Fouille de connaissances en diagnostic mammographique par ontologie et règles d'association

Rihab Idoudi

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

UNIVERSITE BRETAGNE LOIRE

THÈSE / IMT Atlantique

sous le sceau de l'Université Bretagne Loire

pour obtenir le grade de

DOCTEUR D'IMT Atlantique

École Doctorale Sicma

Et en cotutelle avec l'Ecole Nationale d'Ingénieurs de Tunis (Tunisie)

Mention : Sciences et Technologies de l'Information et de la Communication

Présentée par

Rihab Idoudi

Préparée au département Image et Traitement Information

Fouille de connaissances en diagnostic mammographique par ontologie et règles d'association

Ontologies and Association Rules Knowledge Mining, Case Study: Mammographic domain

Thèse soutenue le 24 janvier 2017

devant le jury composé de :

Ali Khenchaf
Professeur, ENSTA Bretagne / président

Amel Grissa
Professeure, ISIG Kairouan - Tunisie / rapporteur

Jean Paul Haton
Professeur, Université Henri Poincaré - Nancy / rapporteur

Karim Saheb Ettabaa
Maître-assistant, ISSAT - Tunisie / examinateur

Kamel Hamrouni
Professeur, ENIT de Tunis - Tunisie / co-directeur de thèse

Basel Solaiman
Professeur, IMT Atlantique / directeur de thèse

Mohamed Mohsen Gammoudi
Professeur, ISSAMM Université de Marouba - Tunisie / invité

Résumé

Face à la complexité significative du domaine mammographique ainsi que l'évolution massive de ses données, le besoin de contextualiser les connaissances au sein d'une modélisation formelle et exhaustive devient de plus en plus impératif pour les experts. C'est dans ce cadre que se situe notre travail de thèse qui s'intéresse à unifier différentes sources de connaissances liées au domaine au sein d'une modélisation ontologique cible.

D'une part, plusieurs modélisations ontologiques mammographiques ont été proposées dans la littérature, où chaque ressource présente une perspective distincte du domaine d'intérêt. D'autre part, l'implémentation des systèmes d'acquisition des mammographies rend disponible un grand volume d'informations issues des faits passés, dont la réutilisation devient un enjeu majeur. Toutefois, ces fragments de connaissances, présentant de différentes évidences utiles à la compréhension de domaine, ne sont pas interopérables et nécessitent des méthodologies de gestion de connaissances afin de les unifier. C'est dans ce cadre que se situe notre travail de thèse qui s'intéresse à l'enrichissement d'une ontologie de domaine existante à travers l'extraction et la gestion de nouvelles connaissances (concepts et relations) provenant de deux courants scientifiques à savoir: des ressources ontologiques et des bases de données comportant des expériences passées.

Notre approche présente un processus de couplage entre l'*enrichissement conceptuel* et l'*enrichissement relationnel* d'une ontologie mammographique existante. Le premier volet comporte trois étapes. La première étape dite de pré-alignement d'ontologies consiste à construire pour chaque ontologie en entrée une hiérarchie des clusters conceptuels flous. Le but étant de réduire l'étape d'alignement de deux ontologies entières en un alignement de deux groupements de concepts de tailles réduits. La deuxième étape consiste à aligner les deux structures des clusters relatives aux ontologies cible et source. Les alignements validés permettent d'enrichir l'ontologie de référence par de nouveaux concepts permettant d'augmenter le niveau de granularité de la base de connaissances. Le deuxième processus s'intéresse à l'enrichissement relationnel de l'ontologie mammographique cible par des relations déduites de la base de données de domaine. Cette dernière comporte des données textuelles des mammographies recueillies dans les services de radiologies. Ce volet comporte ces étapes : i) Le prétraitement des données textuelles ii) l'application de techniques relatives à la fouille de données (ou extraction de connaissances) afin d'extraire des expériences de nouvelles associations sous la forme de règles, iii) Le post-traitement des règles générées. Cette dernière consiste à filtrer et classer les règles afin de faciliter leur interprétation et validation par l'expert vi) L'enrichissement de l'ontologie par de nouvelles associations entre les concepts. Cette approche a été mise en œuvre et validée sur des ontologies mammographiques réelles et des données des patients fournies par les hôpitaux Taher Sfar et Ben Arous.

