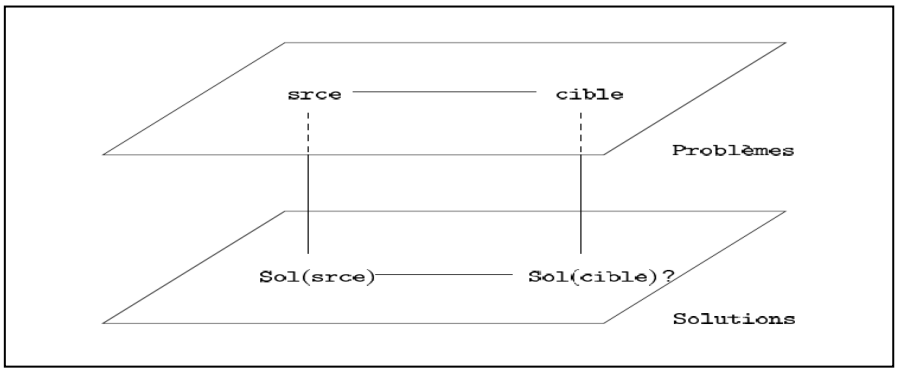

Figure 4.1 : La re-mémoration et l'adaptation en CBR

4.2.3 Processus de raisonnement

Le processus de raisonnement d'un système CBR est décrit dans le cycle de Aamodt et Plaza **[83]**: une phase recherche, une phase adaptation, une phase révision et une phase mémorisation. La figure 4.2 permet de schématiser le cycle du CBR.

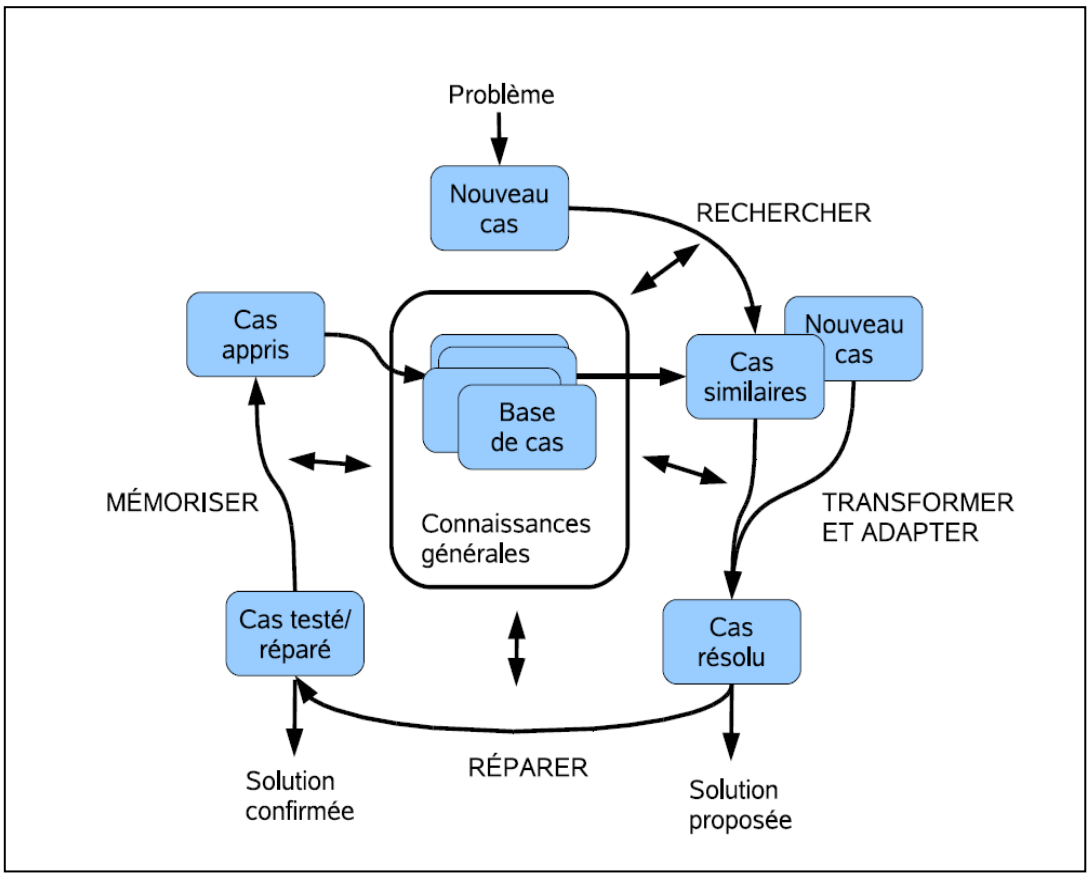

Figure 4.2 : Cycle de CBR proposé par Aamodt et Plaza **[83].**

La recherche : lors de la présentation d'un nouveau cas, cette phase permet de déterminer les cas de la base qui sont les plus similaires au problème à résoudre après l'extraction des indices connus qui vont servir à effectuer la recherche de cas analogues. La procédure de recherche est habituellement implantée par une sélection des plus proches voisins ("k-nearest neighbors") d'une structure de partitionnement obtenue par induction **[84].**

L'adaptation : elle permet de modifier les cas candidats sélectionnés lors de l'étape de la recherche, pour qu'ils répondent aux mieux à la description du cas cible. Dans la plupart des systèmes, l'étape d'adaptation nécessite une intervention humaine pour compléter une solution partielle ou tout simplement pour générer une solution entièrement à partir des cas. Ceci est dû à la difficulté de l'implémentation de cette étape et à la nécessité d'une base de connaissances intensives et un coût en termes de temps et d'efforts **[85], [86]**.

La révision : elle consiste à évaluer la solution proposée en la testant dans un environnement réel ou simulé. Le retour d'information, suite au test, peut alors réorienter en cas d'échec de la solution proposée **[87].** Afin de procéder à la révision d'un échec, il est souvent pertinent de tenter d'expliquer cet échec en analysant les différences constatées entre les résultats des solutions obtenues et ceux qu'on aurait dû obtenir **[88].**

La mémorisation : le nouveau cas est intégré dans la mémoire des cas passés. Cette intégration doit tenir compte des caractéristiques pertinentes par mise en place d'index appropriés **[89]**. La mémoire des cas est ainsi enrichie. L'apprentissage consiste à mémoriser le cas résolu et à réaliser des généralisations à partir des cas pour générer de nouvelles connaissances.

4.2.4 Les concepts de base

Pour effectuer toute tâche d'inférence, un moteur de raisonnement nécessite des connaissances sur le domaine ciblé. Pour le raisonnement à base de cas, ces connaissances forment la *base de cas*.

Les cas : Un cas est représenté par les données du problème (état initial), la solution (état final) et les contraintes qui permettent de passer de l'état initial à l'état final. Il représente l'énoncé d'un problème et sa solution **[90], [91].**

Chaque cas constitue une instance d'une classe qui spécifie l'ensemble des attributs qu'il doit contenir. Une classe peut hériter d'une ou plusieurs classes ainsi que des attributs qu'elle prescrit. Les attributs peuvent se présenter sous des types élémentaires mais aussi complexes.

Composantes d'un cas : Les éléments fondamentaux d'une base de cas constituant les cas eux-mêmes, il est important de bien décrire leur contenu et leur fonction au sein du système. Selon Kolodner **[92],** un *cas* peut être défini comme un ensemble de connaissances contextuelles enseignant une leçon. Par exemple, un cas peut se présenter sous la forme d'un problème et sa solution, de spécifications et d'un plan permettant de les satisfaire.

Contenu : trois éléments principaux apparaissent généralement dans le contenu : le problème traité par le cas, la solution et le résultat de la solution.

Caractéristiques : Afin de pouvoir comparer les cas entre eux, il doit être possible d'extraire des caractéristiques du contenu des cas. Kolodner définit une *caractéristique* ou

un *descripteur* par une paire (attribut, valeur) associée à un cas ou à un ensemble de cas. Par exemple, l'âge du patient et son sexe représentent des caractéristiques utilisées par CASEY dans le but d'effectuer un diagnostic.

Format des cas : Chaque cas constitue une instance d'une classe qui spécifie l'ensemble des attributs qu'il doit contenir. Une classe peut hériter d'une ou plusieurs classes ainsi que des attributs qu'elle prescrit. Les attributs peuvent se présenter sous des types élémentaires mais aussi complexes **[93].**

4.2.5 Stockage des cas

Il existe plusieurs formats pour stocker la base de cas:

4.2.5.1 Stockage dans une base de données relationnelles

Les bases de cas partagent plusieurs points communs avec les bases de données. Leur extensibilité permet le stockage d'un grand nombre de cas. Des mesures assurent la fiabilité du stockage des cas. Dans notre approche, la solution consiste à stocker les cas dans une base de données relationnelle (SGBD), les cas sont alors stockés dans une table de la base de données, chaque attribut correspondant à un champ, chaque cas, à un enregistrement **[94].**

4.2.5.2 CASUEL

Le langage CASUEL **[URL3]** représente une tentative de formatage unifié de bases de cas. Il utilise une structure orientée objet fondée sur les cadres pour stocker les cas dans un fichier ASCII. Chaque cas est représenté comme un objet d'une classe donnée. Malheureusement, CASUEL n'est pas adapté à la conception et à la planification. La conception exige des contenus complexes et hiérarchiques qu'une représentation (attribut, valeur) rend complexe tandis que la planification demande un ordre sur les attributs.

4.2.5.3 Utilisation de XML

Le Case-Base Markup Langages (CBML) **[95]** constitue une tentative de classe de documents XML permettant de décrire les cas par des caractéristiques simples seulement. Le CBML s'avère trop restrictif, car il impose un format pour le contenu des cas et une adaptation des outils actuels de raisonnement à base de cas. De même, Il augmente significativement la somme de données utilisées pour représenter les cas **[96].**

4.2.6 Concept de similarité

Tout système, ayant pour but d'analyser ou d'organiser automatiquement un ensemble de données ou de connaissances, doit utiliser, sous une forme ou une autre, une fonction de similarité dont le but est d'évaluer précisément les ressemblances ou les dissemblances qui existent au sein de ces données. Cette notion de similarité a fait l'objet d'importantes recherches dans des domaines extrêmement divers tels que l'analyse des données, la reconnaissance des formes, les sciences cognitives ou encore l'apprentissage symbolique.

Dans le domaine de l'IA, la similarité est surtout utilisée en apprentissage symbolique où l'on essaie de déterminer une ressemblance entre individus.

4.2.6.1 La distance Euclidienne

La distance euclidienne est un cas particulier de la méthode de la mesure de Minkowski qui s'écrit sous la forme suivante **[97]:**

$$d_p(x, y) = \left(\sum_{i=1}^{k} w_i \times |x_i - y_i|^p\right)^{1/p} \qquad (I)$$

Dans cette formule, les variables xi et yi représentent respectivement les valeurs du i ^{eme} attribut décrivant les individus x et y et le terme W_i représente le poids associé à cet attribut. Cette fonction s'applique lorsque les individus sont décrits par des variables numériques.

La distance euclidienne est calculée par la fonction suivante :

$$d_2(x,y) = \left(\sum_{i=1}^{k} w_i \times |x_i - y_i|^2\right)^{1/2} \qquad (II)$$

Il est à noter que ce que nous appelons une *distance euclidienne* correspond à la notion classique de distance entre deux points dans un espace à deux dimensions ; elle est calculée à l'aide du théorème de Pythagore.

4.2.6.2 Mesures de similarités locales

- **similarité symbolique**

Pour les attributs symboliques, la similitude entre deux attributs peut être calculée en utilisant l'équation (1), pour les valeurs nominales (c'est-à-dire de types disjonctives et catégoriques non ordonnées), et l'équation (2), pour les valeurs ordinales (valeur ordonnée selon une échelle)

$$\text{Sim}(a_i, b_i) = \begin{cases} 1 \text{ si } a_i = b_i \\ 0 \text{ si } a_i \neq b_i \end{cases} \quad (1)$$

$$\text{Sim}(a_i, b_i) = \begin{cases} 1 \text{ si } a_i = b_i \\ 1 - (d / N) \text{ si } a_i \neq b_i \end{cases} \quad (2) \qquad (III)$$

où a et b sont les attributs qui caractérisent deux cas A et B.

- **similarité numérique**

La similarité sur des variables quantitatives (numériques) mesure l'écart entre les deux objets de manière relative par rapport à l'étendue de la distribution de la variable; elle est définie par la formule suivante :

$$\text{Sim}(a_i, b_i) = 1 - [\,|a_i - b_i|\,/\text{range}] \qquad (IV)$$

Nous rappelons que a et b sont les attributs qui caractérisent deux cas A et B et range est la valeur absolue de la différence de la cardinalité minimale entre a et b et la cardinalité maximale entre a et b.

4.2.6.3 Mesures de similarités globales

Dans le CBR le calcul des similarités globales est fait à partir des similarités locales.

Prenons l'exemple de cas modélisés comme des collections de n couples d'<attribut : valeur>. Le calcul de similarité de deux cas de ce type est alors fondé sur n calculs de similarité, dédiés à la comparaison, deux à deux, de chacun des attributs des deux cas.

Ainsi, si Sim_i (cas1, cas2) représente le résultat du calcul de similarité du i-ème attribut des cas 1 et 2, alors la formule du calcul de similarité de deux cas est la suivante :

$$\mathbf{Sim\ (cas1, cas2) = [\sum_{i=1}^{n} Sim_i\ (cas1, cas2)]/\ n} \qquad \mathbf{(V)}$$

Où Sim_i = 1 si l'attribut du cas 1 est égal à l'attribut i du cas 2,

et Sim_i = 0 si les $i^{èmes}$ attributs des cas 1 et 2 sont différents

Notons que ce calcul de similarité est normalisé (divisé par n) afin de donner un résultat obligatoirement compris entre 0 et 1. Cette normalisation est le principe le mieux adapté pour pouvoir donner, ensuite, des résultats d'appariement sous la forme de pourcentages de similarité de deux cas (si Sim=1 alors 100% de similarité et si Sim=0,1, on a 10% de similarité). Il est également possible d'ajouter des pondérations pour faire varier l'importance de la prise en compte de certains attributs. On obtient alors la formule suivante :

$$\mathbf{Sim\ (cas1, cas2) = [\sum_{i=1}^{n} w_i Sim_i\ (a_i, b_i)]/ \sum_{i=1}^{n} w_i} \qquad \mathbf{(VI)}$$

Où Sim_i = 1 si l'attribut du cas 1 est égal à l'attribut i du cas 2,

Et Sim_i = 0 si les i-èmes attributs des cas 1 et 2 sont différents

et où w_i est le poids affecté au i-émes attribut

4.2.6.4 Principe des k-proches voisins

Le principe des k-proches voisins consiste à la recherche des k-proches cas au cas courant en utilisant la mesure de distance, et aussi la sélection de la classe de la majorité de ces k-cas comme étant la plus pertinente. En d'autres termes, pour la classification du cas courant, la confiance de chaque classe est calculée par m_i/k, ou m_i est le nombre de cas parmi les k-proches cas qui appartiennent à la classe i. La classe ayant la confiance la plus élevée est alors assignée au cas courant. Habituellement, pour améliorer les chances d'une prise de décision correcte pour les cas présents qui sont proches des limites entre deux

classes, le seuil β est placé de sorte qu'au moins β parmi les k proches voisins doit convenir à la classification.

Bien que l'algorithme des K-plus proches voisins soit simple, il souffre de plusieurs inconvénients : lorsque le nombre de dimensions des attributs et le nombre de cas dans la base de cas est large, le calcul requis pour la classification est énorme **[98].**

4.3 Systèmes CBR dans le domaine Médical

Un grand nombre de systèmes de raisonnement à base de cas ont été développés pour le domaine médical. Ce sont principalement des systèmes d'aide au diagnostic, c'est-à-dire des applications qui suggèrent des orientations diagnostiques voire qui proposent des diagnostics médicaux.

4.3.1 Caractéristiques des systèmes CBR médicaux

Pour évoquer les propriétés des systèmes CBR médicaux, cette section va s'appuyer sur la description de Nilsson et Sollenborn **[82].**

- **L'adaptabilité** : Adapter un système CBR à un domaine aussi complexe que le domaine médical peut s'avérer une tâche difficile, car les cas médicaux présentent beaucoup d'attributs, et le changement d'attributs mène au changement des cas **[99].**

- **La taille de la base de cas** : Plus la librairie de référence est documentée en cas prototypes, plus le système est capable de faire converger des cas donnés vers un cas général, donc la taille de la base de cas augmente.

- **L'autonomie** : Le degré d'autonomie est important pour un système de diagnostic, car il dénote le niveau d'interaction nécessaire avec le médecin avant et après le diagnostic. Le degré d'autonomie se fait donc en fonction du besoin de l'intervention humaine dans le cycle de raisonnement et lors de l'évaluation des résultats.

- **Les contraintes de sécurité** : Tout système doit se préoccuper de la fiabilité et de la sûreté de ses résultats. Ainsi, un système de diagnostic médical doit présenter un très haut niveau de fiabilité pour être utilisé dans le domaine médical.

4.3.2 Avantages et inconvénients des systèmes CBR médicaux

4.3.2.1 Avantages

Girel et Schmidt **[99]** ont identifié les différents avantages du raisonnement à base de cas en médecine :

– **Suffisance cognitive** : le raisonnement à base de cas ressemble à la démarche diagnostic ou thérapeutique des médecins dans leur pratique quotidienne, donc il facilite le travail des médecins.

– **Expérience explicite** : un système de raisonnement à base de cas répond à certains besoins des cliniciens, par exemple : disposer d'exemples qui existaient déjà et faire la comparaison entre les nouveaux cas.

– **Acquisition automatique de connaissances subjectives** : les systèmes de raisonnement à base de cas présentent une acquisition incrémentielle des connaissances à partir des cas.

– **Système d'intégration** : les dossiers patients sont quotidiennement collectés par les hôpitaux et les praticiens, ils peuvent alimenter ainsi la base de cas d'un tel système. Ils sont alors déjà enregistrés sur des supports faciles à exploiter par un système de raisonnement à base de cas.

– **Dualité de la connaissance subjective et objective** : contrairement aux systèmes experts qui se basent uniquement sur les connaissances subjectives d'experts, les systèmes de raisonnement à base de cas sont construits à partir de cas réels et à partir d'une modélisation du raisonnement (issues de la base de cas ou de l'expert).

4.3.2.2 Inconvénients

Cependant, les systèmes CBR présentent de nombreux inconvénients :

– **L'adaptation** : en raison du grand nombre d'attributs composant un cas médical, l'adaptation d'un cas est problématique. Néanmoins, les méthodes de généralisation et d'identification des attributs pertinents nous aident partiellement à remédier à cela.

– **Le manque de fiabilité** : bien que la fiabilité d'un système de raisonnement à base de cas tend à augmenter avec la proportion de couverture du domaine, elle n'est cependant pas toujours garantie. L'ajout de nouveaux cas ne fera pas forcement convergé le système vers une plus grande fiabilité **[100].**

– **Concentration sur une référence** : les systèmes de raisonnement à base de cas reposent sur des références. Ainsi, ces systèmes ne peuvent fonctionner sans base de cas.

– **L'autonomie** : les systèmes CBR nécessitent un degré élevé d'interaction humaine en particulier dans le cycle de raisonnement et dans l'évaluation des résultats.

– **Volume de stockage** : la base de cas peut s'élargir indéfiniment, ce qui risque de poser des problèmes d'espace de stockage.

Cependant, malgré certains inconvénients, elle reste une méthode très prisée dans le monde médicale grâce à ces différents avantages.

4.3.3 Classification des systèmes CBR médicaux

Nilsson et Sollenborn dans **[82]** classent les systèmes CBR de façon suivante :

4.3.3.1 Les systèmes de diagnostic

La majorité des systèmes de raisonnement à base de cas médicaux appartiennent à la catégorie des systèmes de diagnostic. Ils tentent de fournir une aide aux praticiens dans la détermination d'un diagnostic et thérapeutique, suivant différents degrés d'assistance. Notre système est classé dans cette catégorie bien que son objectif est d'orienter les patients vers un service.

4.3.3.2 Les systèmes de classification

Ces systèmes s'évertuent à identifier le groupe d'affiliation d'un cas donné. Le système de classification d'images dans le domaine de la radiologie est un exemple typique.

4.3.3.3 Les systèmes de tutorat

Adjoindre à un système de raisonnement à base de cas, le concept de l'apprentissage par des exemples, permet de définir un système d'apprentissage. Ainsi un système d'aide médical permet à des étudiants en médecine d'accéder à des cas réels lors de leur apprentissage.

4.3.3.4 Les systèmes de planification

Ce type de systèmes offre une assistance dans la programmation de processus comprenant plusieurs étapes.

4.3.3.5 Les systèmes hybrides

Ce sont des systèmes qui couplent le raisonnement à base de cas à d'autres types d'outils, dans l'optique d'apporter des solutions multi-facettes pour un espace de problème donné. Ainsi, de nombreux systèmes à base de cas médicaux sont dits hybrides, car ils font également appel aux méthodes des réseaux de neurones afin d'identifier des cas difficiles à identifier.

4.3.4 Exemples des systèmes CBR médicaux

Un grand nombre de systèmes de raisonnement à base de cas ont été développés pour le domaine médical. Ce sont principalement des systèmes d'aide au diagnostic, c'est-à-dire des applications qui suggèrent des orientations diagnostiques voire qui proposent des diagnostics médicaux. Malheureusement, ces systèmes ne sont pas utilisés en routine clinique. Ils demeurent en effet des prototypes confinés au domaine de la recherche.

4.3.4.1 Systèmes de diagnostic

- FM-Ultranet **[101], [102]** est un projet médical de CBR mis en application avec CBR Works. FM-Ultranet qui détecte des malformations et des anomalies de fœtus par les examens ultra sonographiques. La détection ou le diagnostic emploie des attributs dérivés des balayages de l'utérus de la mère, et identifie les organes et les extrémités anormaux. Les cas sont arrangés de façon hiérarchique et dans une structure orientée objet.

- Perner **[103]** propose un système qui emploie CBR pour optimiser la segmentation d'image selon le type d'acquisition d'image et la qualité d'image. Le système a été employé pour détecter la maladie dégénérative d'Alzheimer en. Les cas sont composés des images et des attributs d'image aussi bien que des informations sur l'acquisition d'image et le patient. La solution d'un cas s'appuie sur les paramètres de segmentation de l'image.

- CARE-PARTNER **[100], [104]** est un système interactif d'aide à la décision pour le suivi à long terme des patients transplantés de cellules osseuses au centre de recherche sur le cancer de Fred Hutchinson (FHCRC) à Seattle. Le système de CARE-PARTNER donne l'appui médical et de décisionnel aux centres de soin qui suivent les patients transplantés, en utilisant l'internet pour relier les centres de soin aux spécialistes de FHCRC. Une des caractéristiques du système est qu'il emploie une base de connaissances riche en cas prototypes et des directives de pratique pour interpréter des cas médicaux.

- Schmidt et *al* traitent spécialement des prototypes dans **[105],** ou un prototype dénote une généralisation résultant du groupement de cas simples dans un cas plus général. Le stockage de nouveaux cas peut améliorer les capacités de trouver des cas semblables. Schmidt et Gierl ont développé plusieurs systèmes se concentrant sur la généralisation de prototype, comme il a été décrit dans leur enquête médicale de CBR **[99],** tel que : ICONES pour le conseil antibiotique de thérapie, GS.52 pour le diagnostic des syndromes dysmorphiques, COSYL pour appliquer le meilleur traitement aux patients afin de soigner la pathologie dont est atteint leur foie, TeCoMED pour des épidémies de prévision des maladies infectieuses.

- Jaulent et al. **[106]** diagnostiquent l'histopathologie dans le domaine du cancer du sein. Leur système emploie des cas qui sont dérivés de plusieurs rapports médicaux. Un cas a une structure arborescente interne et représente une collection de secteurs macroscopiques. Chaque secteur macroscopique est une collection de secteurs histologiques et chaque secteur histologique contient une description cytologique des attributs. Les cas sont comparés par leur structure (structure de l'arbre histologique), leur surfe (ressemblance sémantique des secteurs microscopiques) et par la similitude des attributs.

4.3.4.2 Systèmes de classification

- Montani et *al* se sont concentrés sur CBR dans des traitements d'hémodialyse pour la dernière phase de la maladie rénale **[107].** Leur système est appliqué aux évaluations des sessions d'hémodialyse. A chaque nouvelle session de dialyse, l'évaluation est représentée comme un cas dans le système. Des modèles temporels des échecs sont conçus à partir de l'historique des patients et du croisement des références d'autres patients. Des attributs statiques et dynamiques sont collectés. Les attributs statiques sont des informations concernant les patients à caractère général (âge etc.) et les attributs dynamiques qui proviennent des mesures obtenues lors d'une session de dialyse, comme par exemple la pression artérielle.

- Costello et Wilson **[108]** se concentrent sur la classification des séquences d'ADN et en utilisant une base de cas de séquences de nucléotide (A, T, G, C). Les séquences stockées sont classées déjà comme exons (fragment codant) et introns (fragment non codant). Le système identifie des exons dans un mélange apparemment aléatoire d'exons et d'introns dans les brins d'ADN. Un calcul de distance d'un montage (d'une insertion,

d'une substitution et d'une suppression de différents nucléotides) dans les exons tests est employé pour évaluer la similitude entre le brin test et les cas d'exon de référence. Des exons assortis sont groupés par les niveaux d'activation (nombre de similitudes) pour trouver de nouveaux segments des exons dans le brin test.

- Nilsson et al **[109]** abordent le domaine des dysfonctionnements psycho-physiologiques, une forme de stress. Le système classifie des mesures physiologiques des sondes. Le système est divisé en plusieurs parties distinctes. Les mesures, comme des signaux d'un ECG, sont filtrées et améliorées. Une bibliothèque de cas des modèles de déformation etc. est appliquée aux filtres. Des attributs sont extraits à partir des signaux filtrés (mesures). Un ensemble additionnel d'attributs est extraits à partir du premier ensemble, pour l'analyse de tendance. Les attributs du premier et du second ensemble et des données patientes sont employés comme des cas. Les cas sont classés à l'aide de la méthode des *k plus proches voisins*.

- Le système TeCoMED de Schmidt et de Girel **[110]** tentent d'employer un modèle de pronostics pour prévoir des vagues d'épidémies d'influenza, basées sur des observations faites les années précédentes. TeCoMED combine CBR selon une abstraction temporelle pour traiter le problème du comportement cyclique mais irrégulier des épidémies. Des tendances sont discrétisées de la façon suivante : *très forte diminution, diminution marquée, diminution, régulier, augmentation, forte augmentation, très forte augmentation*. TeCoMED utilise d'anciens évènements et des cas d'avertissement semblables d'une façon analogue à celle utilisée dans le système IC ôNES. Une petite compagnie de logiciel a réussi de commercialiser un système qui avait incorporé les avertissements produits par le système dans des pages Web d'un plan d'assurance médicale et dans une page du service de santé de l'état fédéral.

- Montani et al **[111]** essayent d'intégrer différentes méthodologies dans un système Multi-Modal du raisonnement (MMR), utilisé dans l'appui de thérapie pour les patients diabétiques. Les auteurs affirment que la plupart des systèmes utilisant plus d'une méthode font cela seulement dans un mode exclusif, avec des méthodes fonctionnant simplement comme prolongements à une des autres méthodes. Montani affirme qu'un système de MMR a besoin d'une intégration beaucoup plus étroite des technologies pour obtenir l'ensemble des bénéfices d'une solution multi-modale. L'intégration laisse aborder les problèmes bien connus des méthodologies simples, c'est-à-dire la qualification du problème dans CBR et le problème de trop petite base de cas dans le raisonnement à base

de cas. Le système proposé essaie d'employer une intégration complète et d'utiliser le CBR, le raisonnement à base de règles et le raisonnement à base de modèles (RBM).

4.3.4.3 Systèmes de tutorat

- WHAT **[112]** est un système de tutorat médical basé sur le raisonnement à base de cas pour l'apprentissage des étudiants en médecine sportive. WHAT est conçu pour donner les prescriptions mieux assorties d'exercices que l'approche basée sur les règles conservatrices enseignées par la plupart des livres. Le système fournit deux recommandations séparées pour les prescriptions d'exercices, une qui est basée sur les règles trouvées dans les livres, l'autre emploie CBR avec une base de cas conçue par un expert. Les exercices prescrits sont appliqués aux patients atteints de problèmes cardiaques et pulmonaires. Les prescriptions sont basées sur des dispositifs d'antécédents médicaux des patients et sur leurs tests physiologiques.

- Bichindaritz et al **[104]** ont transformé CARE-PARTNER en un système médical de formation sur le net. L'intention est d'aider les étudiants en médecine à améliorer leurs connaissances en résolvant des cas pratiques. Les cas prototypes se composent des pistes cliniques, qui peuvent être travaillées pour produire des cas ayant un de niveau variables de complexité. Ce système permet de créer des cas virtuels avec un degré de complexité prédéfini. Le système peut également évaluer les solutions données par les étudiants pour des cas pratiques.

Si la solution de l'étudiant soit assortie à la solution de référence, des points d'exactitude sont calculés et la solution de l'étudiant est placée dans une des trois catégories : n'atteint pas les normes, adéquate, atteint les normes.

4.3.4.4 Systèmes de planification

- Le projet d'Auguste **[113]** fournit une aide à la décision pour la planification des soins continus des patients atteints de la maladie d'Alzheimer. Un premier prototype de système d'aide à la décision a été développé, il permet la prescription des médicaments neuroleptiques pour des troubles du comportement. Le prototype est un système hybride ou la partie CBR décide si un médicament neuroleptique doit être donné, et la partie basée sur le raisonnement à base de règles décide quel neuroleptique est à employer. Le système emploie approximativement 100 attributs, extraits à partir de la charte médicale. Le patient est au commencement interrogé pour des troubles du comportement avant que la méthode du voisin le plus proche ne suggère si le patient doit prendre des neuroleptiques. Si le

module de CBR trouve qu'il est approprié de donner le neuroleptique et qu'aucune contradiction n'est trouvée (par exemple des allergies à certaines médicaments etc.), le module de raisonnement à base de règles détermine quel neuroleptique est à employer.

- Davis et al. **[114]** emploient un système de planification basé sur la structure du ReCall CBR. Le système décide quel genre de dispositifs ont besoin les infirmes et les vieilles personnes dans leur maison pour avoir une vie indépendante. Des dispositifs sont construits, des manuelles et des rapports sont écrits. Le système contient 10 groupes de problèmes et 14 groupes de solutions. Chaque groupe est subdivisé par un arbre de décision C4.5. L'arbre de décision est reconnu pour son efficacité et sa facilité d'interpréter le processus de raisonnement.

4.4 Conclusion

Notre objectif est de concevoir un système de gestion de connaissances dans le domaine médical. Nous avons montré dans les chapitres précédents comment nous avons pu répondre à cet objectif tout en essayant de prendre en compte les relations établies entre les différentes approches telles que les entrepôts de données,l'extraction de connaissances à partir de données et le raisonnement à base de cas, afin de préparer les données qui étaient hétérogènes et distribuées par la construction d'un entrepôt de données médical, puis extraire à partir de ces données des connaissances fiables et pertinentes pour construire notre base de connaissances qui sera utilisée dans le raisonnement à base de cas.

CHAPITRE 5

PROCESSUS D'EXTERACTION ET D'EXPLOITATION DES CONNAISSANCES MÉDICALES : DW, DM ET CBR

5.1 Introduction

La gestion des connaissances médicales vise à améliorer les performances de l'organisation médicale en permettant aux individus de l'établissement de soins (médecins, infirmières, paramédicaux, etc.) de capturer, partager et appliquer des connaissances collectives pour faire prendre des décisions optimales en temps réel. Notre idée est d'orienter des patients vers les différents services et les différentes spécialités dans les grands établissements de santé tels que les CHU (Centre Hospitalo-Universitaire). Si le patient ne peut reconnaître les premiers signes d'un problème médical ou d'une lésion, il peut prendre un rendez-vous de plus de deux ou trois mois chez un médecin spécialiste dans un service sans qu'il soit le bon médecin.

Le besoin de développer un système de gestion des connaissances dans le domaine médical nécessite une connaissance des problèmes dans ce domaine.

Dans le premier chapitre, et après avoir aborder la gestion des connaissances dans le domaine médical, nous avons constaté que la diversité des sources d'informations distribuées et leur hétérogénéité sont une des principales difficultés rencontrées dans le domaine médical.

Dans le deuxième chapitre nous avons proposé une solution pour résoudre les types de conflits afin de combiner ou fusionner correctement les informations et connaissances issues de ces sources médicales. Plusieurs approches ont été proposées dans ce cadre. Parmi elles l'entrepôt des données.

Dans le troisième chapitre nous avons traduit les techniques d'extraction de connaissances à partir de données stockées dans les entrepôts. Les résultats de ce travail sont la base de connaissances qui va nous permettre d'implémenter l'étape de raisonnement à base de cas qui est détaillée dans le quatrième chapitre.

Le présent chapitre décrit l'approche proposée, nous présentons le processus de construction de l'entrepôt de données médicales, ensuite nous choisissons une des techniques d'extraction des connaissances à partir de données, enfin le résultat de la technique proposée est une base de connaissances utilisée pour le système de raisonnement à base de cas.

5.2 Intégration des données hétérogènes

Cette partie consiste à extraire, à l'avance, les données médicales pertinentes pour l'usage des utilisateurs (infirmier, médecin, radiologue, biologiste…), à les filtrer, les transformer et les stocker. Pour répondre à nos besoins, nous allons construire un entrepôt de données médicales qui stocke physiquement les données des sources réparties, ces dernières correspondent généralement aux sources de données opérationnelles de l'établissement de santé. Pour être exploitables, toutes les données provenant des systèmes distribués doivent être organisées, coordonnées, intégrées et enfin stockées pour donner à l'utilisateur une vue globale des informations.

5.2.1 La conception de l'entrepôt de données médical

Dans notre travail, nous avons choisi l'architecture réelle, elle est généralement retenue pour les systèmes décisionnels, le stockage des données est réalisé dans un SGBD séparé du système opérationnel, ces systèmes sont conçus d'une façon autonome, distribuée. Le système intégral regroupe l'ensemble des systèmes de production telles que:

- **La gestion des consultations :** permet la prise en charge des résumés de sortie, des diagnostics et des consultations externes et internes effectuées. Lesquelles données sont pertinentes et sollicitées par les praticiens.

* Les données sont stockées dans le SGBD Microsoft Access avec une implémentation d'une application sous le Delphi.

- **La gestion des RDV médicaux :** suit les rendez-vous médicaux des patients chez un médecin ou un service.

* Les données sont stockées dans le SGBD Access de Microsoft avec une implémentation d'une application sous le Delphi séparé.

- **La gestion des admissions :** consiste à suivre l'épisode d'hospitalisation d'un patient depuis son arrivée à l'établissement jusqu'à sa sortie, Elle vise à *:*

- La saisie du résumé clinique et éventuellement : les antécédents du patient, le diagnostic d'entrée,
- la saisie de l'intervention,
- la saisie du compte rendu opératoire : le préambule clinique, le compte rendu de l'opération, les actes opératoires, les interventions chirurgicales, le diagnostic post et pré – opératoire,
- la saisie des résultats d'analyses, la fiche navette,
- la saisie du résumé de sortie : la conclusion de sortie, le diagnostic de sortie.

* Les données sont stockées dans le SGBD propriétaire de Visuel Basic de Microsoft.

- **La gestion d'hébergement hospitalier :** consiste à suivre l'épisode d'hébergement d'un patient dans l'établissement jusqu'à sa sortie, Elle vise essentiellement à :
  - Gérer les informations médico-administratives du patient.
  - Gérer la localisation du patient dans l'établissement : urgence, salle de soin, bloc opératoire, chambre, lit…

* Les données sont stockées dans des classeurs de Microsoft Excel.

- **La gestion de laboratoires et analyses médicales** : consiste à suivre les examens du laboratoire, elle couvre les spécialités suivantes : Biochimie, Parasitologie, Anatomie, Cytologie, Hématologie, Immunologie, Bactériologie-virologie…, vise à:
  - Suivre les analyses biologiques.
  - Préparer les fiches de l'automate du laboratoire.
  - Préparer les résultats et le bilan.

* Les résultats sont stockés dans des classeurs de Microsoft Excel, et les bilans proviennent de l'automate du laboratoire sous une forme privée

- **La gestion de radiologie et imagerie médicale** : vise à :
  - Suivre les demandes d'examens radiologiques.
  - Préparer les fiches de résultats.
  - Fournir les rapports, les clichés et les comptes rendu des radiologies.
  - Fournir les images radiologiques des scanners, échographies, mammographies…

* Les comptes rendus de ce système sont stockés dans le Microsoft Word, et les images sont stockées dans des formes privées ou bien dans le logiciel DCOM.

- **La gestion gynécologie et bloc obstétrical :** suivre la gestion du dossier médical adopté au niveau de service de gynécologie/obstétrical de l'établissement de santé. Elle permet en premier lieu:

  - L'identification des patients, la prise en charge des fiches médicales.

  - Suivre la fiche Gynécologique, la fiche Obstétrique et la fiche d'accouchement.

  - La prise en charge des données du nouveau né.

  - La génération des états de sortie.

* Les données sont stockées dans des classeurs de Microsoft Excel.

- **La gestion financière et comptabilité** : permet de connaître à tout moment les recettes de l'établissement, ses dépenses et la situation de sa trésorerie. Son but est de:

  - Informatiser la comptabilité générale dans l'établissement hospitalier.

  - Faciliter et informatiser la gestion financière dans l'établissement hospitalier.

* Les données sont stockées dans le SGBD propriétaire de Delphi.

- **La gestion du patrimoine :** consiste à faire l'approvisionnement et la gestion du stock d'articles magasin de l'établissements de santé. Elle vise à :

  - Gérer plusieurs dépôts.

  - Rationaliser et contrôler la consommation des articles.

  - Fournir instantanément la situation des stocks des articles.

  - Valoriser la consommation des articles par période, service, dépôt, famille d'articles.

* Les données sont stockées dans le SGBD propriétaire de Delphi.

- **La gestion de la pharmacie et des stocks de médicaments :** consiste à gérer la pharmacie et les stocks de médicaments de l'établissement de santé, vise à :

  - Rationaliser et contrôler la consommation des médicaments et le stick minimum.

- Valoriser la consommation des médicaments par période, patient, service, dépôt et classe thérapeutique.

- Vérifier les médicaments périmés, et les dates de péremption.

* Les données sont stockées dans le SGBD propriétaire et implémenter sous Delphi.

- **La gestion d'hémodialyse :** consiste a suivre le dossier médical et administratif des patients dialysés, elle vise à :

- Suivre les données médicales du patient (tension artérielle, poids, températures…)

- Suivre les rendez-vous et les séances de dialyses.

* Les données sont stockées dans des classeurs de Microsoft Excel.

- **La gestion des ressources humaines (GRH) :** consiste à gérer le personnel, la paie, et gestion des carrières.

5.2.2 La construction de l'entrepôt de données médical

Pour la construction de l'entrepôt de données médicales, nous avons suivi le processus d'ETL (Extraction, Transformation, Loading). Notons, cependant que cette décomposition est seulement logique. L'étape d'extraction et une partie de l'étape de transformation/intégration ; elles sont groupées dans le même composant logiciel, c'est l'adaptateur. Ces étapes sont couplées dans un même composant logiciel qui réalise le chargement dans l'entrepôt de données.

5.2.2.1 L'extraction des données

Cette phase collecte les données utiles des sources de données opérationnelles à partir des différentes sources hétérogènes.

Il faut poser la question : Quelles sont les données opérationnelles qu'il faut sélectionner pour alimenter notre entrepôt médical ? Toutes les données sources ne sont forcément pas utiles. Doit-on prendre le compte rendu radiologique ou bien le résultat de l'examen radiologique sans détail?

Les données que nous allons extraire peuvent être :

➢ Interaction d'un médecin : cette interaction se traduit en fichier texte, ordonnance ou bien un compte rendu des résultats médicaux,

➢ Bilan d'un test biologique : ce bilan peut être des résultats numériques d'un automate du laboratoire, un bilan pré-opératoire, un bilan post-opératoire stockés dans un fichier texte ou bien un fichier Excel ou bien une base de données.

➢ Fichier image : résultant d'une application d'imagerie numérique (DCOM), d'un radio scanner ou IRM.

➢ Séries temporelles : ces données sont extraites à partir d'appareils spécifiques tels qu'Electro-Cardio-Gramme (ECG), appareils des soins intensifs qui fournissent une mesure de différents paramètres physiologiques (fréquence cardiaque, fréquence respiratoire …).

➢ Commentaire vocal : commentaire du résultat d'un examen interprété au fichier texte.

➢ Données cliniques : sous forme de notes cliniques, stockées dans des documents non structurées.

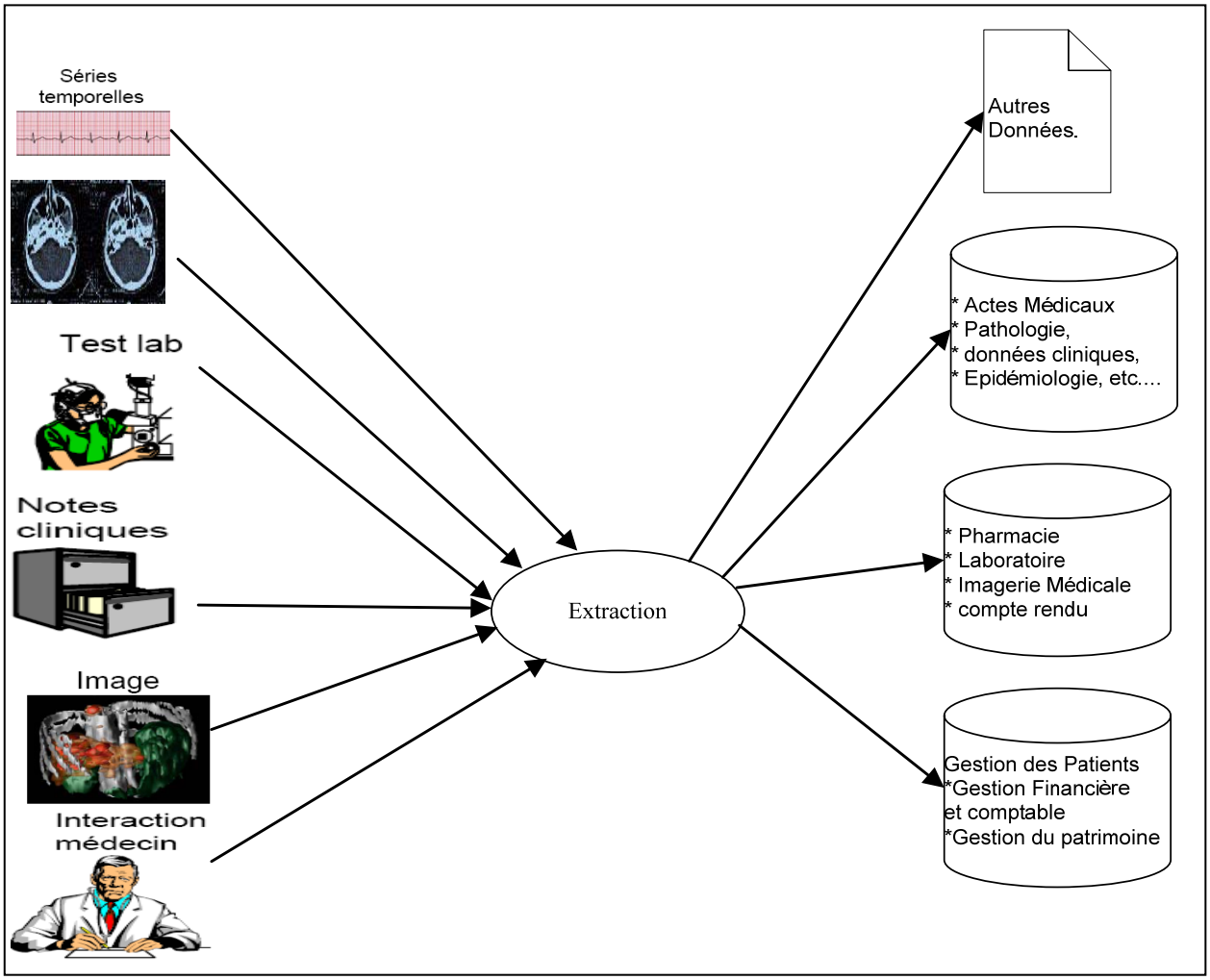

Figure 5.1 : L'extraction des données hétérogènes dans l'entrepôt de données

5.2.2.2 La transformation des données

Les types de conflits rencontrés :

Au cours de la phase de la transformation des données, nous avons rencontré plusieurs types de conflits qui sont traités chacun à part :

➢ **Conflits de classification** : ils apparaissent lorsque les types en correspondance décrivent des ensembles différents. Par exemple, deux sources médicales peuvent contenir chacune une entité nommée *« Médecin »* avec pour extension, pour la première, tous les médecins de l'hôpital, et pour la seconde uniquement les médecins spécialistes. La solution standard pour ce genre de conflit est d'inclure dans le schéma intégré la hiérarchie de généralisation/spécialisation appropriée.

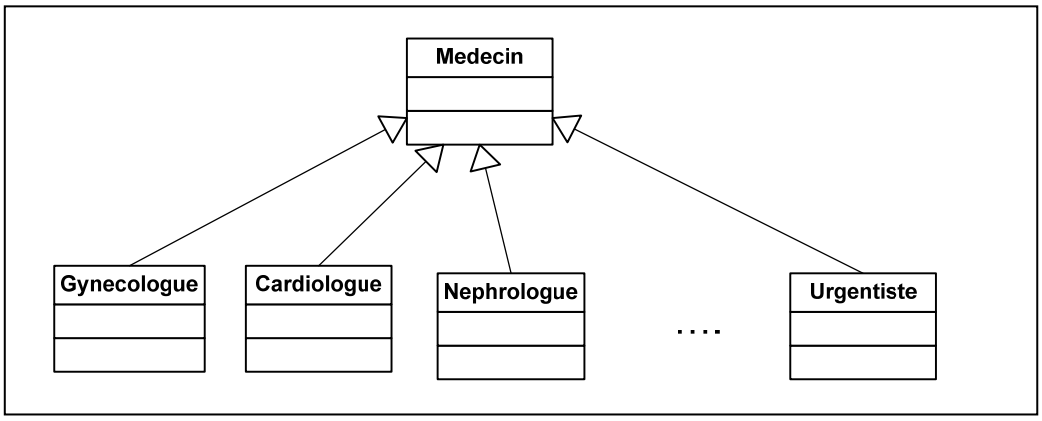

Figure 5.2 : Conflit de classification.

➢ **Conflits descriptifs** : ils surviennent dès qu'il y a une différence entre les propriétés des types en correspondance (les types d'objets peuvent différer selon leurs noms, leurs clés, leurs attributs ; les attributs peuvent aussi différer selon leur noms, leurs structures, etc.). Un exemple de ce type de conflits (différence selon le nom) est l'utilisation de termes synonymes ou homonymes dans la désignation d'un même type d'entité dans deux schémas différents : *« Spécialité d'un médecin »* dans l'un et *« Grade d'un médecin »* dans l'autre.

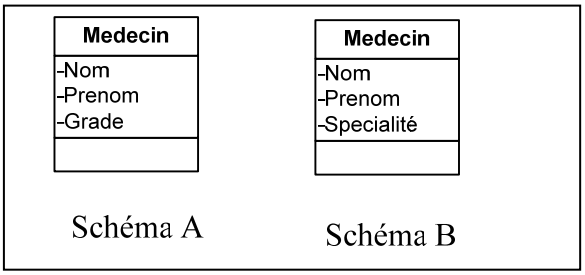

Figure 5.3 : Conflit descriptif dans deux schémas différents.

➢ **Conflits structurels** : un conflit structurel survient lorsque les éléments en correspondance sont décrits par des concepts de niveaux de représentation différents ou soumis à des contraintes différentes. Par exemple, une classe d'objet, et un attribut, ou un type d'entité et un type d'association. Par exemple, adresse qui est un *attribut* d'une table relationnelle dans un schéma A, peut correspondre à une *table* dans un schéma B.

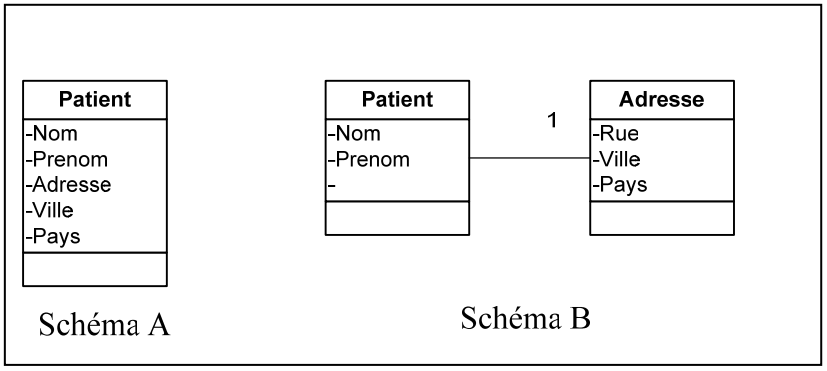

Figure 5.4 : Conflit structurel dans deux schémas différents.

➢ **Conflits données/méta-données** : ils surviennent lorsqu'une donnée dans une base est en correspondance avec une méta-donnée (le nom d'un type) dans le schéma d'une autre base. En prenant comme exemple deux schémas de bases de données de praticiens avec praticien1 (id, neurologue, ...) et praticien2 (id, fonction, ...), certaines valeurs de l'attribut fonction de praticien2 peuvent correspondre au nom de l'attribut neurologue de praticien1.