Les travaux de recherche présentés dans ce manuscrit s'inscrivent dans le cadre de la valorisation et fusion des connaissances issues des sources hétérogènes afin d'améliorer les processus de gestion de connaissances.

Mots clés : Fouille de connaissances, enrichissement conceptuel d'ontologie, alignement d'ontologies, clustering hiérarchique conceptuel flou, enrichissement relationnel d'ontologie, Extraction de règles d'association, post-traitement des règles d'association.

Abstract

Facing the significant complexity of the mammography area and the massive changes in its data, the need to contextualize knowledge in a formal and comprehensive modeling is becoming increasingly urgent for experts. It is within this framework that our thesis work focuses on unifying different sources of knowledge related to the domain within a target ontological modeling.

On the one hand, there is, nowadays, several mammographic ontological modeling, where each resource has a distinct perspective area of interest. On the other hand, the implementation of mammography acquisition systems makes available a large volume of information providing a decisive competitive knowledge. However, these fragments of knowledge are not interoperable and they require knowledge management methodologies for being comprehensive. In this context, we are interested on the enrichment of an existing domain ontology through the extraction and the management of new knowledge (concepts and relations) derived from two scientific currents: ontological resources and databases holding with past experiences.

Our approach integrates two knowledge mining levels: The first module is the conceptual target mammographic ontology enrichment with new concepts extracting from source ontologies. This step includes three main stages: First, the stage of pre-alignment. The latter consists on building for each input ontology a hierarchy of fuzzy conceptual clusters. The goal is to reduce the alignment task from two full ontologies to two reduced conceptual clusters. The second stage consists on aligning the two hierarchical structures of both source and target ontologies. Thirdly, the validated alignments are used to enrich the reference ontology with new concepts in order to increase the granularity of the knowledge base. The second level of management is interested in the target mammographic ontology relational enrichment by novel relations deducted from domain database. The latter includes medical records of mammograms collected from radiology services. This section includes four main steps: i) the preprocessing of textual data ii) the application of techniques for data mining (or knowledge extraction) to extract new associations from past experience in the form of rules, iii) the post-processing of the generated rules. The latter is to filter and classify the rules in order to facilitate their interpretation and validation by expert, vi) The enrichment of the ontology by new associations between concepts.This approach has been implemented and validated on real mammographic ontologies and patient data provided by Taher Sfar and Ben Arous hospitals.

The research work presented in this manuscript relates to knowledge using and merging from heterogeneous sources in order to improve the knowledge management process.

Keywords: Knowledge mining, ontology conceptual enrichment, ontology hierarchical fuzzy conceptual clustering, relational ontology enrichment, association rules extraction, post-processing of association rules.

A mes chèrs parents Abdelhamid & Hedia, …. A qui je dois tout;

A mon cher mari Oussema, pour son soutien inconditionnel;

A mes chères sœurs Safa, Marwa et Wiem;

A toute ma famille et ma belle-famille ;

… Je dédie ce travail

Remerciements

Au terme de cette expérience riche et passionnante, je tiens à exprimer ma reconnaissance et ma gratitude à mes directeurs de thèse Monsieur Kamel HAMROUNI, Professeur à l'Ecole Nationale d'Ingénieurs de Tunis (ENIT) et Monsieur Basel SOLAIMAN Professeur et Directeur du département Image et Traitement de l'Information à Telecom Bretagne, pour la qualité de leurs encadrements et orientations tout au long de l'élaboration de cette thèse.