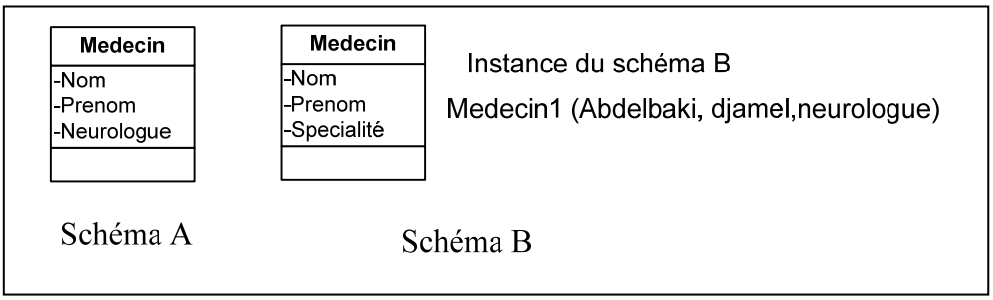

Figure 5.5 : Conflit données/méta-données dans deux schémas différents.

➢ **Conflits de données** : ils surviennent au niveau des instances, lorsque des occurrences en correspondance ont des valeurs en conflit pour des attributs en correspondance. Par exemple, deux sources différentes peuvent stocker la même instance d'un patient de diverses façons (ou avec des erreurs de saisie) : l'adresse dans l'une des sources peut ne pas être représentée de la même façon que dans l'autre. Ce sont les conflits de nommage, de graduation, de confusion, et de conflits de représentation.

1. Les conflits de nommage se produisent lors de l'attribution des noms dans des schémas qui diffèrent de manière significative. Les cas les plus fréquents sont les cas de présence de synonymes et d'homonymes.

2. Les conflits de graduation apparaissent lorsque différents systèmes de graduation sont utilisés pour mesurer une valeur. On peut citer en exemple la mesure de la température (degrés Celsius ou Farenheit).

3. Les conflits de confusion se produisent lorsque les concepts paraissent avoir la même signification mais diffèrent en réalité. Ce type de confusion peut être causé par des contextes temporels différents par exemple.

4. Les conflits de représentation, qui se produisent quand deux schémas sources décrivent le même concept de manière différente. Par exemple, dans une source l'adresse peut être désignée par une chaîne de caractères tandis que dans une autre l'adresse est une structure composée du numéro et du nom de la rue, du code postal et de la ville.

On associe à chaque source un adaptateur /moniteur qui a deux fonctions :

- L'*adaptateur* transforme les données à partir de la représentation locale vers le format objet. Les sources peuvent avoir des modèles de données et des schémas hétérogènes. La partie adaptateur transforme les données d'une source en une représentation intermédiaire et comprend également une interface pour interroger la source. Par exemple, un adaptateur relationnel associé à une source que les données stockeés selon un format relationnel. L'intégrateur tire profit de cette uniformité de représentation pour faire la fusion des sources.
- *Le moniteur* détecte automatiquement les changements de sa source et les propage vers l'intégrateur. Par rapport à la détection automatique de changements, il est possible de classifier les sources**:**

Les sources peuvent également présenter des problèmes d'hétérogénéité sémantique, ces problèmes sont traités en spécifiant un ensemble de règles de transformation pour arriver à une représentation uniforme, par exemple, remplacer le mot « genre » par « sexe » ou bien «Féminin» par «F» et «Masculin» par «M».

	Données sources	Données cibles
Appli 1 :	Genre : Masculin, Féminin	Sexe : M,F
Appli 2 :	Sexe : 1, 0	Sexe : M,F
Appli 3 :	Genre : Homme, Femme	Sexe : M,F

Tableau 5.1 : Exemple de transformation de données

L'intégration entraîne plusieurs activités de « nettoyage » liées à la transformation des données pour les rendre conformes au schéma de l'entrepôt médical et aux critères de qualités choisis**.**

Le développement de techniques et d'outils d'intégration est fait par un intégrateur qui est responsable de la réception des représentations intermédiaires des données sources, en provenance des adaptateurs à partir de spécifications déclaratives.

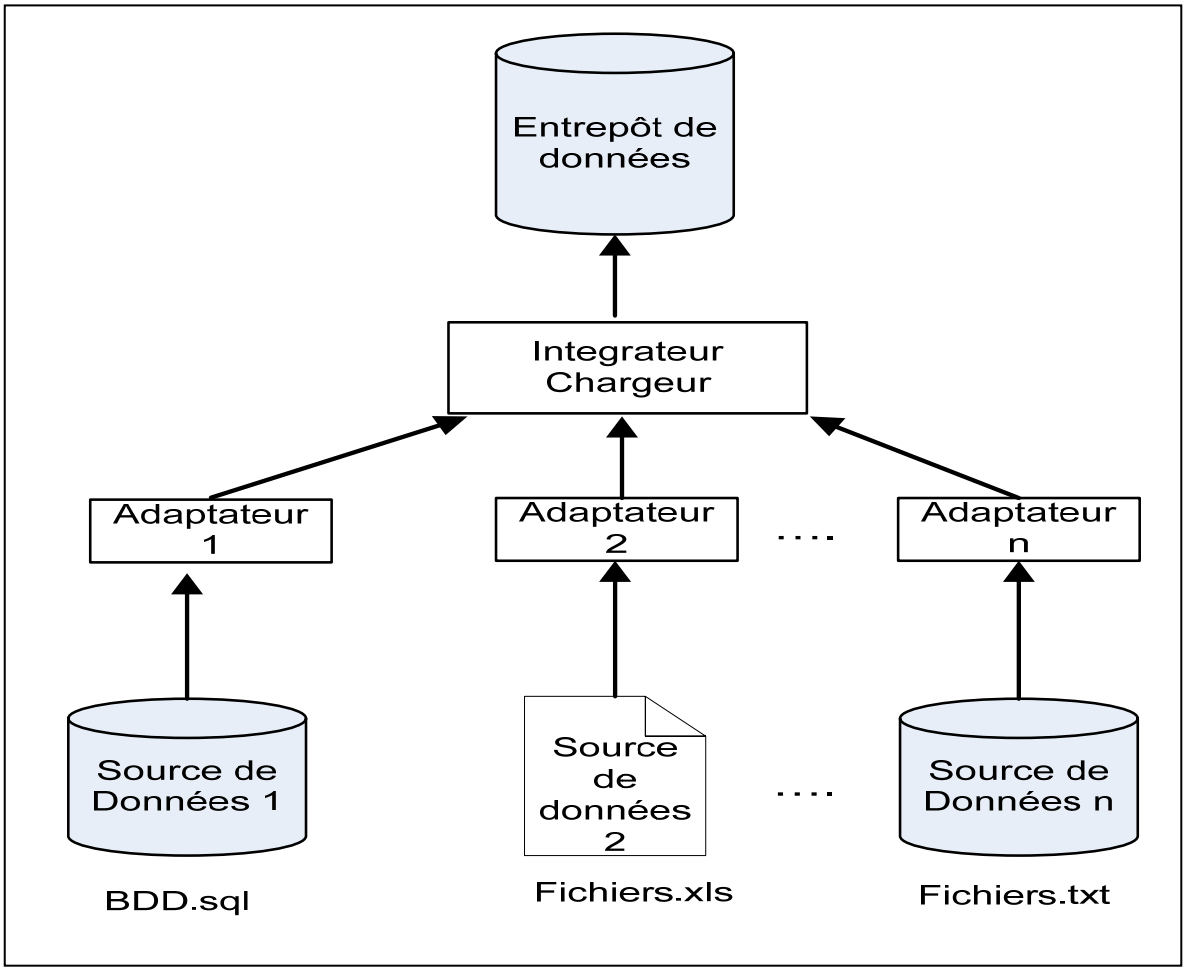

Figure 5.6 : Vue opérationnelle des composants utilisés pour la construction de l'entrepôt.

5.2.2.3 Le chargement des données intégrées dans le système cible

Le processus de rafraîchissement dans notre système est incrémental, la mise à jour utilise les changements dans les sources chaque fois qu'une source change ou bien de manière périodique avec une période qui dépend des besoins des utilisateurs et de la charge d'accès à l'entrepôt.

La figure suivante résume l'enchaînement de ces étapes de traitement :

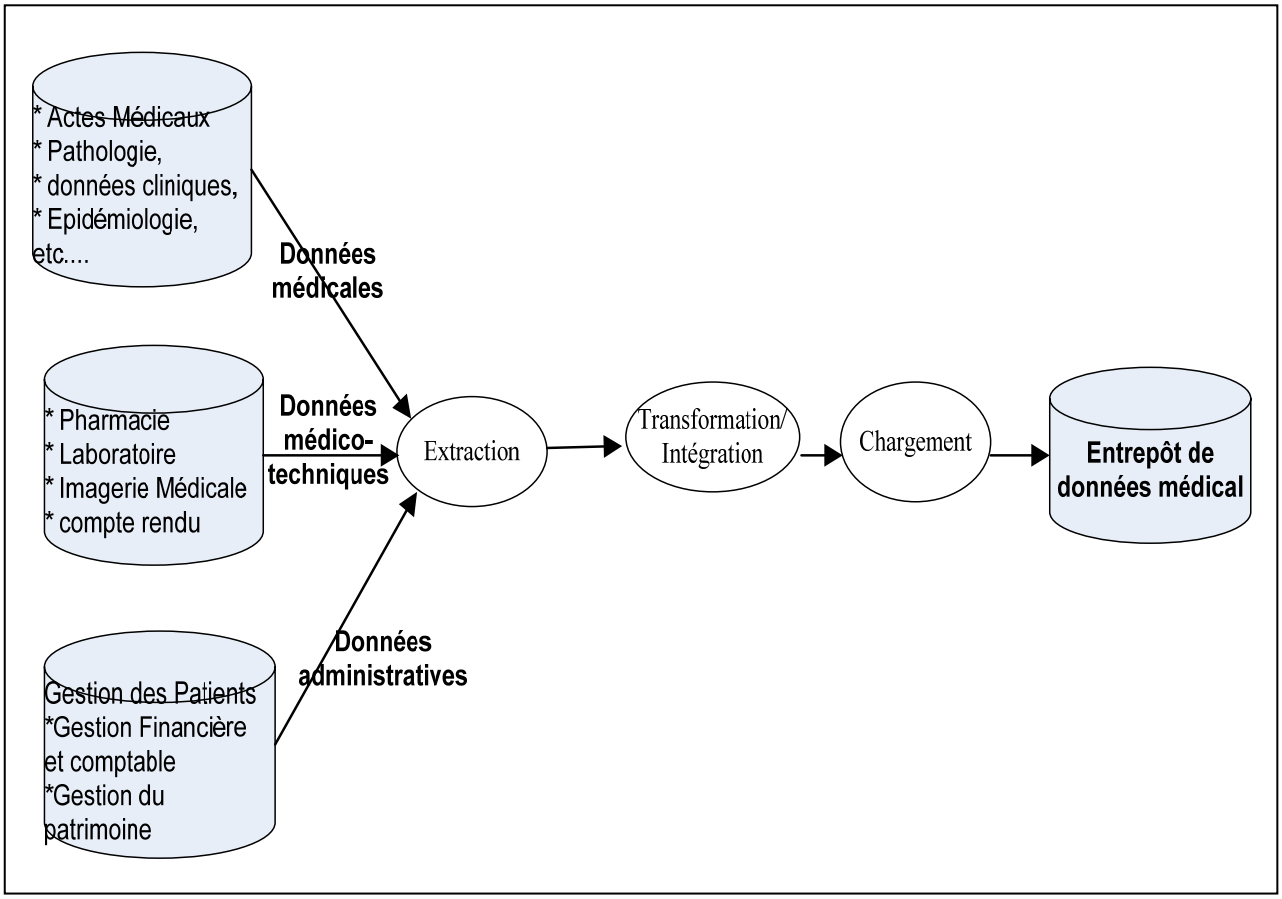

Figure 5.7: Etapes de traitement de construction de l'entrepôt de données médical.

5.2.3 Les caractéristiques de l'entrepôt de données

Les données de l'entrepôt médical que nous avons construit comprennent les caractéristiques suivantes :

- **Orientées sujet :** Les données de l'entrepôt sont organisées par sujet. Cette orientation sujet va également permettre de développer notre système décisionnel via une approche qui sera expliquée plus tard dans les sections suivantes.

Pour matérialiser cette orientation sujet nous avons construit une structure supplémentaire appelée Data Mart « Consultation » (magasin de données) qui est ciblée et pilotée par le besoin de notre système.

- **Intégrées :** Les données provenant des différentes sources hétérogènes sont intégrées, avant leur stockage dans l'entrepôt de données. Par exemple, la consolidation de l'ensemble des informations concernant un patient donné est nécessaire pour donner une vue homogène de ce patient.

- **Historisées :** l'entrepôt garde la trace de toutes les données médicales d'un patient et suit toutes ses consultations, bilans, radios et interventions.

- **Non volatiles :** Les données sont permanentes et ne peuvent pas être modifiées. A chaque rafraîchissement de l'entrepôt, on ajoute de nouvelles données, sans modifier ou perdre celles qui existent.

5.2.4 La construction du schéma de l'entrepôt

5.2.4.1 Le schéma global de l'entrepôt

Le résultat du traitement de construction de l'entrepôt de données consiste à définir un schéma global fournissant une vue intégrée des sources qui vont être exploitées par la suite dans le processus d'extraction des connaissances à partir des données. Ce schéma est représenté dans la figure suivante :

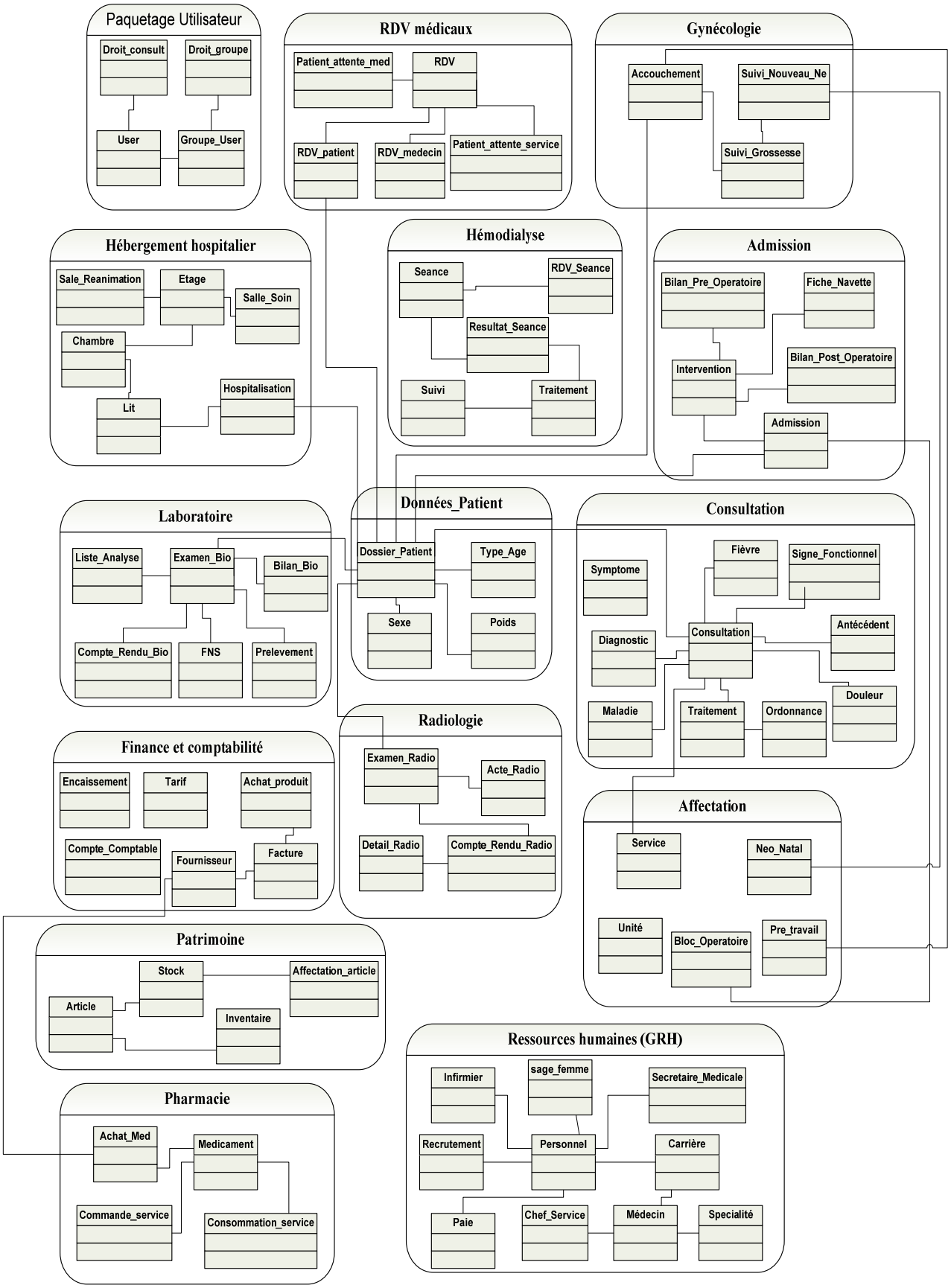

Figure 5.8 : Schéma global de l'entrepôt de données médical.

Nous avons définit dans le schéma global une représentation des différents paquetages et classes qui composent l'entrepôt de données médicales, les paquetages correspondent à des bases plus légères, destinées à quelques utilisateurs suivant leurs besoins.

5.2.4.2 Data Mart Consultation

Le data mart est ciblé et piloté par les besoins de notre système. Il a la même vocation que l'entrepôt de données médicales (fournir une architecture décisionnelle), mais il vise à résoudre notre problématique avec un nombre d'utilisateurs plus restreint

Il est caractérisé par :

- Sa petite taille par rapport à l'entrepôt de données médicales.
- Il est moins complexe et plus facile à déployer que l'entrepôt de données médicales.
- Orienté sujet et piloté par les besoins de notre système, et son champs d'application est restreint.

La figure suivante illustre la décomposition de l'entrepôt de données en plusieurs Data Marts :

1 DM Consultation;

2 DM RDV médicaux;

3 DM Admission;

4 DM Hébergement hospitalier;

5 DM Hémodialyse;

6 DM Gynécologie et bloc obstétrical;

7 DM Laboratoire et analyses médicales;

8 DM Radiologie et imagerie médicale;

9 DM Pharmacie et des stocks de médicaments;

10 DM Finance et comptabilité;

11 DM Patrimoine;

12 DM Ressources humaines (GRH);

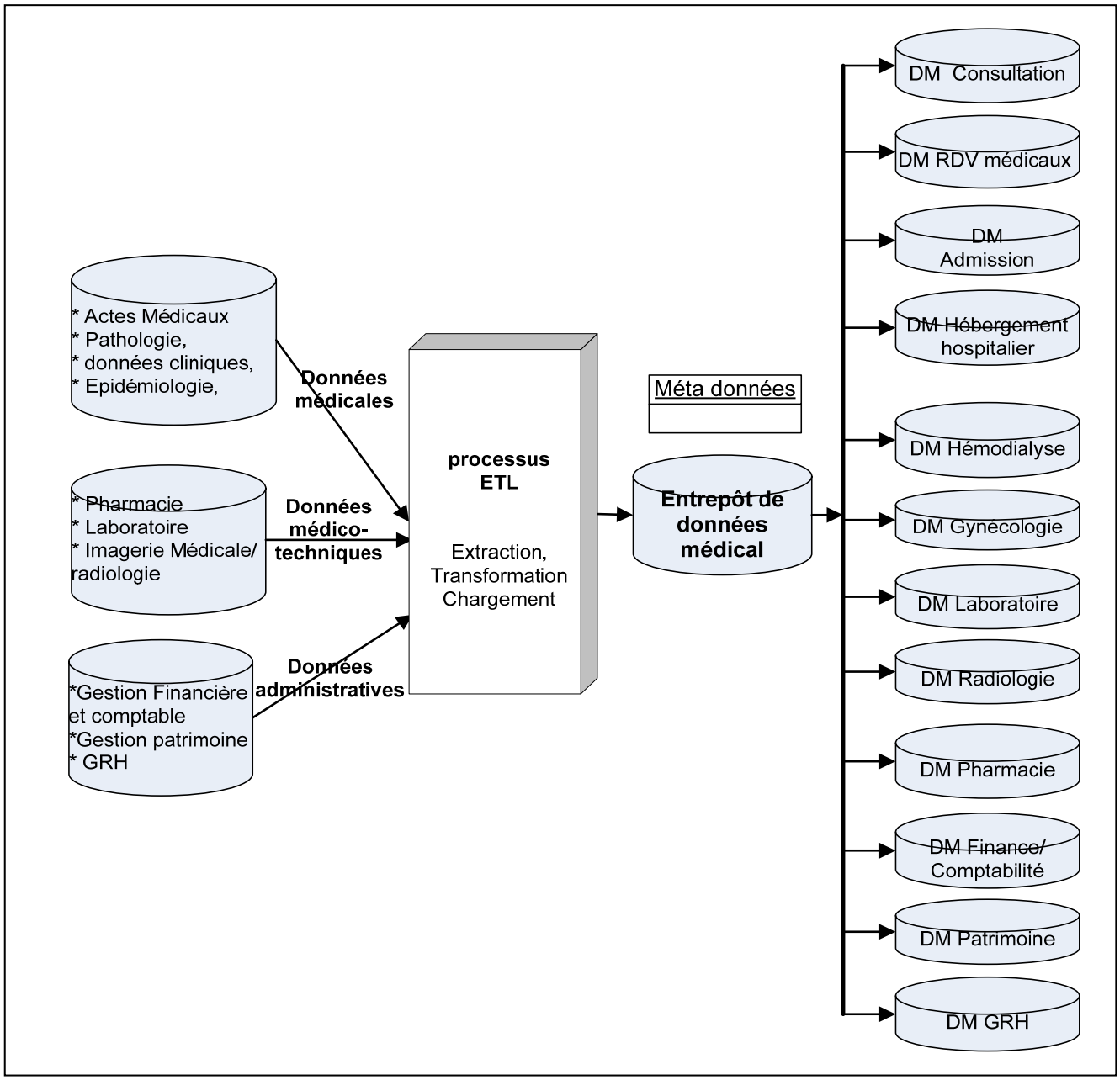

Figure 5.9 : Décomposition de l'entrepôt de données en plusieurs Data Marts.

Notre Travail est basé sur le data mart « Consultation », il vise à répondre à notre besoin d'orientation médicale des patients vers les différents services. Il regroupe trois paquetages :

1- Paquetage Consultation : contient les classes qui détaillent les données de la consultation telles que : code consultation, date consultation, les signes fonctionnels, les douleurs, les symptômes, le diagnostic et le traitement.

2- Paquetage Patient : contient les classes qui détaillent les données du patient telles que : nom, prénom, sexe, age, poids, date de naissance, adresse….

3- Paquetage Affectation : contient les classes qui détaillent l'affectation vers les services de la consultation, blocs opératoire, pré travail...

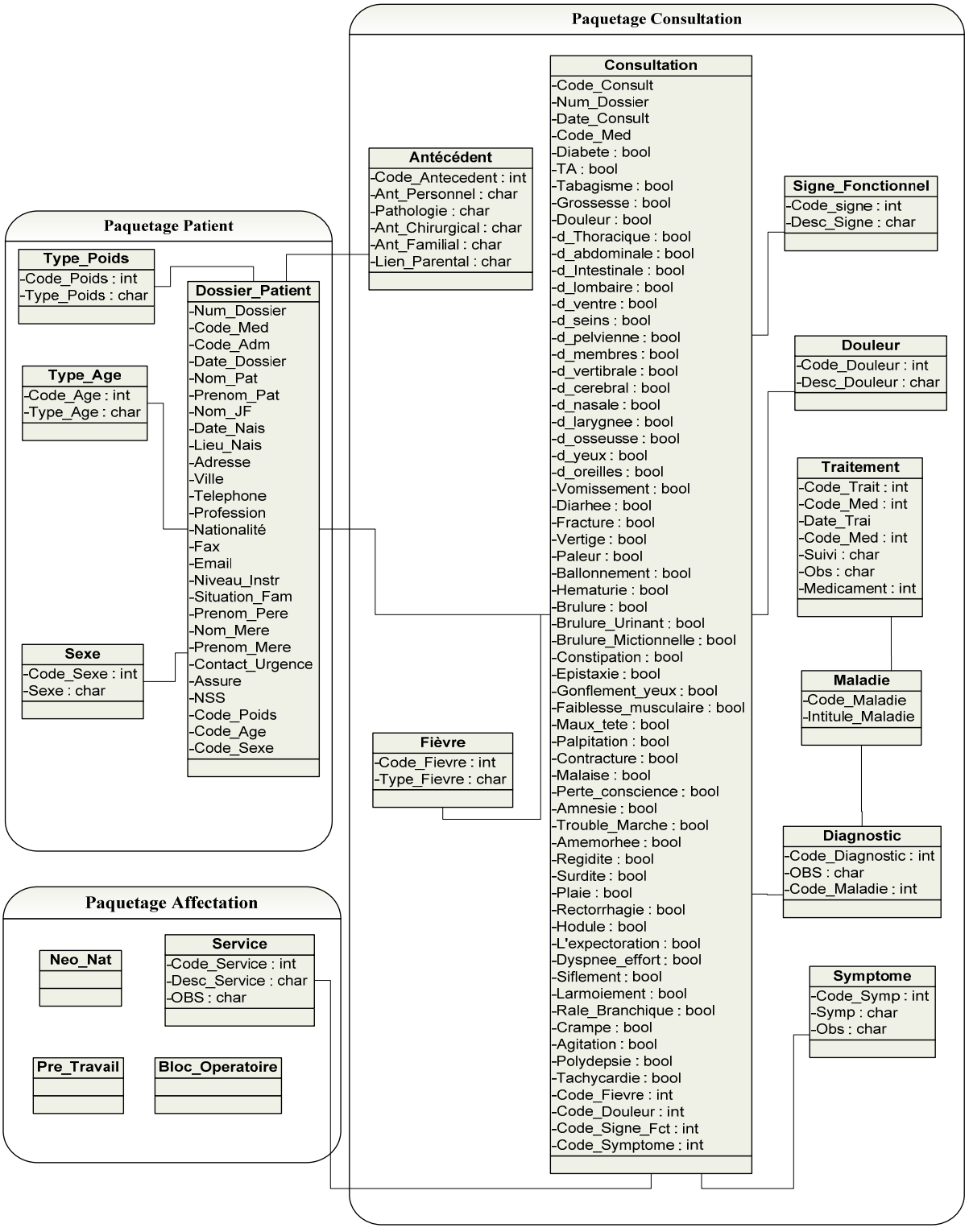

Figure 5.10 : Data Mart « Consultation ».