J'adresse également mes remerciements à Monsieur Karim SAHEB ETTABAA, Maitre-assistant à l'Institut Supérieur d'Informatique et des Technologies de Communication de Hammam Sousse, pour sa disponibilité et ses précieuses instructions dans ce travail de recherche. Qu'il reçoit ici en un peu de mots toute ma gratitude.

Mes remerciements s'adressent également à Monsieur Ali KHENCHAF, Professeur à l'ENSTA Bretagne, pour l'honneur qu'il m'a accordé en présidant mon jury de thèse.

Je tiens à exprimer ma haute reconnaissance à Monsieur Jean Paul HATON, Professeur à l'Université Henri Poincaré Nancy 1, et Madame Amel GRISSA, Professeur à l'Institut Supérieur d'Informatique et de Gestion de Kairaouen, d'avoir accepté la fastidieuse tâche de rapporteurs. Je les remercie pour les remarques constructives qui m'ont permis de bien améliorer le présent manuscrit.

Je remercie également Monsieur Mohamed Mohsen GAMMOUDI qui m' a fait plaisir d'examiner ce travail.

Mes remerciements vont également à tous mes collègues du laboratoire Signal, image et traitement de l'information et à tous ceux qui m'ont aidé et encouragé tout au long de cette thèse.

Enfin, je tiens à montrer toute ma reconnaissance à mes parents qui ont permis l'aboutissement de mes "très" longues années d'étude. A toute ma famille et ma belle-famille pour leurs encouragements et soutien tout le long de ce travail.

Table des matières

Liste des Figures

Liste des Tableaux

Liste des Acronymes

ACR : American college of Radiology

API : Application Programming Interface

Bi-RADS : Breast Imaging Reporting And Data System de l' American College of Radiology

BRCA : Breast Cancer Gene

DL : Logique de description

ECD : d'extraction de connaissances à partir de données

FCMdd : Fuzzy C-Medoid

OWL: Web Ontology Langage

RA : Règle d'association

RDF : Resource Description Framework - Modèle de graphes pour la représentation d'ontologies.

RDFS : RDF Schema - Langage standard pour la modélisation d'ontologies sous forme de triplets.

RO : Règle ontologique

SWRL : Semantic Web Rule Language - Langage alliant la modélisation OWL-DL et les règles Rule ML.

W3C : World Wide Web Consortium - International community that standardizes web languages

XML : Extensible Markup Language - Langage informatique à balises.

Introduction Générale

Contexte du travail

L'accès aux informations mammographiques est un enjeu majeur pour les experts de santé ainsi que les ingénieurs de connaissance à cause de la complexité des relations significatives du domaine ainsi que le degré élevé d'exigence de spécifications. De ce fait, le besoin d'un support non contextuel et non ambigu des « connaissances mammographiques» permettant d'une part l'interopérabilité des informations et d'autre part l'organisation des objets du domaine dans une modélisation conceptuelle, est indispensable. Ceci demande, ainsi, d'extraire les informations qu'une mammographie peut porter et de représenter formellement son contenu pour une meilleure compréhension du domaine.

Une telle modélisation peut être assurée par une ontologie qui consiste en une spécification explicite d'une conceptualisation sous forme d'une hiérarchie des concepts associés avec des relations. A ses intérêts génériques, une ontologie dans le contexte de la modélisation des connaissances mammographiques permet de: (i) établir une terminologie commune du vocabulaire spécifique admis par la communauté du domaine, (ii) représenter les connaissances afin d'être intégrées dans des systèmes à base de connaissances, indépendamment de l'hétérogénéité des sources de connaissances et (iii) palier au problème du fossé sémantique par la mise en correspondre des caractéristiques visuelles des objets radiologiques (forme, textures … etc) avec leurs interprétations sémantiques correspondantes.

Dans ce contexte, plusieurs ontologies dans le domaine mammographique ont été proposées dans la littérature. Bien que ces modélisations soient riches en information, elles sont plus ou moins complètes par rapport aux connaissances liées à la mammogrpahie (en termes de concepts pertinents du domaine ainsi que les relations significatives les reliant). Une des raisons principales de ce fait est que ces bases de connaissances sont conçues d'une façon indépendante, représentant de différentes perspectives de domaine.

En revanche, l'amélioration des technologies d'information rend disponible un grand volume des données dont l'exploitation devient un enjeu majeur. Aujourd'hui, dans les services de radiologies, toutes les mammographies diagnostiquées (constituant des expériences passées) sont présentées, chacune, sous la forme d'un dossier médical comportant la description du cas associée avec le diagnostic approprié. Ces grandes quantités d'information recueillies disposent d'un fort potentiel pour révéler de nouvelles connaissances implicites. Néanmoins, elles nécessitent des méthodologies de gestion pour être utiles et généralisées.

Problématiques

De ce fait, la gestion de ces connaissances constitue un facteur essentiel dans de nombreux domaines (particulièrement le domaine médical); elle doit rendre possible un processus de création de valeur à partir des différentes sources de connaissances disponibles sous des formes divergentes: les modélisations de connaissances existantes, les données antérieurement

diagnostiquées, les connaissances de l'expert, etc. Il existe, en conséquence, un besoin réel de mettre en œuvre un processus de fouille de connaissances pour la découverte et la représentation des connaissances d'intérêt au sein d'une base de connaissances cohérente, exhaustive et formelle. A travers cette thèse, ce processus est employé principalement sur les deux sources de connaissances : les ontologies existantes et les données du domaine. Les connaissances découvertes servent pour l'enrichissement d'une ontologie existante.

La fouille de connaissances est fortement limitée aux divergences des modèles de connaissances employés. Même les ontologies représentant le même domaine, sont hétérogènes. En effet, chacune, est conçue différemment par différents ingénieurs et pour des objectifs distincts. De ce fait, l'unification de ces bases de connaissances nécessite le traitement des différentes formes d'hétérogénéités. Toutefois, la complexité du domaine d'étude et la taille des ontologies rendent la tâche d'unification de ces bases de connaissances beaucoup plus complexe. Ceci nous amène à nous poser la question : Comment peut-on réduire la complexité de mise en correspondance des ontologies ?

Les grandes quantités d'informations recueillies dans les systèmes d'acquisition portent des connaissances utiles pour la compréhension du domaine mais inexploitables à leurs états bruts. En effet, ces données sont contextualisées et nécessitent des méthodologies de fouille pour pouvoir les employer. Toutefois, les techniques actuelles d'extraction de connaissances présentent des inconvénients à savoir : les nouvelles connaissances déduites se présentent en des grandes quantités d'où la difficulté d'étudier leur pertinence et leur utilité. Ceci nous amène à se poser la question : Comment peut-on exploiter et adapter les connaissances découvertes afin d'être intégrées dans les base de connaissances globale?

Motivations

Ce travail de recherche vise la mise en œuvre d'un processus de fouille de connaissances permettant la découverte et la représentation des connaissances d'intérêt au sein d'une base de connaissances ontologique cohérente, exhaustive et formelle. L'approche proposée est basée sur le développement d'une ontologie du domaine basée sur l'enrichissement *conceptuel* et *relationnel* d'une ontologie noyau. Ceci en exploitant de manière conjointe deux courants scientifiques considérés comme étant deux sources de connaissances possibles à notre problématique d'enrichissement. Le premier type d'enrichissement concerne la fouille de connaissances des ontologies de domaine pour l'extraction de nouveaux concepts différents des entités de référence. Le deuxième type d'enrichissement se base sur l'extraction des corrélations entre les attributs (ou les concepts initiaux de l'ontologie) mis en jeux dans les données recueillies.