5.2.4.3 Modélisation multidimensionnelle de l'entrepôt

La modélisation que nous avons adoptée pour notre Data Mart est le schéma en étoile. Ce schéma est composé d'une relation de fait et de dix relations de dimensions. Nous définissons la structure de chaque relation qui intègre le schéma.

Fait :

CONSULT = {Code_Consult, Num_Dos, Date_Consult, Code_Med, Sexe, Type_Age, Fievre, Type_Fievre, Type_Poids, Diabete, TA, Tabagisme, Grossesse, Code_Douleur, d_thoracique, d_thoracique_embole_pulmonaire, d_thoracique_angine_poitrine, d_abdominale, d_abdominale_colique, d_intestinal, d_lombaire, d_ventre, d_seins, d_pelvienne, d_membres, d_osseuse, d_rapportee_membres, d_vertebrale, Epoule_douleureuse, Jambe_douleureuse, d_yeux, d_oreilles, d_cerebral, d_nasale, d_laryngee, Vomissement, Diarrheen, Fracture, Vertige, Paleur, Ballonnement, Hematurie, Besoin_uriner, Brulure, Brulure_urinant, Brulure_mictionnelle, Constipation, Dysphagie, Coliques_nephritiques, Dyspnee_paroxystique, Dyspnee_permanante, Dysurie, Epistaxis, Faiblesse_musculaire, Gonflement_yeux, Lemoptysie, Maux_tete, oeil_rouge, Palpitation, Paresies, Pouls_faible, Prurit, Rhinoree, Spasme_musculaire, Trouble_rythme_cardiaque, Trouble_vision, Amorexie , Asthenie, Dysphonie, Contracture, Malaise, Perte_conscience, Tremblement , Amnesie, Trouble_marche, Amemorhee, Metrorragie, Tachycardie, Polydepsie, Vesicule, Agitation, Rale_Branchique, Trouble_voix, Lexpectoration_purulente, Sifflement, Crampe_douleureuse, Dyspnee_effort, La_toux_quinteuse, La_vomique_purulente, Larmoiement, Les_nausees, Lexpectoration_muqueuse, Rectorrhagie, Regidite, Hodule_palpation, Surdite_perception, Plaie, Signe1, Signe2, Signe3 , Signe4, Code_service}

Dimensions :

1- **DOSSIER_PATIENT=** {NUMDOS, CODEMED, CODE_ADM, DATEDOS, NOMPAT, PRENPAT, NOMJF, CODE_AGE, AGE, SEXE, SF, DATNAIS, LIEUNAIS, ADRPAT, TELPAT, PROFPAT, NATIO, FAXPAT, EMAILPAT, NIVINST, NSS, GS, NB_GROSS, OBSERVATION, DATE_NOTE, NOTE, PRENPERE, NOM_MERE, PREN_MERE, CONTACT_URG, ORGANISME_EMPL, ADR_PARENT, ORIGINE, ORG_CARTE_GROUPAGE,

CONFIDENTIEL, RELAT_ASSURE, NOM_ASSURE, PREN_ASSURE, DATE_NAIS_ASSURE, CAISEE8ASSURE, AGENCE_ASSURE, NUM_AGENCE, EPOUSE, DATE_ARRIVE, TEL_PARENT}

2- **SERVICE** = {CODESERV, DES_SERV, OBS}

3- **AGE** = {CODEAGE, TYPE_AGE}

4- **TYPE_PIODS** = {CODE_POIDS, TYPE_POIDS}

5- **FIEVRE** = {CODE_FIEVRE, TYPE_FIEVRE}

6- **DOULEUR** = {CODEDOULEUR, DES_DOULEUR}

7- **SIGNE**1 = {CODESIGNE, DESCSIGNE, CODESERV}

8- **SIGNE2** = {CODESIGNE, DESCSIGNE, CODESERV}

9- **SIGNE3** = {CODESIGNE, DESCSIGNE, CODESERV}

10- **SIGNE4** = {CODESIGNE, DESCSIGNE, CODESERV}

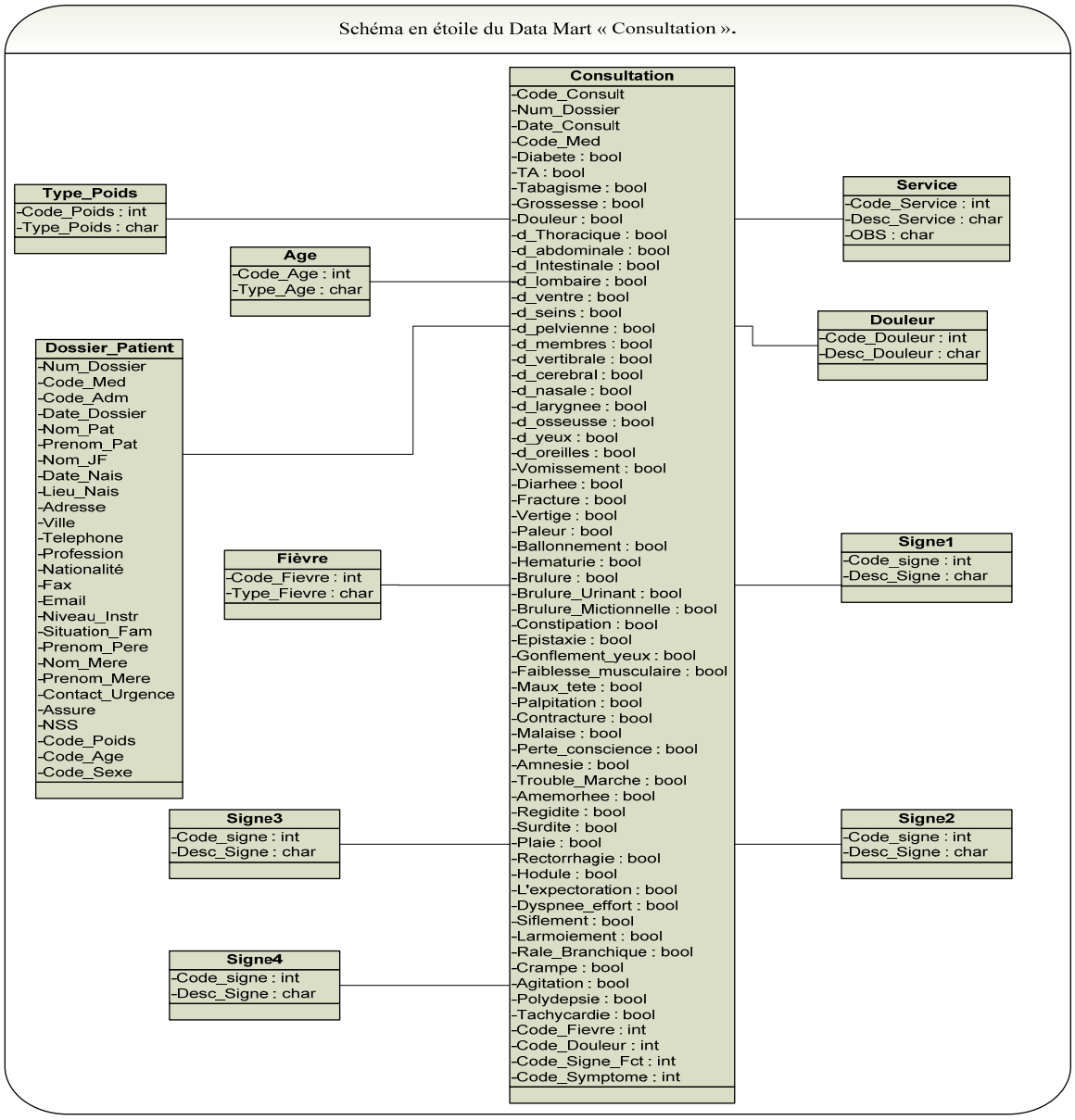

Figure 5.11 : Schéma en étoile du Data Mart « Consultation ».

Le tableau suivant contient le type et le nombre d'enregistrements de chaque relation :

Relation	**Fait/Dimension**	**Taille**
CONSULT	fait	10588
DOSSIER_PATIENT	Dimension	25200
SERVICE	Dimension	25
AGE	Dimension	6
TYPE_POIDS	Dimension	7

FIEVRE	Dimension	4
DOULEUR	Dimension	55
SIGNE	Dimension	147
SIGNE1	Dimension	147
SIGNE2	Dimension	147
SIGNE3	Dimension	147

Tableau 5.2 : Liste de relations de schéma en étoile

5.2.4.4 Interrogation de l'entrepôt de données médicales

Dans cette section, nous essayons de faire l'interrogation de l'entrepôt de données basé sur nos besoins, donc nous nous focalisons sur la génération des vues matérialisées pour construire les agrégats.

Apres le traitement du schéma en étoile illustré dans la Figure 5.11, nous avons extraire les vues suivantes :

1- Vue1 : La liste de CONSULTATION avec les dates nominée par les PATIENTS :

```
SELECT      dbo.CONSULT.CODE_CONSULT, dbo.CONSULT.DATE_CONSULT,
dbo.DOSSIER_PATIENT.NOMPAT, dbo.DOSSIER_PATIENT.PRENPAT
FROM        dbo.CONSULT INNER JOIN dbo.DOSSIER_PATIENT
ON          dbo.CONSULT.NUMDOS = dbo.DOSSIER_PATIENT.NUMDOS;
```

2- Vue2 : La liste des CONSULTATIONS par SERVICE

```
SELECT      dbo.CONSULT.CODE_CONSULT, dbo.SERVICE.CODESERV,
.           dbo.SERVICE.DESSERV
FROM        dbo.CONSULT INNER JOIN dbo.SERVICE
ON          dbo.CONSULT.CODESERV = dbo.SERVICE.CODESERV
```

3- Vue3 : La liste de CONSULTATION / PATIENT/ SEXE/AGE/ CODE_FIEVRE/ FIEVRE/TYPE_POIDS/ DIABETE/ TA/ GROSSESSE PAR SERVICE :

```
SELECT      TOP 100 PERCENT dbo.CONSULT.CODE_CONSULT,
.           dbo.DOSSIER_PATIENT.NUMDOS,
dbo.DOSSIER_PATIENT.NOMPAT,
.           dbo.DOSSIER_PATIENT.PRENPAT, dbo.DOSSIER_PATIENT.SEXE,
```

. dbo.CONSULT.CODEAGE, dbo.AGE.TYPE_AGE,
dbo.CONSULT.FIEVRE,

. dbo.FIEVRE.CODE_FIEVRE,dbo.FIEVRE.TYPE_FIEVRE,

. dbo.TYPE_POIDS.CODE_POIDS,dbo.TYPE_POIDS.TYPE_POIDS, .
dbo.CONSULT.DIABETE,dbo.CONSULT.TA,dbo.CONSULT.TABAGISME, .
dbo.CONSULT.GROSSESSE, dbo.SERVICE.CODESERV, dbo.SERVICE.DESSERV

FROM dbo.CONSULT

INNER JOIN dbo.DOSSIER_PATIENT

ON dbo.CONSULT.NUMDOS = dbo.DOSSIER_PATIENT.NUMDOS

INNER JOIN dbo.FIEVRE

ON dbo.CONSULT.CODE_FIEVRE = dbo.FIEVRE.CODE_FIEVRE

INNER JOIN dbo.TYPE_POIDS

ON dbo.CONSULT.CODE_POIDS = dbo.TYPE_POIDS.CODE_POIDS

INNER JOIN dbo.SERVICE

ON dbo.CONSULT.CODESERV = dbo.SERVICE.CODESERV

INNER JOIN dbo.AGE ON dbo.CONSULT.CODEAGE = dbo.AGE.CODEAGE

4- Vue 4 : Calcul de nombre d'apparitions des signes fonctionnels sur chaque service par la fonction « GROUP BY »:

SELECT TOP 100 PERCENT SEXE, TYPE_AGE, FIEVRE, TYPE_FIEVRE,
. TYPE_POIDS, DIABETE, TA, TABAGISME, GROSSESSE,
DES_DOULEUR, Signe1, Signe2, Signe3, Signe4, COUNT(NUMDOS) AS
Nombre_apparaission

FROM dbo.[Consult-final]

GROUP BY SEXE, TYPE_AGE, TYPE_FIEVRE, TYPE_POIDS, DES_DOULEUR,
Signe1, Signe2, Signe3, Signe4, FIEVRE, DIABETE, TA, TABAGISME,
GROSSESSE.

5.3 Extraction des Connaissances à partir des Données

Dans cette partie du système, nous allons réaliser un processus d'extraction des connaissances médicales à partir de l'entrepôt de données (data mart « consultation ») construit dans la partie précédente, le processus que nous avons retenu se décompose en plusieurs phases :

- La première phase se résume en la sélection des données dont l'exploitation permet de répondre à la problématique.
- La seconde phase concerne la préparation des données sélectionnées lors de la phase précédente afin qu'elles soient facilement exploitées par les méthodes du data mining.
- La phase trois correspond à l'utilisation d'une ou des méthodes intelligentes du data mining appliquées sur les données traitées. Cette phase est le coeur du processus d'ECD. Notre travail consiste à appliquer les règles d'associassion dans le but d'extraire les connaissances pertinentes qui répondent à nos besoins.
- La dernière phase consiste à valider et à évaluer les modèles, les connaissances (les règles déduites) de la phase précédente, ces connaissances seront exploitées pour la construction d'une Base de cas pour la partie raisonnement à base de cas.

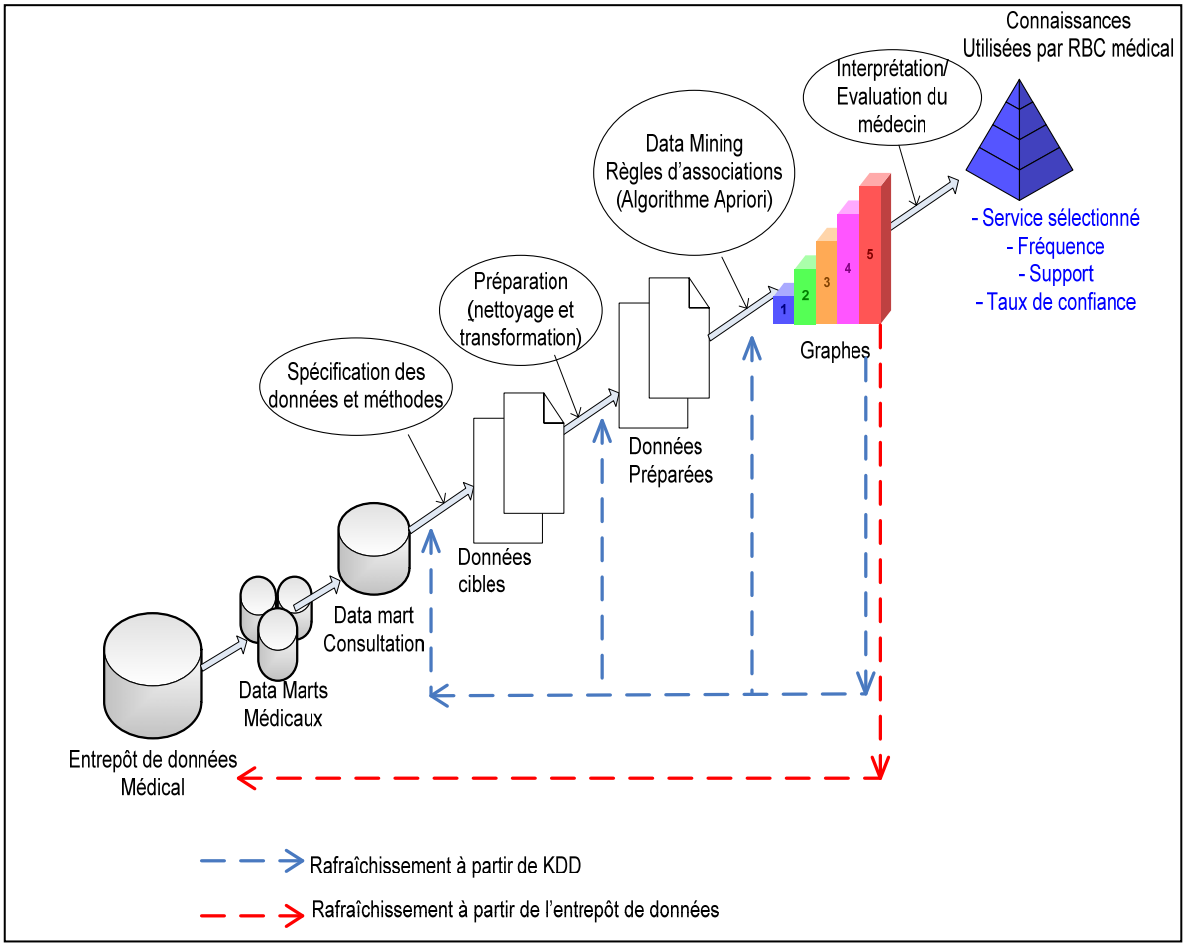

Figure 5.12 : Processus d'Extraction des connaissances à partir de données médicales.

5.3.1 La sélection des données : cette phase se résume en la sélection des données dont l'exploitation permet de répondre à notre problématique.

5.3.1.1 Spécification des données

La spécification des données nous permet de déterminer l'ensemble des données et de faire le choix des tables qui se trouvent dans l'entrepôt de données médicales et qui seront utiles pour la résolution de notre problème.

Dans la partie intégration des données nous avons décomposé notre entrepôt de données en plusieurs data marts. Le data mart « Consultation » contient les données qui résoud notre problème.

5.3.1.2 Spécification de méthodes de data mining

Cette étape permet de définir la ou les familles de méthodes de data mining qui peuvent être exécutées dans le processus ECD, telles que : les méthodes de classification, les méthodes de segmentation, les méthodes de règles d'association, etc.

La technique de *règles d'association* est la méthode appropriée pour répondre à notre objectif, l'algorithme *Apriori* est la solution adéquate pour résoudre notre problème avec des améliorations et des changements.

5.3.1.3 Spécification de la Mesure

La mesure permet d'indiquer le degré de fiabilité d'une connaissance extraite. Cette mesure a pour conséquence de permettre de distinguer parmi les connaissances extraites, celles qui sont intéressantes.

Pour la méthode de règles d'association, la fonction de mesure intéressante que nous avons suivi est la fréquence relative d'apparition d'une règle dans la base de données par apport à d'autres règles.

5.3.1.4 Représentation des résultats du data mining

Ceci réfère à la forme sous laquelle le résultat du data mining sera présenté. On peut choisir différentes techniques de visualisation (graphe, tableur, cube, etc.). Dans notre système les connaissances extraites seront représentées sous forme d'histogrammes.

5.3.1.5 Représentation de la connaissance extraite

La connaissance extraite peut être représentée sous la forme d'une règle, arbre ou bien table d'une base de données, Notre solution consiste à stocker les connaissances obtenues à partir du processus de data mining sous forme d'enregistrements dans une base de données relationnelle.

5.3.2 La Préparation des données

Les données à analyser par les méthodes de data mining sont parfois incomplètes, inconsistantes, erronées, incompatibles entre elles, inadaptées ou encombrantes. Ces types de données sont courants et se retrouvent régulièrement dans les bases de données et d'entrepôts de données.

Dans cette étape, plusieurs procédures sont nécessaires et chacune d'entre-elles a des tâches bien précises dans le traitement des données médicales :

5.3.2.1 La procédure de nettoyage des données

Elle permet de compléter les données manquantes et de régulariser les données erronées et inconsistantes.

Nous avons procédé à Plusieurs méthodes qui permettent d'accomplir les données manquantes dans l'entrepôt de données médical. Le choix de chaque méthode dépend des données :

- Ignorer les instances incomplètes: dans le cas ou les classes des instances ne sont pas indiquées dans l'entrepôt de données, la solution est d'ignorer totalement cette données, par exemple : le champ « Adresse » dans la table « patient » n'est pas renseigné et cette données est inutile dans notre étude.
- Compléter les données manuellement: Il y a des données importantes dans notre étude mais elles sont incomplètes dans l'entrepôt de données dues aux problèmes de la saisie, la solution est de compléter la saisie de ces données manuellement, par exemple le champ « sexe » dans la table « patient » n'est pas renseigné dans un enregistrement, on le rempli manuellement on se basant sur le prénom.
- Compléter les données incomplètes à l'aide d'une fonction ou d'une constante globale:

Par exemple: le champ « Age » d'un patient n'est pas défini dans une consultation de pédiatrie, donc on le rempli par :

Type Age=Enfant

5.3.2.2 Procédure de transformation

Cette procédure est déjà réalisée dans la construction de l'entrepôt de données, nous avons regroupé des données saisies dans les différentes tables du data mart « Consultation » par des méthode d'agrégation dans une seule vue matérialisée, Ceci permet d'avoir une vue d'ensemble de toute les tables et de modeler les données sous une forme exploitable.

5.3.3 Le Data Mining

La fouille de données (*data mining*), est le coeur du processus d'ECD. Il s'agit à ce niveau de trouver des connaissances à partir des données. Le travail consiste à appliquer des méthodes intelligentes dans le but d'extraire cette connaissance. Tout le problème de la

fouille de données réside dans le choix de la méthode adéquate à un problème donné. La méthode que nous avons optée dans notre étude est les règles d'association.

5.3.3.1 Règles d'association

Les règles d'associations sont traditionnellement liées au secteur de la distribution car leur principale application est l'analyse du panier de la ménagère qui consiste en la recherche d'association entre les produits présents sur chaque ticket de caisses. L'objectif de la méthode est d'étudier les achats des clients pour obtenir des informations sur leur profil et les tendances des clients. Toutefois, cette méthode peut être généralisée à tout secteur d'activité pour lequel il est pertinent de s'intéresser au regroupement de produits ou de services: services bancaires, services de télécommunication, services médicaux, etc.

Dans notre cas la technique les règles d'associations est appliquée pour la recherche des règle : ensemble (signes fonctionnels) → service approprié.

5.3.3.2 Choix de l'algorithme : *Algorithme Apriori*

Notre data mart de données contient une masse très importante de données, et notre objectif est de découvrir la connaissance pertinente et utile pour pouvoir orienter les patients vers les différents services.

Chaque enregistrement stocké dans le data mart « Consultation » contient les données d'un patient :

1- Identification : nom, prénom, age, sexe, poids, adresse…

2- Signes fonctionnels du patient : fièvre, diabétique, tension artérielle, la liste des douleur, la liste des autres signes…

3- Données administratives : code consultation, numéro de dossier, le médecin traitant, date de consultation, Service de consultation…

Notre algorithme a pour principe la recherche de règles intéressantes parmi toutes les règles de type : *Si X et Y alors Z.*

Pour ce faire, il effectue le calcul des critères globaux : fréquence, support et le taux de confiance de tous les ensembles des enregistrements qui existent dans le data mart de données. Ainsi, un enregistrement ayant un taux de confiance supérieur à un seuil est qualifié de fréquent.