Dans une première étape, nous proposons une méthode d'enrichissement d'ontologie fondée sur l'utilisation des bases de connaissances ontologiques. L'essence de cette méthode se base sur une approche d'alignement d'ontologies permettant d'unifier les différentes représentations. En effet, ce processus permet de sélectionner une ontologie source et l'enrichir avec de nouvelles connaissances à partir d'autres ontologies locales. Néanmoins, à cause de la voluminosité et la sensibilité des ontologies d'étude, la phase d'alignement devient de plus en plus compliquée. Pour ce fait, nous avons proposé d'introduire une étape de pré-

alignement des ontologies à manipuler. Cette étape consiste à réorganiser la structure définissant l'ontologie et regrouper les concepts sémantiquement liés en des clusters flous qui sont organisés en une structure hiérarchique des clusters. Cela nous a permis non seulement de simplifier la visualisation hiérarchique des connaissances mais également de réduire le problème de recherche de correspondances en transformant le problème d'alignement de deux ontologies entières en alignement des paires de clusters conceptuels de tailles réduites.

Une fois l'étape de pré-alignement (appelée aussi l'étape de préparation d'ontologies) est achevée, nous procédons à l'étape d'alignement qui se compose de deux phases : la phase d'ancrage et la phase de dérivation. La phase d'ancrage permet de déterminer et rapprocher les clusters les plus similaires de deux ontologies (source et cible), quant à la phase de dérivation ; elle permet d'identifier les différents alignements entre les éléments des clusters similaires à travers l'utilisation de différentes techniques de similarité. La phase d'alignement a pour objectif de découvrir les nouvelles connaissances qui sont rajoutées dans l'ontologie cible afin d'obtenir une ontologie de domaine exhaustive.

Dans une deuxième étape, nous avons proposé d'exploiter les évidences provenant des données des patients antérieurement diagnostiqués dans le but de les transformer en de nouvelles connaissances qui servent de support d'enrichissement. Ce processus d'extraction de connaissances se base sur deux étapes : i) la modélisation des expériences recueillies afin de faciliter leur exploitation ultérieure, ii) l'application de la technique de fouille de données: *les règles d'association* afin d'extraire de nouvelles connaissances sous la forme de règles. Ce type de règles permet non seulement d'identifier les cooccurrences des items mais aussi les relations d'implications entre les observations dans une mammographie et la classification correspondante. Toutefois, le nombre de règles extraites est tellement volumineux que la tâche d'analyse devient impossible pour les experts du domaine. A cette fin, nous proposons une méthode de fouille de connaissances des règles générées. Cette méthode se base, d'une part sur le filtrage des connaissances connues au préalable, d'autre part, le *ranking* des règles afin d'aider l'expert à se focaliser sur les connaissances de grande importance. Les deux actions proposées se basent sur l'utilisation de l'ontologie du domaine afin de guider la tâche d'extraction des connaissances utiles. Ceci, nous amène à comparer les deux sources de connaissances pour élaguer les règles déjà connues. L'importance des règles restantes est évaluée en utilisant une nouvelle distance sémantique basée sur la structure hiérarchique des clusters conceptuels proposée dans la première partie de notre travail. La sortie de cette étape consiste à une liste de règles filtrées et classées qui sera présentée à l'expert du domaine, qui seul peut juger la pertinence de ces connaissances. Finalement, les sous-ensembles de relations validées seront employés pour l'enrichissement de l'ontologie noyau.

Ainsi, la complémentarité des deux approches permet d'enrichir l'ontologie du domaine au niveau relationnel et conceptuel avec de nouvelles connaissances fiables et prouvées par l'expert de domaine.

Concernant le processus d'enrichissement, nous faisons une place aux experts du domaine, des personnes ayant une grande compréhension du sens des connaissances représentées. Afin de préserver la cohérence de l'ontologie, les résultats validés par l'expert sont introduits manuellement dans l'ontologie de référence.

Contributions

L'objectif de notre thèse consiste en un processus de couplage entre le processus d'enrichissement conceptuel et le processus d'enrichissement relationnel d'une ontologie mammographique cible.

Les contributions apportées dans le cadre de ce travail sont :

Dans le volet « enrichissement relationnel »:

- Le pré-alignement ou la préparation d'ontologies basé sur le clustering hiérarchique flou des concepts de l'ontologie et ayant pour objectif la réorganisation de la structure ontologique de manière à grouper les concepts similaires dans des clusters répartis sur différents niveaux de granularité.
- L'alignement des ontologies en se basant sur les structures hiérarchiques des clusters qui leurs correspondent.
- L'enrichissement de l'ontologie de référence par de nouveau concepts.

Dans le volet « enrichissement relationnel »:

- L'extraction des connaissances à partir des données sous la forme de règles,
- La confrontation des deux sources de connaissances, provenant de l'ontologie de référence et des règles d'association extraites, afin d'élaguer les règles portant des connaissances déjà connues
- L'évaluation et le *ranking* des règles d'association en utilisant des mesures d'intérêts. Le calcul de la mesure d'intérêt est basé sur une nouvelle distance sémantique exploitant la structure hiérarchique des clusters proposée dans le premier volet.
- L'enrichissement de l'ontologie par de nouvelles relations.

Plan du rapport

Cette mémoire de thèse est structurée en quatre chapitres :

Le **chapitre I**, présente une analyse l'état de l'art concernant les généralités du courant scientifique employé dans ce travail : les *ontologies*. Nous présentons tout d'abord ses notions générales, ses composantes et notamment les processus de construction des bases ontologiques. Une attention particulière a été portée aux ontologies relatives au domaine mammographique. Nous exposons ensuite le processus d'alignement des ontologies qui sert de base à notre proposition. Nous révélons un état de l'art sur les techniques d'alignement proposées dans la littérature, tout en mettant l'accent sur les méthodes qui se sont intéressées à réduire les espaces de recherche des correspondances des entités relatives aux ontologies à aligner.

Le **chapitre II** aborde le processus d'extraction de connaissances (appelées aussi processus de fouille de données), ses étapes et ses principaux outils permettant d'analyser les

bases de données. Une attention particulière a été portée aux règles d'association qui servent de base à notre proposition. Nous définissons, dans un premier temps, leurs principes et les algorithmes d'extraction. Ensuite, nous présentons les méthodes proposées pour l'analyse et le post traitement des règles produites.

Le **chapitre III** présente notre démarche originale de fouilles de connaissances qui s'articule autour d'un couplage entre le processus d'enrichissement conceptuel de l'ontologie de référence à partir des bases ontologiques existantes et le processus d'enrichissement relationnel à partir d'une base de données du domaine.

Nous présentons dans un premier temps la démarche d'enrichissement conceptuelle proposée. Cette partie se base sur trois étapes : (i) Le pré-alignement ou la préparation d'ontologies : cette étape se base sur le clustering hiérarchique flou des concepts de l'ontologie. Elle a pour but la réorganisation de la structure ontologique de manière à grouper les concepts similaires dans des clusters répartis sur différents niveaux de granularité. (ii) L'alignement des ontologies en se basant sur les structures hiérarchiques des clusters qui leurs correspondent. (iii) l'enrichissement de l'ontologie de référence par de nouveau concepts.

Dans un deuxième temps, nous exposons les étapes de notre méthodologie pour l'enrichissement relationnel d'ontologie à partir d'une base de données. Cette partie se base, d'une manière générale, sur trois étapes : (i) l'extraction des connaissances à partir des données sous la forme des règles, (iii) la confrontation des deux sources de connaissances (provenant de l'ontologie de référence et les règles d'association extraites) afin d'élaguer les règles portant des connaissances déjà connues (iii) l'évaluation et le ranking des RAs en associant des mesures d'intérêts. Le calcul de la mesure d'intérêt est basé sur une nouvelle distance sémantique en utilisant sur la structure hiérarchique des clusters déterminée dans la première partie. (iiii) l'enrichissement de l'ontologie par de nouvelles relations.