(a) La fréquence

La *fréquence* d'une règle correspond au nombre d'apparitions simultanée des signes fonctionnels d'un enregistrement sans tenir compte du service de consultation dans le data mart :

- ***Fréquence = freq(condition)***
- ***Fréquence= freq (signe fonctionnel$_1$, signe fonctionnel$_2$,… signe fonctionnel $_n$) (VII)***

(b) Le Support

Le *support* d'une règle correspond à la fréquence simultanée d'apparition des signes fonctionnels d'un enregistrement qui figurent dans la condition et qui donnent le même service de consultation qui correspond au résultat :

- ***Support = freq(condition et résultat).***
- ***Support =freq (signe fonctionnel$_1$, signe fonctionnel$_2$,… signe fonctionnel$_n$ et Service $_k$) (VIII)***

❖ La somme des supports qui contiennent la même condition = la fréquence de la condition :

Support $_1$ =freq (signe fonctionnel $_1$, signe fonctionnel $_2$,… signe fonctionnel $_n$ et Service $_1$)

Support $_2$ =freq (signe fonctionnel $_1$, signe fonctionnel $_2$,… signe fonctionnel $_n$ et Service $_2$)

……

Support $_k$ =freq (signe fonctionnel $_1$, signe fonctionnel $_2$,… signe fonctionnel $_n$ et Service $_k$)

Fréquence= Freq((signe fonctionnel $_1$, signe fonctionnel $_2$,… signe fonctionnel $_n$)

Fréquence = Support $_1$ + Support $_2$ +… Support $_k$

(c) le taux de confiance

La *confiance* est le rapport entre le support et la fréquence, Soit :

- ***Confiance = freq(((signe fonctionnel$_1$, signe fonctionnel $_2$,… signe fonctionnel $_n$) et Service$_k$)/ freq(signe fonctionnel $_1$, signe fonctionnel $_2$,… signe fonctionnel $_n$)***
- ***Confiance k= Support k / fréquence (IX)***

Confiance $_1$=Support $_1$ / Fréquence

Confiance $_2$=Support $_2$ / Fréquence

……

Confiance $_k$=Support $_k$ / Fréquence

Confiance $_1$+ Confiance $_2$+… Confiance $_k$= 1

La liste des Attributs qui constitues La partie condition de la règle d'association:

- ✓ Sexe : F (féminin), M (Masculin)
- ✓ Type d'Age

N° Ordre	Type Age	
1	Nouveau Né	Age < 1 an
2	Enfant	1 <= Age<15 ans
3	Jeune	15 <= Age<25 ans
4	Adulte	25 <= Age<55 ans
5	Agé	55 <= Age<70 ans
6	Très Agé	70<= Age

Tableau 5.3 : La liste de type d'age.

- ✓ Fièvre : Oui/ Non
- ✓ Type de fièvre

N° Ordre	Type Fièvre
0	ND (Non défini)
1	aigu
2	progressif
3	insidieux

Tableau 5.4 : La liste de type de fièvre.

- ✓ Type de poids

N° Ordre	Type Poids
0	ND (non défini)
1	poids normal
2	obésité
3	oedème
4	myxoedème
5	amaigrissement
6	déshydratation

Tableau 5.5 : La liste de type de poids.

- ✓ Diabète : Oui/ Non
- ✓ Tension Artérielle : Oui/ Non
- ✓ Tabagisme : Oui/ Non
- ✓ Grossesse : Oui/ Non
- ✓ Douleur : Oui/ Non
- ✓ Liste des douleurs

N° d'ordre	Description Douleur	Oui/Non
1	douleur thoracique	Oui/Non
2	douleur thoracique d'angine de poitrine	Oui/Non
3	douleur thoracique embole pulmonaire	Oui/Non
4	douleur abdominale	Oui/Non
5	douleur abdominale colique	Oui/Non
6	douleur intestinale	Oui/Non
7	douleur lombaire	Oui/Non
8	douleur de ventre	Oui/Non
9	douleur des seins	Oui/Non
10	douleur pelvienne	Oui/Non
11	douleur des membres	Oui/Non
12	douleur osseuse	Oui/Non
13	douleur rapportée des membres	Oui/Non
14	douleur vertébrale	Oui/Non
15	Epaule douloureuse	Oui/Non
16	Jambe douloureuse	Oui/Non
17	douleur des yeux	Oui/Non
18	douleur des oreilles	Oui/Non
19	douleur cérébrale	Oui/Non
20	douleur nasale	Oui/Non
21	douleur laryngée	Oui/Non

Tableau 5.6 : La liste des douleurs.

- ✓ Liste des signes :

N° d'ordre	Description Signe	Oui/Non
1	Besoin fréquent d'uriner	Oui/Non
2	Brûlure	Oui/Non
3	Brûlure en urinant	Oui/Non
4	Brûlure mictionnelle	Oui/Non
5	Coliques néphrétiques	Oui/Non
6	Constipation	Oui/Non
7	Crampe	Oui/Non
8	Diarrhée	Oui/Non
9	Dysphagie	Oui/Non

10	Dyspnée d'effort	Oui/Non
11	Dyspnée paroxystique	Oui/Non
12	Dyspnée permanente	Oui/Non
13	Dysurie	Oui/Non
14	Epistaxis	Oui/Non
15	Faiblesse musculaire	Oui/Non
16	Fracture	Oui/Non
17	Gonflement des yeux	Oui/Non
18	Hématurie	Oui/Non
19	La pâleur	Oui/Non
20	La toux quinteuse	Oui/Non
21	La vomique purulente	Oui/Non
22	Larmoiement	Oui/Non
23	L'hémoptysie	Oui/Non
24	Les nausées	Oui/Non
25	L'expectoration muqueuse	Oui/Non
26	L'expectoration purulente	Oui/Non
27	Maux de tété	Oui/Non
28	Oeil rouge	Oui/Non
29	Palpitation	Oui/Non
30	Parésies	Oui/Non
31	Pouls faible	Oui/Non
32	Prurit	Oui/Non
33	Rhinorée	Oui/Non
34	Sifflement	Oui/Non
35	Spasme musculaire	Oui/Non
36	Trouble de la voix	Oui/Non
37	Trouble de rythme cardiaque	Oui/Non
38	Trouble de vision	Oui/Non
39	Vertige	Oui/Non
40	Vomissement	Oui/Non
41	Anorexie	Oui/Non
42	Asthénie	Oui/Non
43	Dysphonie	Oui/Non
44	Contracture	Oui/Non
45	Ballonnement	Oui/Non
46	Râle Bronchique	Oui/Non
47	Malaise	Oui/Non
48	Rectorhagie	Oui/Non
49	Perte de conscience	Oui/Non
50	Tremblement	Oui/Non
51	Amnésie	Oui/Non
52	Trouble de marche	Oui/Non
53	Rigidité	Oui/Non
54	Amemorhée	Oui/Non
55	Metrorragie	Oui/Non
56	Hodule a la palpation	Oui/Non
57	Tachycardie	Oui/Non

58	Surdité de perception	Oui/Non
59	polydipsie (Soif intense)	Oui/Non
60	Vésicule	Oui/Non
61	Plaie	Oui/Non
62	Agitation	Oui/Non

Tableau 5.7 : La liste des signes.

✓ Liste autre signe 1 :

N° d'ordre	Signe fonctionnel 1
1	Affection des organes génitaux
2	Affection des seins
3	Hémorragie génitale
4	L'hématémèse
5	Urine foncé (oligurie)
6	Urine rouge
7	Vergeture
8	Saignement

Tableau 5.8 : La liste des autres signes «1».

✓ Liste autre signe2 :

N° d'ordre	Signe fonctionnel 2
1	Angine de poitrine d'effort
2	Eternuement
3	La toux émétisante
4	La toux humide
5	L'expectoration hémoptysique

Tableau 5.9 : La liste des autres signes «2».

✓ Liste autre signe3 :

N° d'ordre	Signe fonctionnel 3
1	Otorrhées
2	Perte de mémoire
3	Rétention vésicale
4	Trouble de transit intestinal
5	Voire double
6	Dyskinésie
7	Retard statural
8	Purpura

Tableau 5.10 : La liste des autres signes «3».

✓ Liste autre signe 4 :

N° d'ordre	Signe fonctionnel 4
1	Anémie
2	Déshydratation
3	Oedème

Tableau 5.11 : La liste des autres signes «4».

Le « Service » et ses critères globaux représentent la partie résultat.

5.3.4 Evaluation et présentation des résultats

➢ Cette phase est constituée de l'évaluation, qui mesure l'intérêt des motifs extraits, ce n'est qu'à partir de cette phase qu'on peut employer le terme de *connaissance* à condition que ces motifs soient validés par les experts du domaine. Il y a principalement deux techniques de validation qui sont la technique de validation statistique et la technique de validation par expertise. l'interaction avec l'expert du domaine est optée dans notre système. La validation par expertise, est réalisée par un expert du domaine (médecin) qui jugera la pertinence des résultats produits (la pertinence des règles d'association)

➢ La phase de présentation consiste à interpréter les résultats à l'utilisateur grâce à différentes techniques de visualisation. Les résultats obtenus du data mining sont représentés sous forme d'un histogramme.

5.4 Raisonnement à Base de Cas

Dans cette partie, nous présentons notre système de raisonnement à base de cas d'orientation médicale qui aide l'utilisateur à orienter le patient vers le service adéquat, sa base de connaissance est le résultat de la partie précédente.

5.4.1 Construction de la Base de cas

La structure des enregistrements, de forme (attribut, valeur) dans un SGBD s'apparente bien avec la forme des caractéristiques dans une base de cas. Notre solution consiste à stocker les cas obtenus à partir du processus de data mining dans une base de données relationnelle. Les cas sont alors stockés dans une table de la base de données, chaque attribut correspondant à un champ, chaque cas, à un enregistrement.

5.4.2 Connaissances de cas

Les connaissances de *cas*, dans le cadre de la prise en charge de l'orientation médicale, sont celles des cas cliniques déjà stockés dans notre *base de cas*. Trois éléments principaux apparaissent dans le contenu du cas :

1- *Le but :* représente l'objectif que tentera d'accomplir la solution, il constitue l'orientation d'un patient vers un service.

2- Les caractéristiques : quelque soit la représentation des connaissances au niveau « symbole » choisie, la caractérisation est décrite par un ensemble fini de couples <attribut, valeur>. Il s'agit de donner le numéro d'ordre de la valeur. La valeur *i* correspond à la $i^{ème}$ valeur de l'attribut telle que définie précédemment :

Exemple :

- ✓ **Sexe** = 1 pour masculin et 2 pour féminin.
- ✓ **Type de poids** = 0 pour ND, 1 pour poids normal, 2 pour obésité, 3 pour oedème, 4 pour myxoedème, 5 pour amaigrissement, 6 pour déshydratation.
- ✓ Trouble de marche = 0 pour Non, 1 pour Oui

3- La solution au problème : dépend bien de l'utilité du cas dans le raisonnement. Dans notre travail la solution est le « *service médical* » et *« le taux confiance »* du service par rapport au cas, ainsi que « *la fréquence »* du cas et *« le support* » du cas.

- ➢ **Liste des services médicaux :**

N° Ordre	**Service**
1	Pneumologie
2	Service Cardio-vasculaire
3	Chirurgie vasculaire et thoracique
4	Gastro-entérologie
5	Neurochirurgie
6	Neurologie
7	Urologie
8	Gynécologie
9	Hématologie
10	Traumatologie/ orthopédie
11	Néphrologie

12	Ophtalmologie
13	ORL
14	Dermatologie
15	Service des maladies infectieuses
16	Chirurgie générale
17	Pédiatrie
18	NEO NATAL
19	Chirurgie Infantile CCI
20	Endocrinologie
21	Rhumatologie
22	Psychiatre
23	Anesthésie /Réanimation
24	Sérologie
25	Pavillon des urgences

Tableau 5.12 : La liste des services médicaux.

5.4.3 Processus du raisonnement à base de cas

Le fonctionnement de cette partie de notre travail repose sur les quatre parties qui composent les systèmes CBR: partie recherche, partie adaptation, partie révision et la partie de mémorisation:

5.4.3.1 Partie recherche

Lors de la présentation d'un nouveau cas, cette phase permet de déterminer les cas de la base qui sont les plus similaires au problème donné du patient, après l'extraction des indices connus qui vont servir à effectuer la recherche de cas analogues.

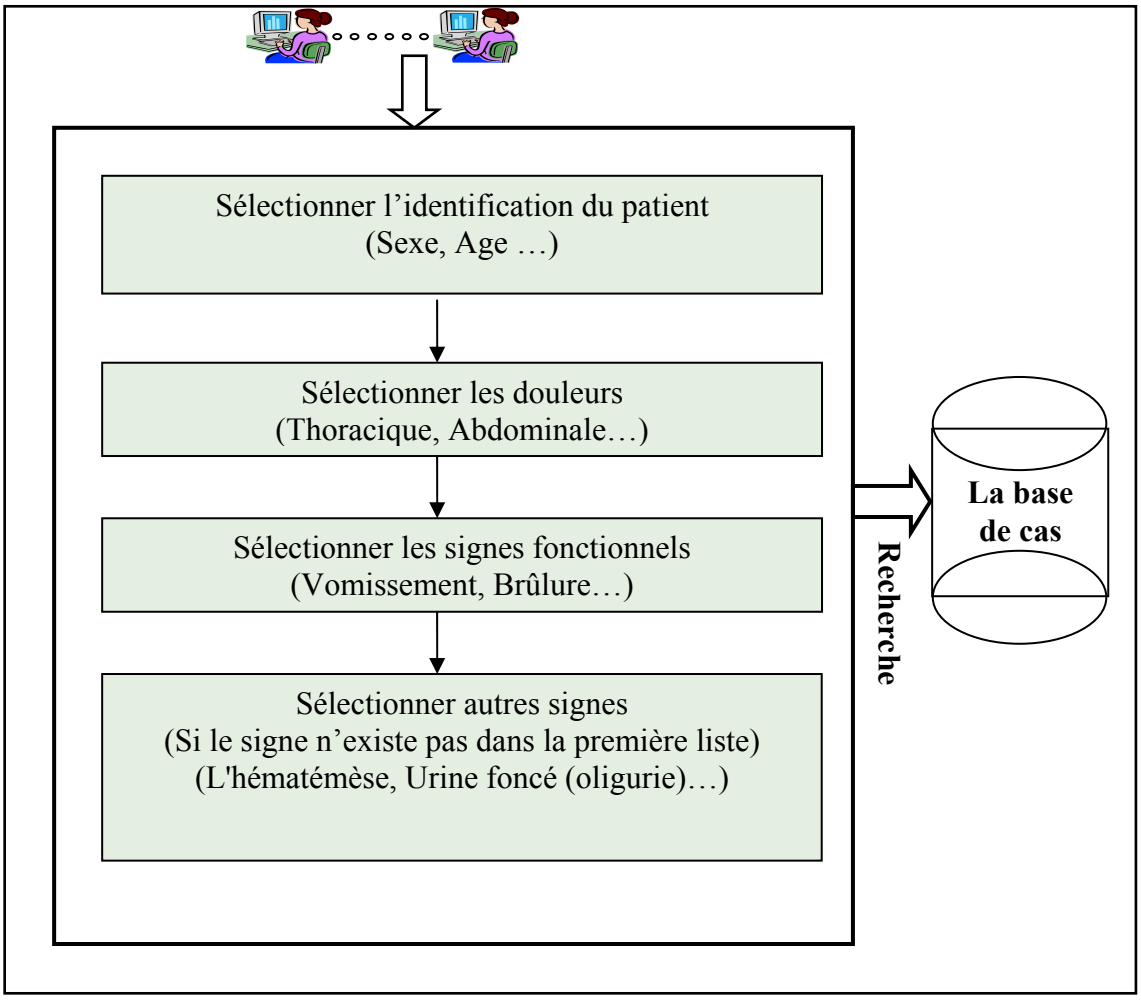

Figure 5.13: Le processus de sélection des attributs

Après la sélection des attributs de cas cible, le système recherche les cas analogues qui se trouvent dans la base de cas par le calcul de similarité.

Types d'attributs

Il existe deux types d'attributs :

- Les attributs à valeurs discrètes (Sexe, type d'age, type de poids, type de fièvre, signe1, signe2, signe3, signe4…)
- Les attributs à valeurs booliennes (diabète, tabagisme, tension artérielle, liste des douleurs, liste des signes fonctionnels…).

Calcul de similarité locale

✓ Pour les attributs à valeurs discrètes, la similitude entre deux attributs est calculée en utilisant l'équation (1) (c'est-à-dire de types disjonctifs et catégoriques et nominaux).

$$\textbf{Sim (ai, bi)} = \begin{cases} \textbf{1 si ai = bi} \\ \textbf{0 si ai} \neq \textbf{bi} \end{cases} \quad \textbf{(1)}$$

✓ Pour les attributs à valeurs booliennes, la similitude entre deux attributs est calculée en utilisant l'équation (2) :

$$\textbf{Sim (ai, bi)} = \begin{cases} \textbf{1 si ai = bi} \\ \textbf{0 si ai} \neq \textbf{bi} \end{cases} \quad \textbf{(2)}$$

Calcul de similarité globale

La similarité globale de deux cas *C1* et *C2* est calculée par la formule suivante qui présente la distance euclidienne détaillée dans le chapitre précèdent:

$$d_2(x,y) = \left(\sum_{i=1}^{k} w_i \times \left| x_i - y_i \right|^2\right)^{1/2} \quad \text{(distance euclidienne)}$$

Où w_i est le i^{ème} valeur du vecteur de poids *W* et Sim_i est la i^{ème} similarité locale selon un ordre croissant des attributs.

$$\text{Sim } (C1, C2) = \sum_{i=1}^{k} w_i \, \text{Sim}(a_i, b_i) \qquad \text{(X)}$$

La valeur de poids est calculée par le data mining, donc on considère que le vecteur W est égal à 1 (w_i=1),

$$\text{Sim } (C1, C2) = \left[\sum_{i=1}^{k} \text{Sim}_i \, (a_i, b_i)\right] / \, n \qquad \text{(XI)}$$

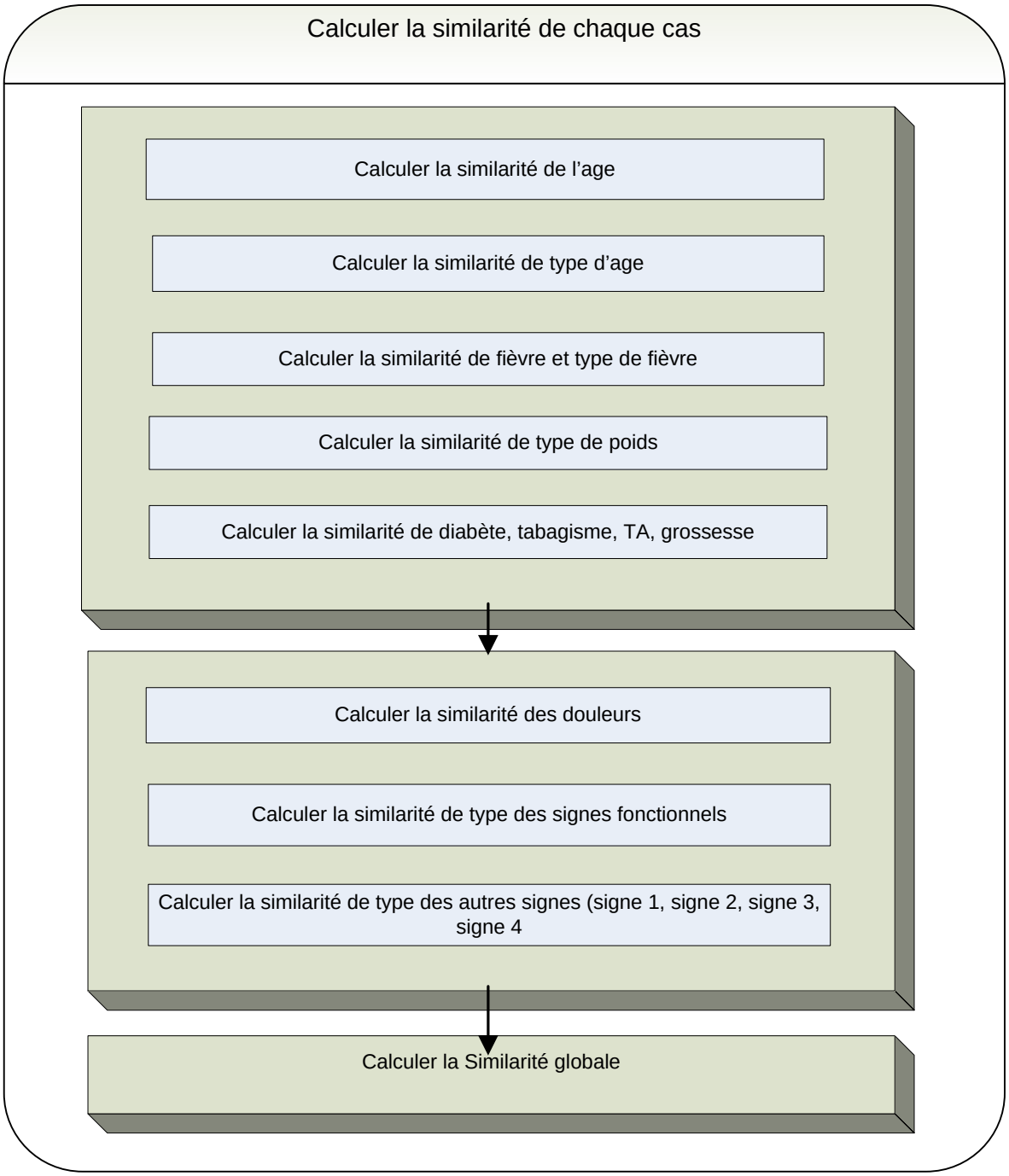

Figure 5.14 : Les étapes de calcul la similarité entre les cas

Sélection des cas similaires

La procédure de recherche est implantée par une sélection des plus proches voisins ("k-proches voisins "). Les cas sont classés selon leur mesure de similarité (degré) et le cas qui a un degré de similarité supérieur au seuil précisé par l'expert est proposé comme solution. Dans notre travail la partie du raisonnement à base de cas est basé sur la partie précédente du data mining et le problème de définition des degrés de similarité des cas est résolus, on

considère que le degré de chaque cas dans le raisonnement à base de cas est le taux de confiance extrait par la partie précédente.

5.4.3.2 Partie adaptation

Cette partie n'est pas prise en charge dans notre système, pour cela nous allons la proposer comme un travail futur dans les perspectives.

5.4.3.3 Partie révision

La révision consiste à évaluer la solution proposée en la testant dans un environnement réel. Dans notre système la révision consisterait à tester si le service médical proposé par le système est efficace, elle est faite par une intervention humaine.

5.4.3.4 Partie mémorisation

Si après le calcul des mesures de similarité entre le nouveau cas et ceux de la base, le système génère un message qu'il n'existe pas de cas similaire ou bien le degré de similarité des cas similaires sont inférieurs au seuil proposé par l'expert, ce dernier a accès au système pour insérer le nouveau cas. Il doit prendre en compte tous les attributs du cas de la base tels que (le sexe du patient, l'age, le poids, les signes et douleurs), ainsi que les index appropriés tels que le service proposé et le taux de confiance du service (dans le cas du CBR le taux de confiance est le degré de similarité de la solution).

5.5 Conclusion

Nous avons présenté dans ce chapitre notre approche pour résoudre le problème d'orientation des patients vers les services médicaux. Cette approche repose sur la gestion des connaissances dans le domaine médical.

Dans le chapitre suivant nous allons essayer de valider cette approche sur la base de cas résultant du data mining.

CHAPITRE 6

VALIDATION ET DISCUSSION DES RESULTATS

6.1 Introduction

Le présent chapitre décrit les phases de développement du système proposé, il présente le processus de développement utilisé et décrit les outils réalisés avec ce système médical. Nous montrons d'abord une vision globale de son architecture en insistant sur la structure et les données de l'entrepôt de données, puis la construction du module data mining avec les règles d'association (les résultats de l'algorithme **Apriori**), enfin l'exploitation et l'exploration des résultats pour le module du raisonnement à base de cas.

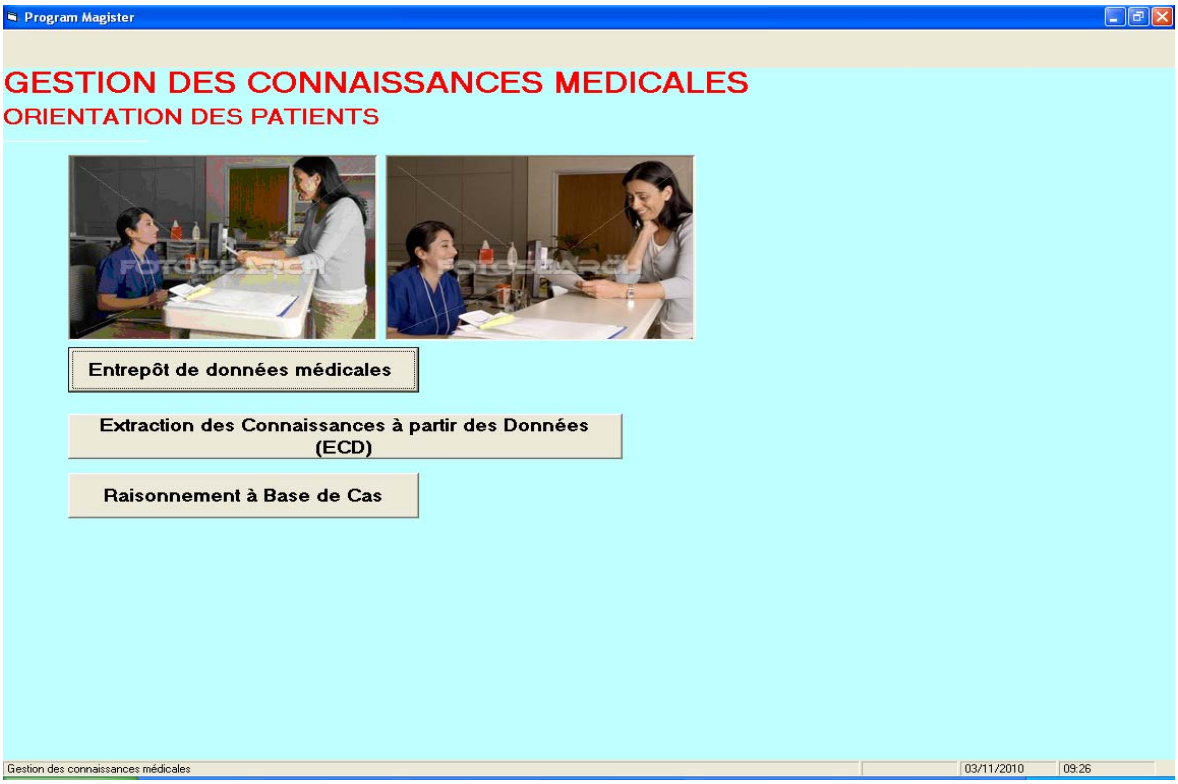

Figure 6.1 : Fenêtre principale de l'application « Gestion des connaissances médicales ».

6.2 Entrepôt de données médicales

L'entrepôt de données médicales construit stocke physiquement les données des sources réparties dans un système SGBD **« SQL Server »**,

6.2.1 Les méta-données

Dans ce module nous avons fait une visualisation de la structure de l'entrepôt qui présente les **méta-données** de l'entrepôt, ces méta-données constituent une véritable aide permettant de connaître l'information contenue dans l'entrepôt de données. Par exemple la liste de toutes les tables, les champs de chaque table et les caractéristiques de chaque champ (Null, Type, Précision, Echelle, Max caractères, Index…).

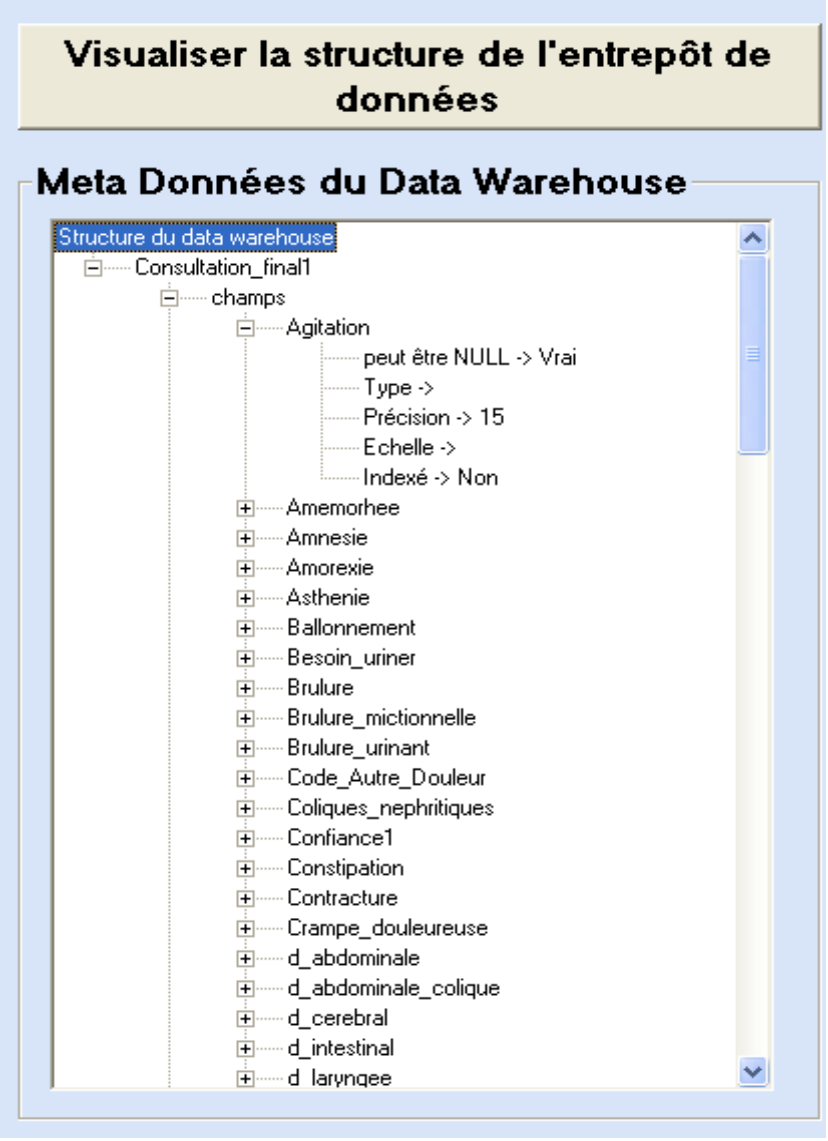

Figure 6.2 : Méta données de l'entrepôt de données médicales.

6.2.2 Interrogation de l'entrepôt de données médicales

Pour préparer les données de l'entrepôt nous avons calculé des agrégats représentées et stockées sous formes des vues matérialisées, ces données pourront être utilisées par le système lors du traitement d'autres vues.

Nous avons procédé au précalcul sélectif qui se base sur l'observation qu'il y a des agrégats qui peuvent servir à en calculer d'autres.

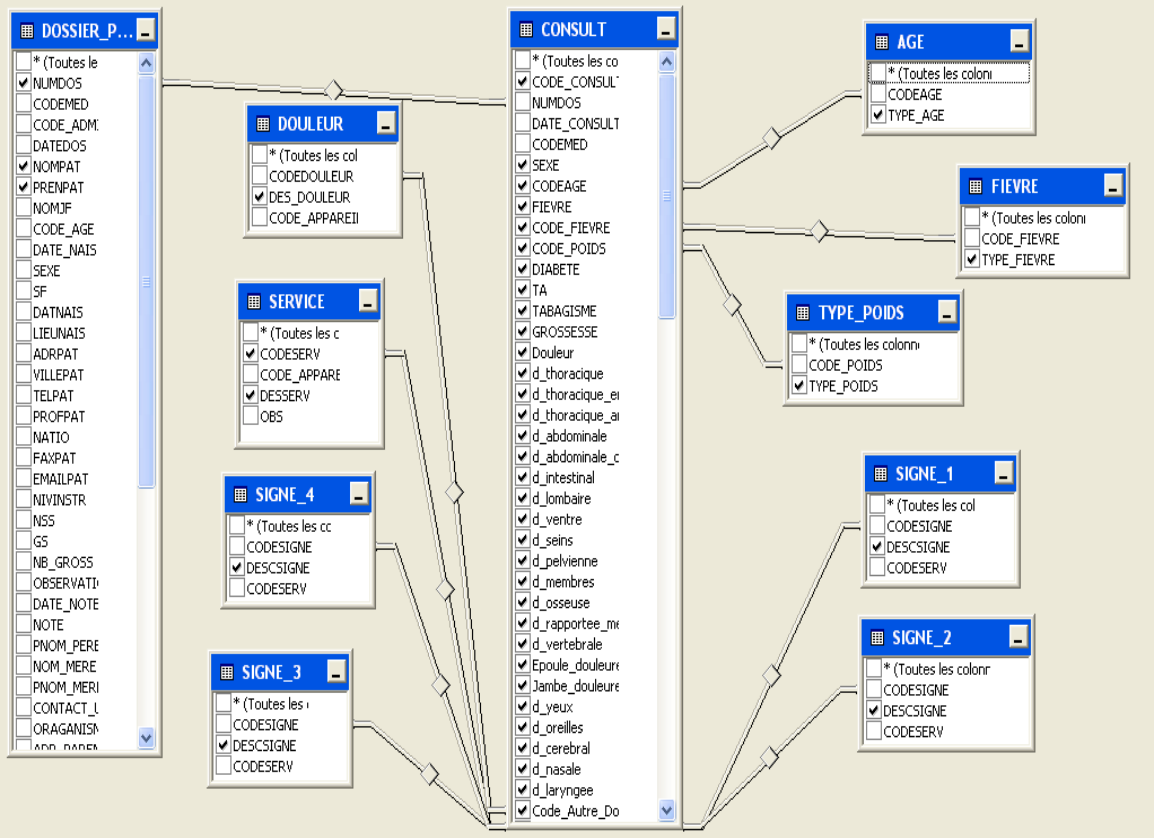

Figure 6.3 : Vue matérialisée « consultation ».

Le résultat final de l'interrogation est la construction de la vue « consultation » présentée dans la figure 6.3. Dans cette vue nous avons considéré une centaine d'attributs avec plus de 10588 enregistrements.

6.3 Extraction des connaissances à partir des données

Dans ce module, nous avons développé la partie de l'extraction des connaissances à partir de données de l'entrepôt, avec l'implémentation de l'algorithme ***Apriori***.

6.3.1 Implémentation de l'algorithme Apriori

Les résultats de cet algorithme donnent 163 règles d'association sans considération de la partie ***résultat*** de la règle, et 327 règles avec considération des deux parties ***condition et résultat*** de la règle (précision du service).

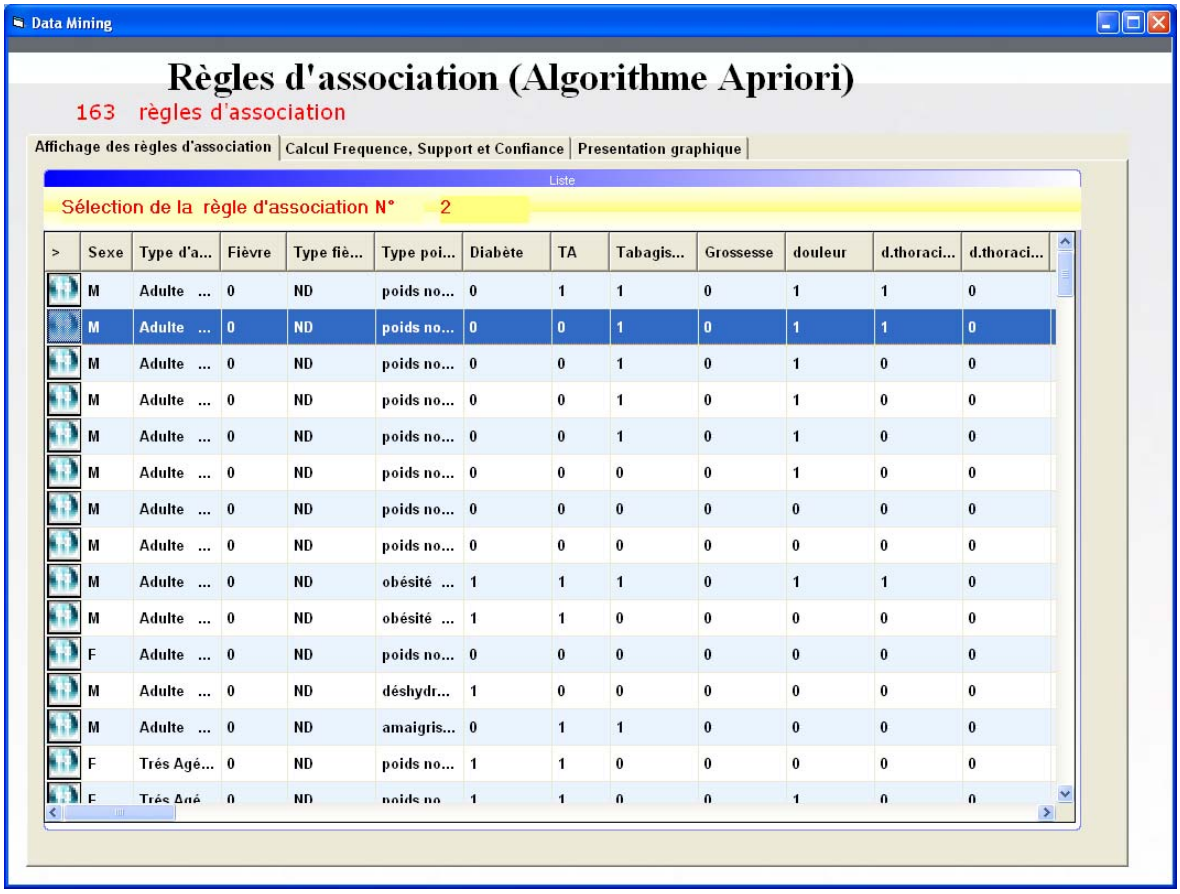

Figure 6.4 : Sélection d'une règle d'association N° 2.

Cette figure présente la partie ***condition*** des règles d'association obtenues de l'algorithme ***Apriori***.

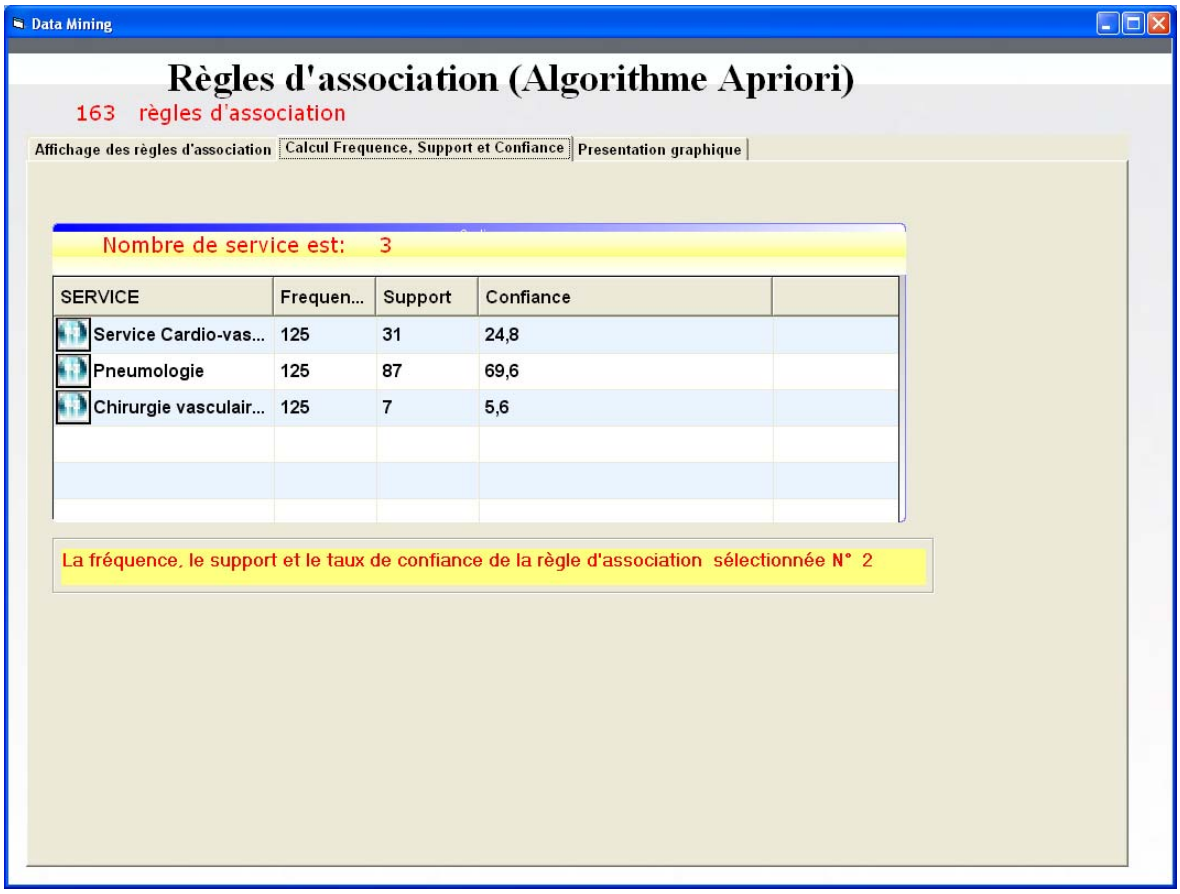

Figure 6.5 : La fréquence, le support et le taux de confiance de la règle sélectionnée.

La sélection de la règle N° 2, donne une fréquence de 125, avec trois services différents :

1- Service cardio-vasculaire (support =31, taux de confiance=24.8) ;

2- Service pneumologie (support =87, taux de confiance=69.6) ;

3- Service chirurgie vasculaire (support =7, taux de confiance=5.6).

6.3.2 Evaluation et présentation des résultats

Dans notre système les résultats obtenus du data mining sont interprétés sous forme d'histogramme qui présente l'ensemble des services dans l'axe des abscisses et les taux de confiance dans l'axe des coordonnées.

La validation du modèle présenté est effectuée par un expert (le médecin).

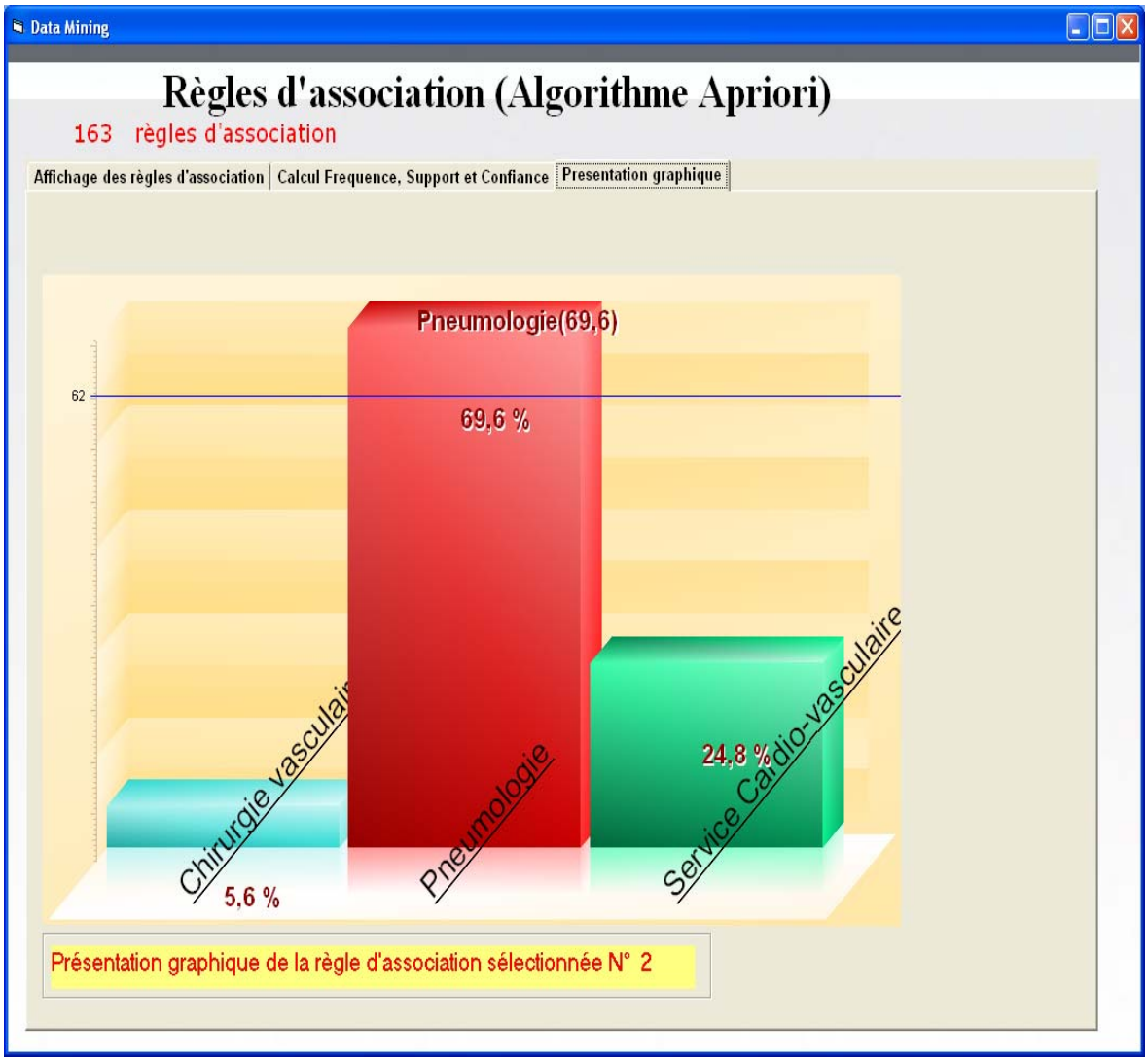

Figure 6.6 : Présentation graphique de la règle d'association sélectionnée.

La figure 6.6 illustre la présentation graphique de la règle d'association sélectionnée N° 2.

6.4 Raisonnement à base de cas

6.4.1 Recherche des cas similaires

Dans ce module nous avons implémenté la partie du système de raisonnement à base de cas, le système permet de saisir un nouveau cas. La figure 6.8 présente la saisie d'identification du patient (sexe, type d'age …),

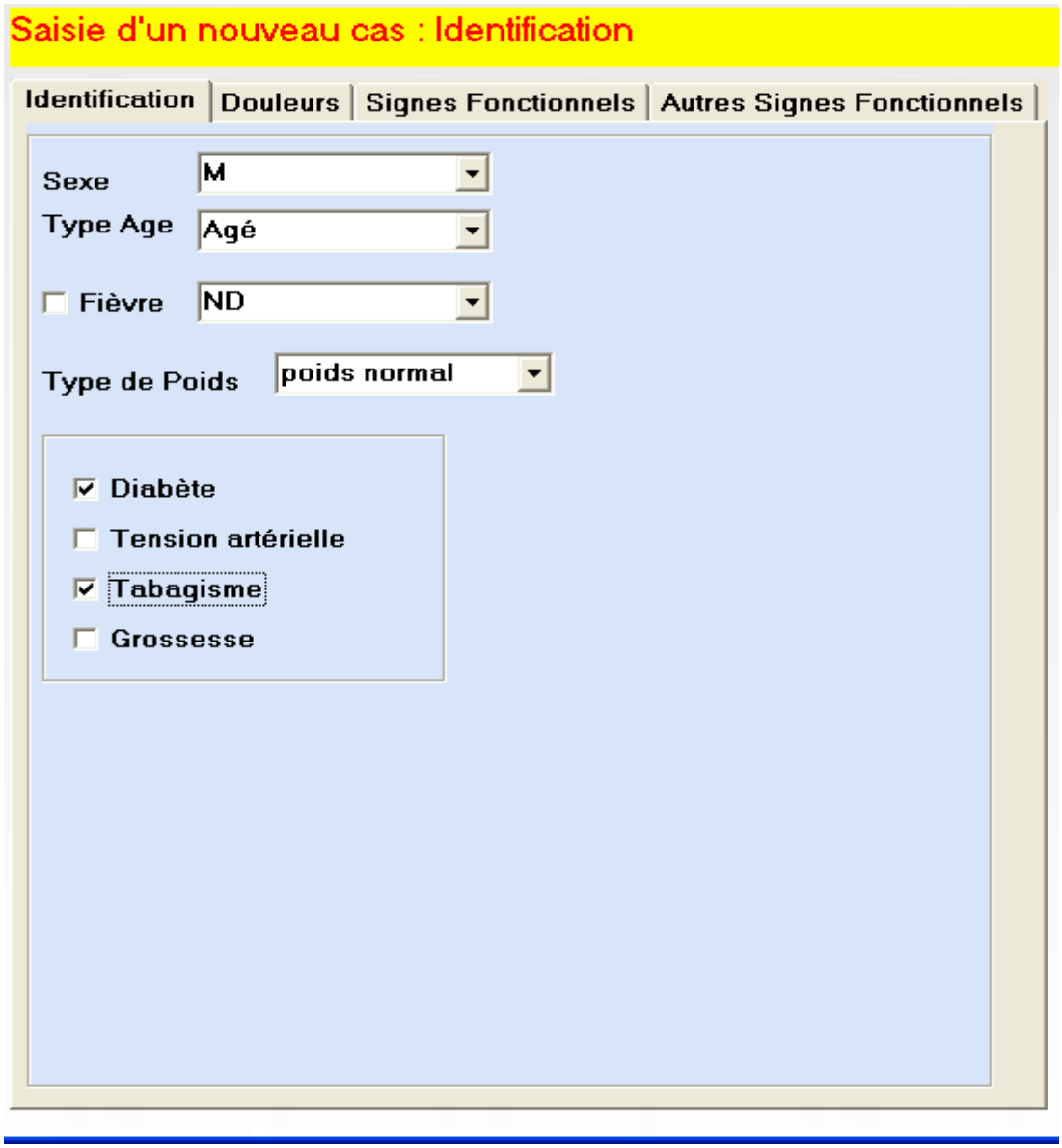

Figure 6.7 : Saisie de l'identification du patient du nouveau cas.

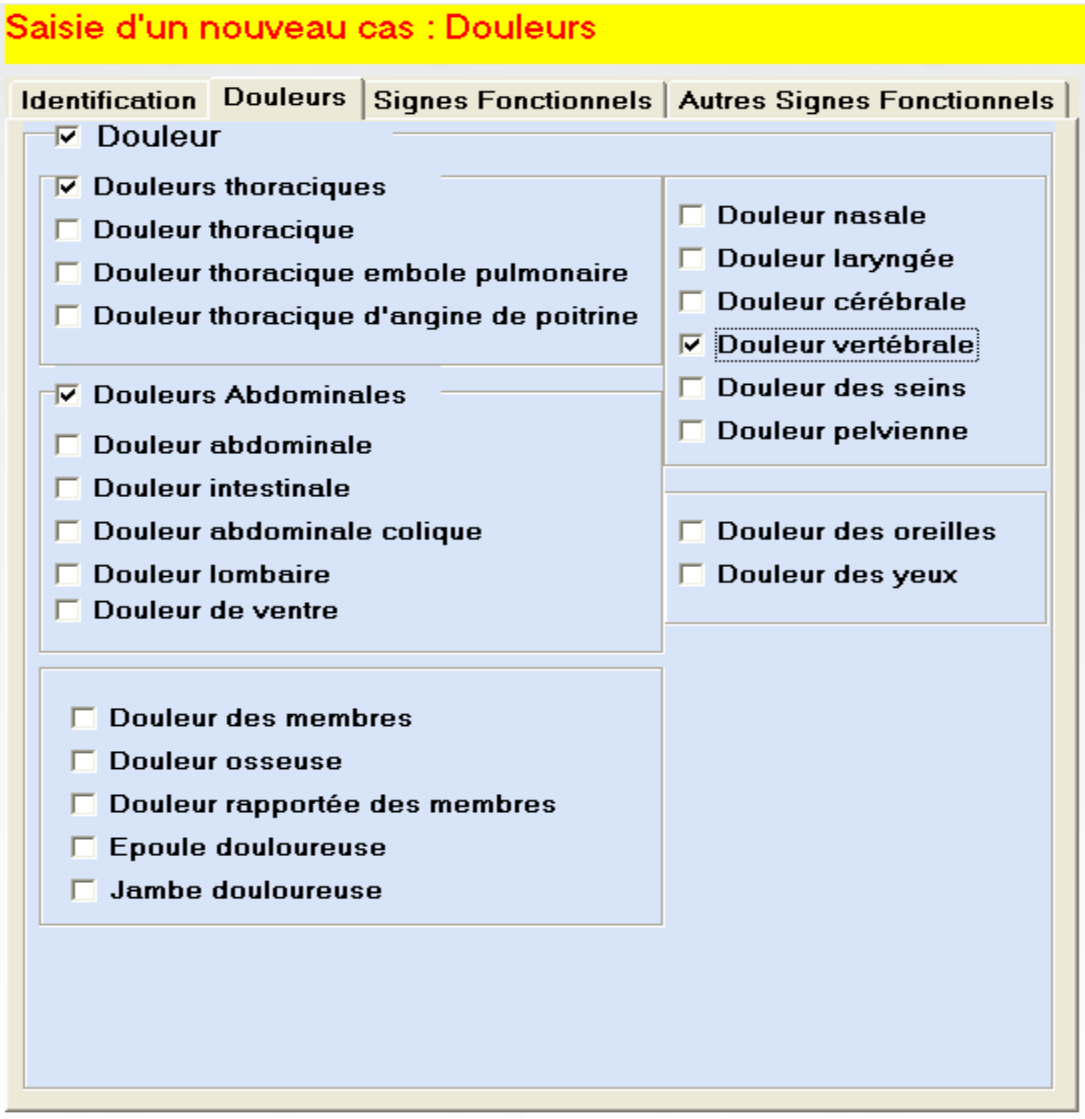

Figure 6.8 : Saisie de la liste des douleurs du nouveau cas.

La figure ci-dessus présente la saisie de la liste des douleurs du nouveau cas.

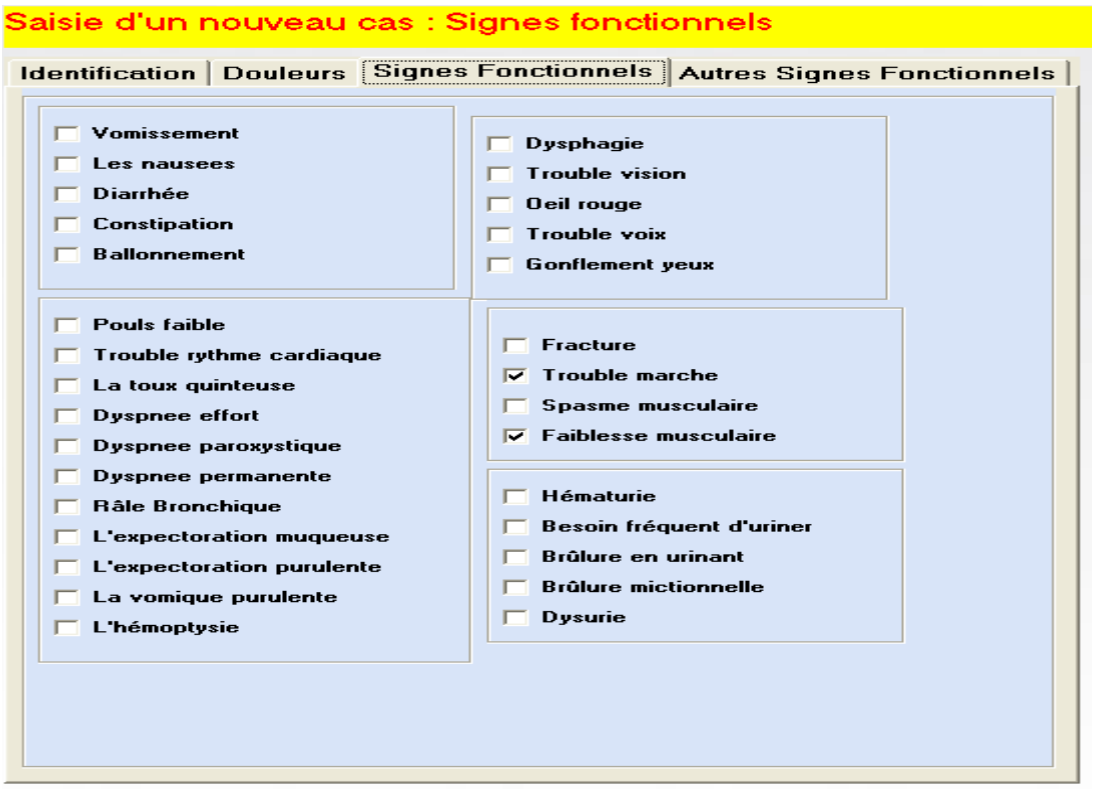

Figure 6.9 : Saisie des signes fonctionnels d'un nouveau cas.

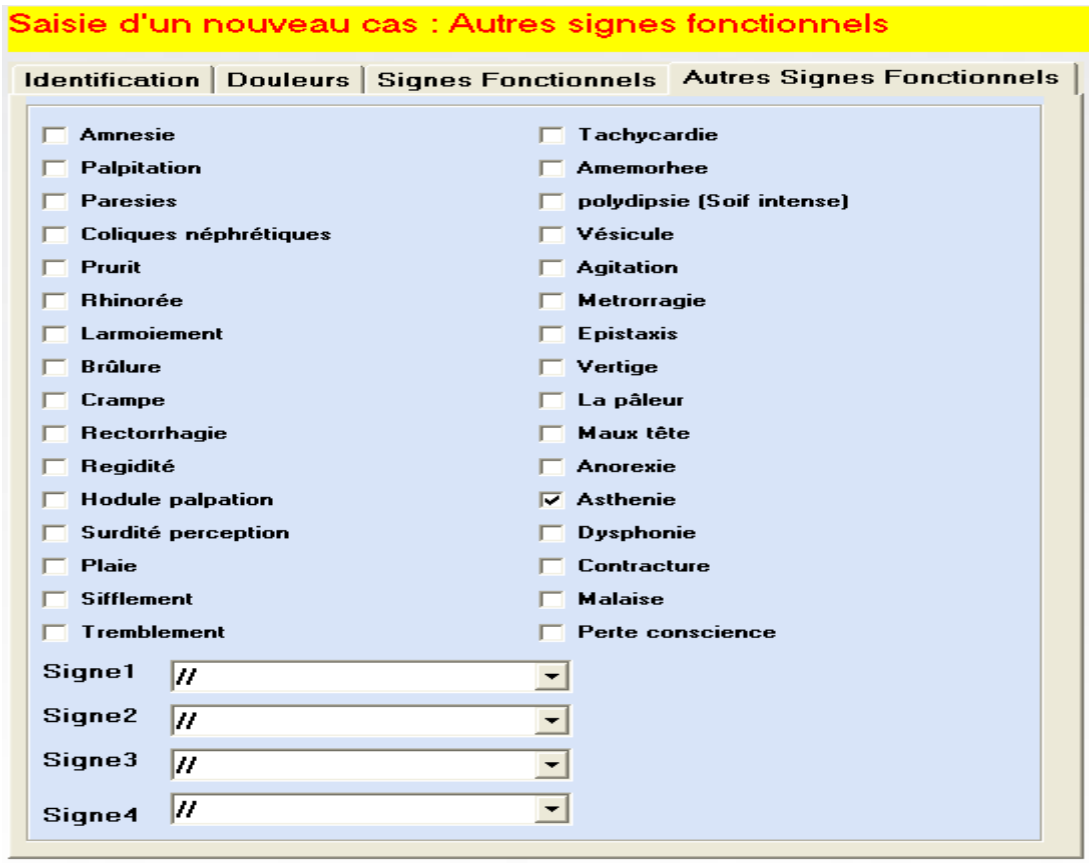

Figure 6.10 : Saisie d'autres signes fonctionnels d'un nouveau cas.

Les figures 6.9 et 6.10 présentent la saisie des signes fonctionnels du nouveau cas.

Une fois le cas saisi, le système cherche les cas similaires au nouveau cas avec le calcul du degré de similarité (le degré de similarité de chaque cas est obtenu par les résultats de l'algorithme ***Apriori***).

Pour proposer le service adéquat, le système compare le degré de similarité de chaque cas similaire par le seuil précisé par l'expert (seuil=50%), s'il trouve un cas similaire avec un taux de confiance (degré de similarité) >= seuil, il présente ce cas comme solution, sinon le système génère un message.

Dans le cas saisi, les cas similaires sont :(Traumatologie/orthopédie, 77.77%), (Neurologie, 11.11) et (Neuro-chirurgie, 11.11). La solution du nouveau cas est :
(Traumatologie/orthopédie, 77.77%)

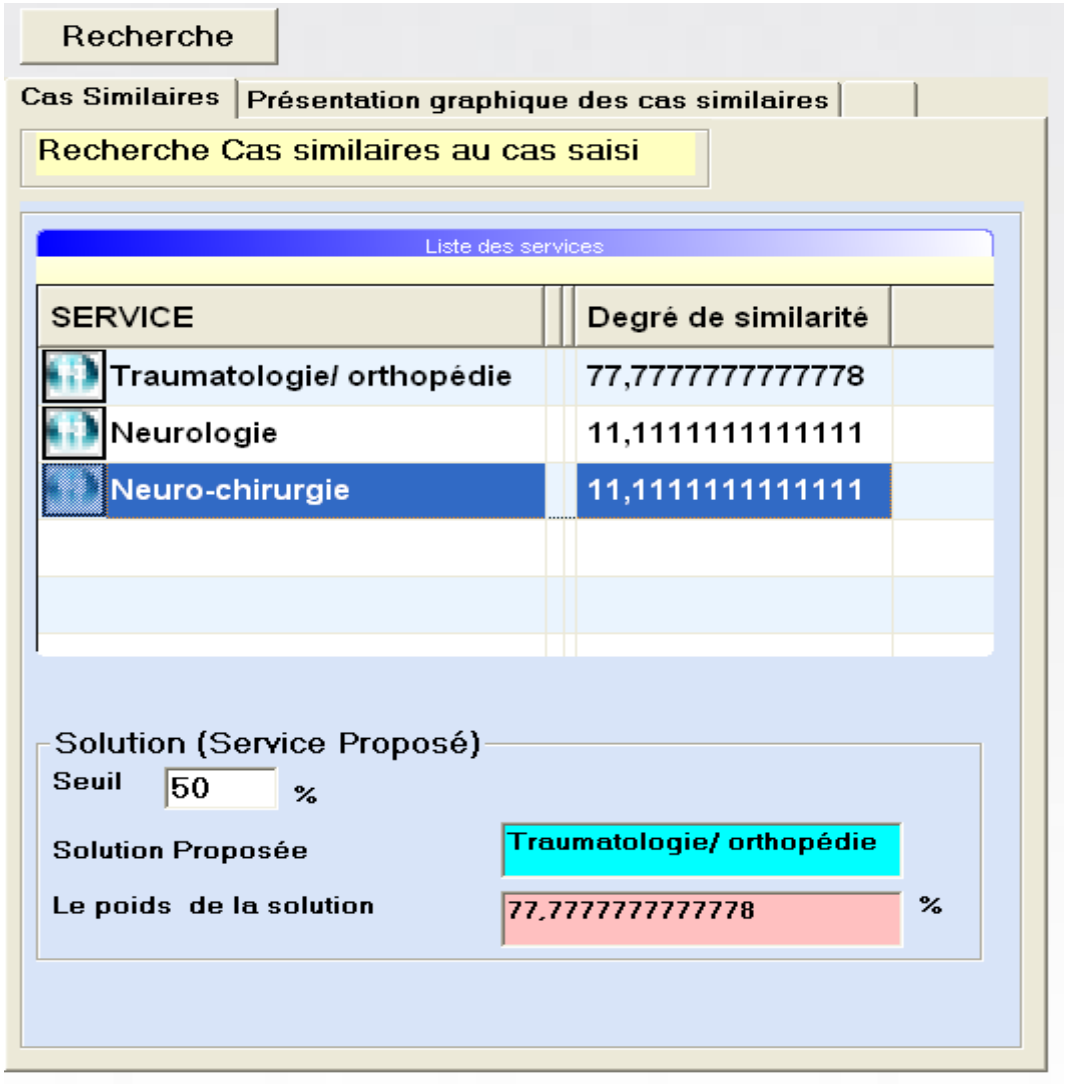

Figure 6.11 : Cas Similaires et la solution du nouveau cas.

6.4.2 Présentation Graphique

La figure suivante illustre la présentation graphique des cas similaires au cas saisi :

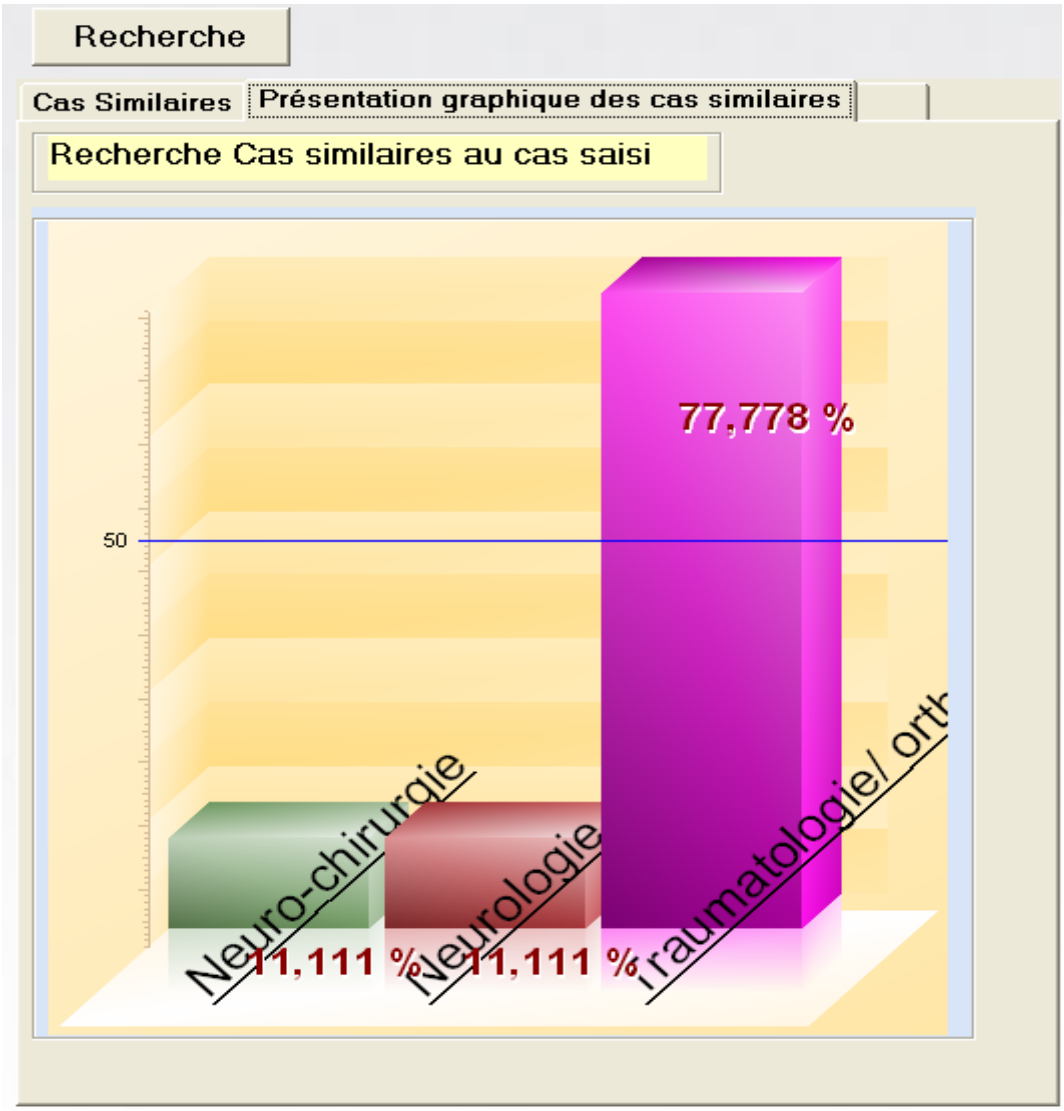

Figure 6.12 : Présentation graphique des cas similaires au cas saisi.

6.4.3 Mémorisation d'un cas

Après le calcul des degrés de similarité entre le nouveau cas et ceux de la base des cas, le système cherche les cas similaires, s'il ne trouve aucun cas similaire ou bien les degrés de similarité des cas similaires sont inférieurs au seuil, le système génère un message « pas de solution, contactez l'expert » ; dans ce cas l'expert peut faire la mémorisation d'un nouveau cas en précisant le service adéquat au cas et un taux de confiance maximal.

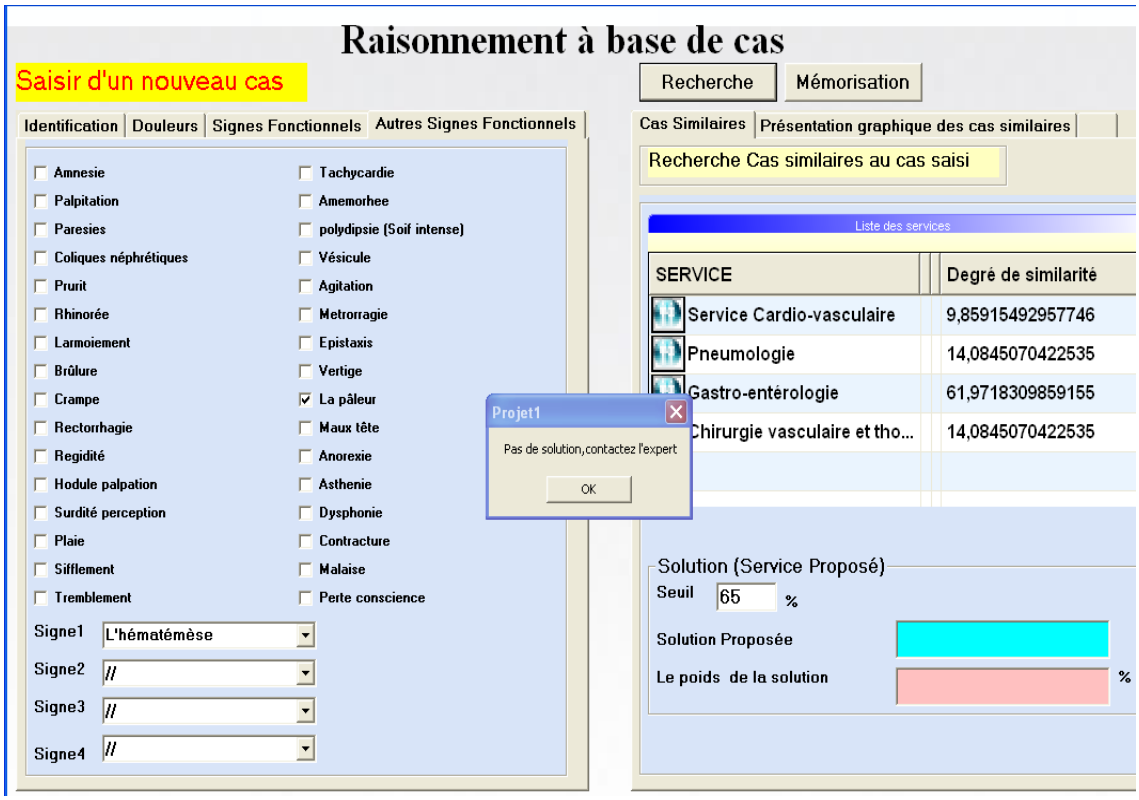

Figure 6.13 : Absence de la solution du cas saisi.

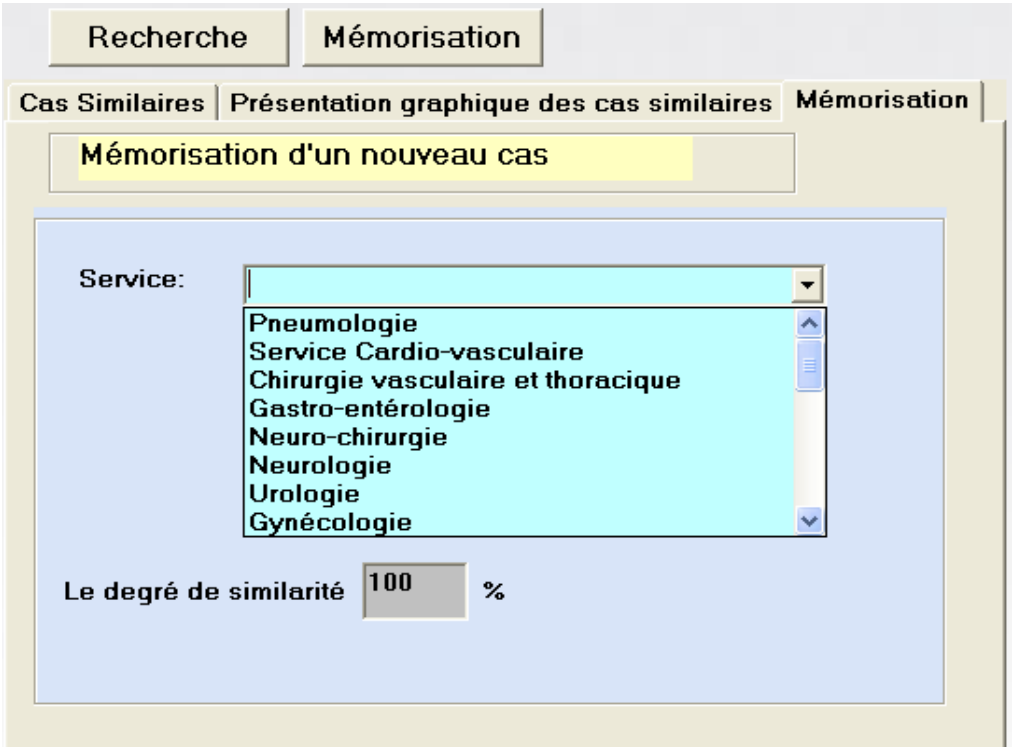

Figure 6.14 : Mémorisation d'un nouveau cas.

6.5 Conclusion

Nous avons présenté dans ce chapitre toutes les fonctionnalités de notre système ***Gestion des connaissances médicales,*** en commençant par le module de l'intégration des données médicales, dans ce module nous avons essayé de présenter les méta-données de l'entrepôt de données et le précalcul des agrégats par la construction des vues matérialisées. Après nous avons développé le module d'extraction des connaissances à partir des données par l'implémentation de l'algorithme Apriori pour générer des règles d'association. Enfin nous avons implémenté un système de raisonnement à base de cas d'orientation médicale qui aide l'utilisateur à orienter le patient vers le service adéquat, la base des connaissances de ce module exploite les connaissances extraites par data mining.

CONCLUSION

Ces dernières années, la gestion des connaissances a connu un grand développement dans divers domaines. Cette inspiration s'illustre effectivement à travers de nombreux articles et ouvrages consacrés à ce concept. Ainsi, nous retrouvons la gestion des connaissances dans différents domaines d'application, et notamment en médecine où les données à traiter sont nombreuses et disparates. Or, il est intéressant d'exploiter ces données pour en extraire des connaissances nouvelles qui par la suite pourraient aider les praticiens dans leurs décisions et gestes.

Dans notre travail nous avons essayé de montrer comment définir une procédure générique de mise en place d'un système d'orientation médicale s'appuyant sur des expériences passées et sans l'introduction de connaissances de l'expert. Cette procédure n'intègre aucune connaissance *a priori* pour la mise en place des bases de connaissances sur lesquelles le système s'appuie pour permettre l'extraction automatique de nouvelles connaissances ou l'apprentissage automatique.

Notre travail de recherche a montré la relation qui existe entre les trois technologies : Intégration des données hétérogènes, Extraction des connaissances à partir des données (ECD) et le raisonnement à base de cas (CBR) dans la gestion des connaissances médicales.

Afin de mettre en relief les principales contributions de nos travaux, nous rappelons les différentes étapes de notre démonstration.

Nous avons montré dans la première partie pourquoi et comment il est possible d'intégrer toutes les technologies de la gestion des connaissances dans un système complet et cohérent dans le domaine médical; elle présente le contexte du problème, elle a pour but de montrer la synergie qui existe entre les entrepôts de données, l'extraction des connaissances à partir de données et le raisonnement à base de cas.

Dans le premier chapitre, et après avoir exposé le contexte de notre travail, nous avons abordé la gestion des connaissances dans le domaine médical, ou nous avons constaté que la diversité des sources des données distribuées et leur hétérogénéité pose des difficultés dans la gestion des connaissances dans le domaine médical.

Dans le deuxième chapitre nous avons proposé une solution pour résoudre les types de conflits afin de combiner et fusionner correctement les informations et connaissances issues de ces sources médicales. Plusieurs approches ont été proposées dans ce cadre. Parmi elles l'entrepôt des données.

Dans le troisième chapitre nous avons mis en oeuvre les techniques d'extraction de connaissances à partir de données stockées dans les entrepôts. Les résultats de cette étape vont constituer la base de connaissances qui permet d'implémenter l'étape de raisonnement à base de cas qui est détaillée dans le quatrième chapitre.

Nous avons, pour finir, illustré dans la deuxième partie, la conception de notre système de gestion des connaissances médicales en s'appuyant sur les correspondances établies aux chapitres précédents. Pour cela, nous avons détaillé notre méthode de réalisation de toutes les étapes du système : depuis l'intégration des données hétérogènes jusqu'à l'implémentation du système de raisonnement à base de cas en passant par l'extraction des connaissances à partir des données intégrées.

1 Contributions

Dans ce modeste travail, nous avons conçu et réalisé un système intelligent de gestion des connaissances médicales, permettant d'orienter les patients vers les différents services hospitaliers. Notre travail est décomposé en trois grandes parties :

- ✓ La première partie consiste à construire un entrepôt de données médicales pour résoudre le problème de l'hétérogénéité des sources de données médicales. Pour cela nous avons suivi le processus ETL (Extraction, transformation, Loading), puis nous avons préparé ces données par la construction d'un schéma global de l'entrepôt et faire une modélisation multidimensionnelle et interrogation de l'entrepôt par les vues matérialisées.
- ✓ Dans la deuxième partie du système, nous avons réalisé un processus d'extraction des connaissances médicales à partir de l'entrepôt de données déjà construit. Le processus que nous avons retenu se décompose en plusieurs phases :

- La sélection des données,
- La préparation des données sélectionnées,
- L'utilisation de la technique des règles d'associations du data mining appliquée sur les données traitées, avec l'implémentation de l'algorithme Apriori.
- Evaluation et présentation des résultats.

✓ Dans la troisième partie, nous avons présenté notre système de raisonnement à base de cas d'orientation médicale qui aide l'utilisateur à orienter le patient vers le service adéquat, nous avons exploité les connaissances extraites du data mining.

2 Perspectives

Dans l'état actuel, ce travail propose un cadre permettant de sensibiliser les chercheurs développant des applications à l'intérêt de gestion des connaissances médicales. Il offre plusieurs perspectives de recherche :

- Par rapport aux connaissances extraites et leurs utilisations ;
- Par rapport à l'organisation de la base de connaissances (de cas) qui doit être plus élaborée;
- Par rapport au rafraîchissement de l'entrepôt de données à partir de la base des connaissances.

REFERENCES

[1] Baizet.Y « La Gestion des Connaissances en Conception Application à la simulation numérique chez Renault » - DIEC, Thèse de doctorat de l'université Joseph Fourier, Grenoble1, Soutenue le 12 mars 2004.

[2] Karl Wiig, « People-Focused Knowledge Management », Elsevier, 2004.

[3] Slodzian M. Wordnet : «What about its linguistic relevancy ? »In R. DIENG, Coordinateur, Proc. of the EKAW conference, Juan-les-Pins, France.2007.

[4] Tanguy, C. et Villavicencio, D., « Apprentissage et innovation dans l'entreprise, Une approche socio-économique des connaissances », Eres, Aix en Provence, 254 pages.2005.

[5] Nonaka I., Takeucchi H. "The Knowledge-Creating Company", Oxford: Oxford University Press, 1995. 2000.

[6] Polanyi, M, "Personal Knowledge: Toward a Post-Critical Philolosophy" , University of Chicago Press, Chicago, IL.1962.

[7] Leonard, D. et Sensiper, S. Wellsprings of knowledge: "Building and sustaining the source of innovation", California Management Review, Vol. 40, n°3, pp. 112-132. 1998.

[8] Hall, R., et Andriani, P., "Managing knowledge associated with innovation", Journal of Business Research, Vol. 56, pp 145-152. 2003.

[9] Von Krogh G., Ichijo K., Nonaka I. "Enabling Knowledge creation : How to unlock the mystery of tacit knowledge and release the power of innovation?" Oxford University Press, Inc., 2002.

[10] Probst G., Raub S., Romhardt K. "Managing Knowledge: Building Blocks for Success". Chichester: John Wiley & Sons Ltd, pp. 5–7, 28-37. 2000.

[11] Degoulet.P. & Fieschi M. . « Traitement de l'information médicale : Méthodes et applications hospitalières », chapitre Informatisation des dossiers médicaux. Manuels Informatiques. Masson Entreprise. 1991.

[12] CLEMMER T. « The role of médical informatics in telemedicine ». Journal of Médical System. 2000.

[13] Renard J., Beuscard R., Delerue d. & Geib j. Le réseau ville-hôpital : « Une nouvelle forme de communication entre professionnels de santé ». In P. DEGOULET & M. FIESCHI, Coordinateurs, Revue européenne de biotechnologie médicale, volume 21 of Innovation et technologie en biologie et médecine, p. 275–280. Springer-Verlag. 2000.

[14] Voros S., Orvain e., Long j. & Cinquin p. "Automatic detection of instruments in laparoscopic images : a first step towards high level command of robotized endoscopic holders" . In International Conference on Biomédical Robotics and Biomechatronics, Pisa, Italy, 2006.

[15] Soualmia L. & Darmoni S. "Combining different standards and different approaches for health information retrieval in a quality-controlled gateway". International Journal of Médical Informatics (IJMI), p. 141–150. 2005.

[16] Shiffman R., Michel G. & Essaihi a. "Bridging the guideline implementation gap : a systematic approach to document-centered guideline implementation". J Am Med Inform Assoc, 11, 418–426. 2004.

[17] Audrey Baney, Construire « Une ontologie de la pneumologie Aspects théoriques, modèles et expérimentations », Thèse de Doctorat de L'université Paris 6, 2007.

[18] Stefanelli M. "Knowledge management to support performance-based medicine. Methods of Information in Medicine", 1, 36.43. 2002.

[19] Ermine J.-L. « Les systèmes de connaissance ». Hermès. 1996.

[20] Florence GONOD BOISSIN, »L'usage de l'information numérique en médecine générale », Diplome De Doctorat mars 2007, L'université Claude Bernard

[21] Jean Charlet « Mission de recherche en sciences et technologies de l'information médicale ». Assistance Publique.Hôpitaux de Paris, 2002.

[22] Perrier a. & Similowski T. « Une série sur la médecine factuelle « evidence-based medicine »: mais pour quoi faire ? Revue des maladies respiratoires », 19, 395.398. 2002.

[23] Lenay O. Régulation, « Planification et organisation du système hospitalier : la place des outils de gestion dans la conception des politiques publiques ». Thèse de doctorat, École de smines de Paris. 2001.

[24] Lenay o. & Moisdon J.-C. « Croître à l'abri de la gestion ? le cas du système hospitalier public français ». Cahiers de recherche du centre de gestion scientifique, École des Mines de Paris, (17). ISSN 1268-4317. 2006.

[25] Naiditch M. Modélisation des trajectoires : « Problèmes méthodologiques. Innovation et technologie en biologie et médecine », 21(5), 307.12. 2005.

[26] Grosjean M.. Les communications collectives : un mode d'approche des compétences du collectif. l'exemple du collectif hospitalier. Psychologie du travail et des organisations, 6(3-4), 103.31. Numéro spécial compétences collectives au travail. 2000.

[27] Beuscart R., Renard J.-M., Delarue D. & Souf A. « Le reseau ville-hôpital : une nouvelle forme de communication entre professionnels de santé ». Informatique et Santé, 11.2009.

[28] Weis J.-C. « Du réseau ville-hôpital au réseau distributionne l ». Rapport interne, STIM/DPA/DSI/AP-HP. DEA informatique médicale. 2002.

[29] Inmon W.H., Conklin E., "Loding data into the warehouse". Tech Topic, 1 (11), 1994.

[30] Aberer K. and Hemm K., “A Methodology for building Data Warehouse in a Scientific Environment”, 2005.

[31] J-M Franco. “Le Data Warehouse”. Editions Eyrolles, Paris, 1997.

[32] A.Doucet and S.Gangarski.. « Entrepôts de données et Bases de Données Multidimensionnelles », Chapitre 12 du livre : Bases de Données et Internet, Modèles, langages et systèmes. Editions Hermès, 2001.

[33] Thomas Zurek and Markus Sinnwell. “Datawarehousing Has More Colours Than Just Black & White”. In VLDB ’99 : Proceedings of the 25th International Conference on Very Large Data Bases, pages 726–729. Morgan Kaufmann Publishers Inc., 1999.

[34] W. Inmon and R.D. Hackathorn. « Using the Data Warehouse ». 1994.

[35] E. Benitez. « Infrastructure adaptable pour l’évolution des entrepôts de données ». Phd thèse, Université Joseph Fourier, Grenoble, France, Septembre 2002.

[36] Jennifer Widom. “Research problems in data warehousing”. In CIKM ’95 : Proceedings of the fourth international conference on Information and knowledge management, pages 25–30, New York, NY, USA, 1995. ACM Press.

[37] Christine Collet, Genovera Vargas-Solar et Helena Graziotin-Riboiro. « Open Active Services for Data Intensive Distributed Applications », dans 16 em journées Bases de Données Avancées (BDA 2000), pages 43-60 Blois Octobre 2000.

[38] Ralph Kimball. The Data Warehouse Toolkit. John Wiley, USA, 1996.

[39] Helena Galhardas, Daniela Florescu, Dennis hasha et Eric Simon. « An Extensible Data Cleaning Tool”. Dans Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, page 590, Dallas, USA, May 2000.

[40] Clement T. Yu et Weiyi Meng, « Principales of Database Query Processing. Morgan Kaufmann Publishers », Inc., San Francisco, USA, 1998.

[41] María Trinidad Serna Encinas. « Entrepôts de données pour l'aide à la décision médicale : conception et expérimentation ». Thèse phd. Université Joseph Fourier, Grenoble, France, Juin 2005.

[42] Ralph Kimboll et Kevin Strehlo. "SQL is Our Language", Fix it now. SIGMOD Record, 24(3), 1995.

[43] H V. Jagadish, Laks V S Lakshmanan et Divesh Srivastava. "What can Hierarchies do for Data Warehousing". Dans Proceeding of the 25th International Conference on Very Large Data Bases (VLDB 99). Edinburg, Scotland, September 1999.

[44] Jérôme Darmont, « Optimisation et évaluation de performance pour l'aide à la conception et à l'administration des entrepôts de données complexes ». Université Lumière Lyon 2 Ecole Doctorale de Sciences Cognitives. Novembre 2006.

[45] V. Harinarayan, A . Rajaraman. "Data Cube Efficiently". Dans proceedings of the 1996 ACM SIGMOD International Conference on management of Data (SIGMOD 96), 205 , Montereal Canada , 1996

[46] C. Zhang, X. Yao, and J. Yang. "An evolutionary approach to materialized view selection in a data warehouse environment". IEEE Trans. Systems, Man, Cybernetics, PART C, 31 :282–294, September 2001.

[47] Michel Adiba. Derived Relations : "A Unified Mechanism for Views, Snapshots and Distributed Data". Dans Proceedings of the 7th International Conference on Very Large Data Bases (VLDB81) page 293, Cannes, France, September 1981.

[48] Luca Cabibbo and Riccardo Torlone. "Querying Multidimensional Databases. In Workshop on Database Programming Languages", pages 319–335, 1997.

[49] R. Kimball. Entrepôts de données, « Guide pratique du concepteur de data warehouse ». John Wiley and Sons, Inc., 1996.

[50] Rakesh Agrawal, A. Gupta, and Sunita Sarawagi. ”Modeling Multidimensional Databases”. In Alex Gray and Per-Åke Larson, editors, Proc. 13th Int. Conf. Data Engineering, ICDE, pages 232–243. IEEE Computer Society, 7–11 1997.

[51] William Inmon. Building the Data Warehouse. QED Technical Publishing Group, Wellesley, Massachusetts, U.S.A., 1992, 1992.

[52] William Inmon. “What is a Data Warehouse”, White paper, 1995 .

[53] E. Benitez, C. Collet, and M. Adiba. “Entrepôts de données : caractéristiques et problématique ». Revue TSI, 20(2), 2001.

[54] R. Kimball and M. Ross. “Entrepôts de données, Guide pratique de modélisation dimensionnelle ». Vuibert, Paris, 2003.

[55] Ralph Kimball and Kevin Strehlo. “Why Decision Support Fails and How To Fix It”. SIGMOD Record, 24(3) :92–97, 1995.

[56] M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. “From data mining to knowledge discovery: An overview. Advances in Knowledge Discovery and Data Mining”, pages 1–34, 1996.

[57] Evangelos Simoudis. “Reality check for data mining”. IEEE Expert: Intelligent Systems and Their Applications, 11(5) :26–33, 1996.

[58] K. J. Hammond. “On functionally motivated vocabularies: An apologia. In Proc. Of a Workshop on Case-Based Reasoning”, pages 52–56, Pensacola Beach, FL, 1989.

[59] J. Han, Y. Fu, W. Wang, K. Koperski, and O. Zaiane. Dmql : “A data mining query language for relational databases”, 1996.

[60] Jiawei Han and Micheline Kamber. “Data Mining : Concepts and Techniques. Morgan Kaufmann”, 2000.

[61] A.A. Freitas. Generic, “Set-Oriented Primitives to Support Data-Parallel Knowledge Discovery in Relational Database Systems”. Phd thèse, University of Essex, UK, 1997.

[62] Y Kodratoff., "techniques et outils de l’extraction de connaissances à partir des données", Signaux n°92 pp 38-43, Mars 1998.

[63] Eva Andràssyovà and Jàn Paralic. “Knowledge discovery in databases : A comparison of different views”. 2002

[64] Heikki Mannila. “Methods and problems in data mining”. In ICDT, pages 41–55, 1997.

[65] J. Brachman and T. Anand. “The process of knowledge discovery in databases. Advances in knowledge discovery and data mining”, pages 37–57, 1996.

[66] J. Lieber inspire du cours d'Amedeo Napoli, ,"Fouille de données : notes de cours ",2007.

[67] Agrawal R, Imielinski T, Swami A, "Mining Association rules between sets of items in large database", Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, pp 207-216, May 26-28, 1993.

[68] G. Piatetsky-Shapiro. “Discovery, analysis, and presentation of strong rules”. In G. Piatetsky-Shapiro and W. J. Frawley, editiors, Knowledge Discovery in databases, pages 229-238. AAAI/ MIT press,1991.

[69] S Tufféry,"data mining et scoring, Bases de données et gestion de la relation client," Groupe bancaire francais, universités de Rennes 1 et paris-Dauphine, 2002.

[70] M Kantardzic. "Data Mining – Concepts, Models, Methods,and Algorithms". IEEE Press, Piscataway, NJ, USA, 2003.

[71] Med, H, Haddad, "Extraction et impact des connaissances sur les performances des systèmes de recherche d’information", these Université Joseph Fourier-

Grenoble1, 2002.

[72] Fayyad U, Piatetsky-Shapiro G, Smyth P, "From Data Mining to Knowledge Discovery in Databases", Advices in Knowledge Discovery and Data Mining, MIT Press, 1:pp 1-36, 1998.

[73] S Tufféry, "data mining et statistique décisionnelle, l'intelligence dans les bases de données", Groupe bancaire francais, universités de Rennes 1 et paris-Dauphine 2005.

[74] Douglas H. Fisher, J. Cios." Knowledge acquisition via incremental conceptual clustering". Mach. Learn., 2(2) :139–172, September 1987.

[75] Fisher D. Gennari J., Langley P. "Model of incremental concept formation". Artificial Intelligence Journal, 40: 11–61, 1989.

[76] Gierl and Schmidt. "Cbr in medicine, case-based reasoning technology". From Foundations to Applications, pages 273–297, 1998.

[77] Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. Freespan : "Frequent pattern-projected sequential pattern mining". In KDD, pages 355–359, 2000.

[78] Jiawei Han, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". In Weidong Chen, Jeffrey Naughton, and Philip A. Bernstein, editors, 2000 ACM SIGMOD Intl. Conference on Management of Data, pages 1–12. ACM Press, 05, 2000.

[79] P. Gancarski and C. Wemmert. "Collaborative multi-strategy classification: application to per-pixel analysis of images". In MDM '05: Proceedings of the 6th international workshop on Multimedia data mining, pages 15–22, New York, NY, USA, 2005. ACM Press.

[80] F. Azuaje, C. Gertosio and A. Dussauchoy. "Knowledge discovery from industrial databases". Journal of Intelligent Manufacturing, 15: 29–37, 2004.

[81] David Heckerman. “Bayesian networks for data mining”. Data Min. Knowl. Discov., 1(1) :79–119, 1997.

[82] Markus Nilsson and Mikael Sollenborn. Advancements and trends in médical casebased reasoning: An overview of systems and system development. In FLAIRS Conference, 2004.

[83] Aamodt, A and Plaza, E ‘Case-Based Reasoning: Foundational Issues, Methodological Variations and Systems Approaches’ AI Communications, Vol 7 No i pp 39-59.1994.

[84] P. Koton. “Reasoning about evidence in causal explanations”. In Proc. of a Workshop on Case-Based Reasoning, pages 260–270, Holiday Inn, Clearwater Beach, FL, 1988.

[85] D'aquin M., Brachais S., Lieber J. & Napoli A. « Vers une acquisition automatique de connaissances d'adaptation par examen de la base de cas - une approche fondée sur des techniques d’extraction de connaissances dans des bases de données ». In 12ème Atelier de Raisonnement à Partir de Cas - RàPC'04. p.41-52. Université Paris Nord, Villetaneuse. 2004.

[86] Elorriaga J. & Fern Ndez-Castro I. “Using case-based reasoning in instructional planning: towards a hybrid self-improving instructional planner”. In International Journal of Artificial Intelligence in Education, p. 416–449. 2000.

[87] Eric Buist, « Les éléments fondamentaux du raisonnement à base de cas », 2004.

[88] K. J. Hammond. “On functionally motivated vocabularies : An apologia”. In Proc. Of a Workshop on Case-Based Reasoning, pages 52–56, Pensacola Beach, FL, 1989.

[89] Fuchs, Béatrice et al. « Towards a unified theory of adaptation in case-based reasoning ». In Althoff et al. (1999), pages 104–117.

[90] Schmid Ute, Carbonell Jaime.” Empirical evidence for derivational analogy”. 21st Annual Conference of the Cognitive Science Society, Simon Fraser

University-Vancouver, British Columbia, august 19-21, 1999.

[91] Blumenthal Brad, Porter Bruce. “Analysis and empirical studies of derivational analogy”. Journal of Artificial Intelligence, volume 67, 1994.

[92] Kolodner, Janet L. et Leake, David B. « A tutorial introduction to case-based reasoning ». In Leake, David B., editor, Case-Based Reasoning Experiences, Lessons, & Future Directions, pages 31–66. American Association for Artificial Intelligence, Menlo Park, 1996.

[93] Goodman, M. « CBR in battle planning ». In Hammond, K., editor, Workshop on case-based reasoning (DARPA), San Mateo, 1989b. Morgan Kaufmann.

[94] Schumacher, J¨urgen et Bergmann, Ralph. « An efficient approach to similarity-based retrieval on top of relational databases ». In Blanzieri. pages 273–284. 2000.

[95] Eckstein, Robert et Casabianca, Michel. « XML Pocket Reference”. O’Reilly & Associates, Inc., Sebastopol, second edition, 2001.

[96] Hayes, Conor et Cunningham, P. « Shaping a CBR view with XML ». In Althoff et al. (1999), pages 468–481.

[97] Ruet, Magali et Geneste, Laurent. « Search and adaptation in a fuzzy object oriented case base ». In Craw et Preece, pages 350–364. 2002.

[98] Keung Shiu, Simon Chi et al. «Maintaining case-based reasoning systems using fuzzy decision trees ». In Blanzieri, pages 285–296.2000.

[99] Gierl and Schmidt. “Cbr in medicine, case-based reasoning technology. from Foundations to Applications”, pages 273–297, 1998.

[100] Isabelle Bichindaritz, Carol Moinpour, Emin Kansu, Gary Donaldson, Nigel Bush, and Keith M. Sullivan. “Case based reasoning for médical decision-support in a safety critical environment” . In AIME, pages 314–323, 2003.

[101] Ziad El Balaa and Ralph Traph¨oner. “Case-based decision support and experience management for ultrasonogrphy”. In Wissensmanagement, pages 277–278, 2003.

[102] Z. E. Balaa, A. Strauss, P. Uziel, K. Maximini, and R. Traphoner. “Fm-ultranet :a decision support system using case-based reasoning, applied to ultrasonography”” . In Lorraine McGinty, editor,Workshop on Case-Based Reasoning in the Health Sciences, Trondheim, Norway, June. NTNU Department of Computer and Information Science. 2003.

[103] P. Perner. “An architeture for a Cbr image segmentation system”. K.-D. Althoff, R. Bergmann, and K. Branting (Eds.) Case-Based Reasoning Research and Development, lnai 1650, Springer Verlag, pages 525–535, 1999

[104] I.Bichindaritz, E. Kansu, and K. M. Sullivan. “Case-based reasoning in care-partner: Gathering evidence for evidence-based médical practice”. j-LECT-NOTES-COMPSCI, 1488 :334–345, 1998.

[105] Rainer Schmidt, Bernhard Pollwein, and Lothar Gierl. “Experiences with case-based reasoning methods and prototypes for médical knowledge-based systems”. In AIMDM, pages 124–132, 1999.

[106] M.-C. Jaulent, C. L. Bozec, E. Zapletal, and P. Degoulet. “A case-based reasoning method for computer-assisted diagnosis in hisopathology”. In Artificial intelligence in Medicine, pages 239–242. AIME’97, 1997.

[107] S. Montani, L. Portinale, G. Leonardi, and Riccardo Bellazi. Applying case-based retrieval to hemodialysis treatment. In In Workshop on CBR in the Health Sciences, pages 53–62. International Conference on Case-Based Reasoning (ICCBR’03), June 2003.

[108] E. Costello and D. C. Wilson. “A case-based approach to gene finding”. In Proceedings of the Fifth International Conference on Case-Based Reasoning Workshop on CBR in the Health Sciences, pages 19–28. ICCBR’03, 2003.

[109] M. Nilsson, P. Funk, and M. Sollenborn. “Complex measurement classification in médical applications using a case-based approach” . In Workshop on CBR in the Health Sciences, pages 63–72. ICCBR’03, 2003.

[110] R. Schmidt and L. Gierl. “Prognostic model for early warning of threatening influenza waves”. In German Workshop on Experience Management, pages 34–46. GWEM’02, 2002.

[111] Stefania Montani, Paolo Magni, Abdul V. Roudsari, Ewart R. Carson, and Riccardo Bellazzi. “Integrating different methodologies for insulin therapy support in type 1 diabetic patients” . In AIME’01: Proceedings of the 8th Conference on AI in Medicine in Europe, pages 121–130, London, UK, Springer-Verlag. 2001.

[112] K. Evans-Romaine and C. Marling. “Prescribing exercise regimens for cardiac and pulmonary disease patients with Cbr. In Workshop on CBR in the Health Sciences, pages 45–52. ICCBR’03, 2003.

[113] C. Marling and P. Whitehouse. “Case-based reasoning in the care of alzheimer’s disease patients”. In Case-Based Research and Development, pages 702–715. ICCBR’01, 2001.

[114] C. Marling, Whitehouse. “Case-based reasoning in the care of alzheimer’s disease patients”n. In Case-Based Research and Development, pages 702–715. ICCBR’ 01, 2003.

[URL1] <http://www. ebm.lib.ulg.ac.be/prostate/ebm.htm>.

[URL2] <http://www.cnr.asso.fr/>.

[URL3] CASUEL, http://www.agr.informatik.uni-l.de/bergmann/casuel/.

[URL4] XML.org, http://www.xml.org.