Finalement, le **chapitre IV**, aborde le contexte d'application qui sert de cadre à nos travaux de recherche et expose le domaine mammographique comme étant le cas d'étude permettant d'instancier la méthodologie proposée dans le chapitre III. La première partie s'intéresse à l'application de notre démarche proposée à des données réelles (à savoir les ontologies mammographiques et les données mammographiques collectées des hôpitaux Taher Sfar de Mahdia et Ben Arous de Tunis). La deuxième partie est consacrée à l'évaluation de notre approche générale en la comparant à des méthodes existantes et en utilisant des métriques d'évaluation standard.

La conclusion de ce manuscrit permet de tirer un bilan préliminaire de notre étude, et de dresser des perspectives pour l'avenir proche de cet axe de recherche. La figure ci-dessous résume la structure et le contenu de chaque chapitre dans ce manuscrit.

I. Chapitre 1 : Les Ontologies, Les méthodes de construction d'Ontologies et Les Ontologies Mammographiques

Sommaire :

1. Introduction

Avec l'augmentation des données dans le domaine médical, le besoin d'un support non contextuel et non ambigu des connaissances permettant d'une part l'interopérabilité des informations et, d'autre part, l'organisation des objets du domaine dans une modélisation conceptuelle est indispensable. Une telle modélisation peut être assurée par une ontologie qui apporte une spécification explicite d'une conceptualisation sous forme d'une hiérarchie des concepts associés avec des relations. A ces intérêts génériques, une ontologie dans le contexte de la modélisation des connaissances en imagerie médicale permet également d'établir une terminologie commune du vocabulaire spécifique admis par la communauté du domaine, de représenter les connaissances afin d'être intégrées dans des systèmes à base de connaissances, indépendamment de l'hétérogénéité des sources de connaissances, et de palier aux problèmes du fossé sémantique par la mise en correspondance des caractéristiques visuelles des objets radiologiques (forme, textures, etc) avec leurs interprétations sémantiques correspondantes.
Dans ce chapitre, nous abordons, dans un premier temps, la notion d'ontologie et ses composants. En deuxième temps, nous présentons les méthodes de construction d'ontologies, particulièrement les travaux d'alignement d'ontologies comme solution au problème d'interopérabilité. Pour finir, nous exposons un état de l'art sur les ontologies existantes liées au domaine mammographique.

2. Les ontologies

En Intelligence Artificielle, particulièrement en ingénierie de connaissances, l'ontologie est introduite pour traiter la représentation et la manipulation des connaissances d'un domaine spécifique pour le partage des connaissances de ce domaine au sein des systèmes informatiques.

2.1 Définitions

Plusieurs définitions ont été attribuées à la notion d'Ontologie. Dans [Neeches, et al., 1991], les auteurs furent les premiers à introduire une définition formelle de l'ontologie : « *une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire*».

[Gruber, 1993] définit une ontologie comme étant : «*Une spécification explicite d'une conceptualisation* ». Cette définition est la plus communément admise où la conceptualisation se réfère à une structure abstraite de concepts pertinents au domaine et associés à des relations explicitement définies.

[Borst, 1997] affine la définition de Gruber en décrivant : «*une ontologie comme une spécification explicite et formelle d'une conceptualisation partagée* » où la spécification formelle

signifie que l'ontologie est compréhensible sans ambiguïté par les humains ainsi que les machines. Une ontologie est une conceptualisation qui répond aux critères suivants:

-« **Formelle** » : L'ontologie doit être introduite par une représentation explicite, formelle, axiomatique et systématique des propriétés du modèle permettant un certain niveau de raisonnement automatique.

-« **Explicite »**: Les concepts, les contraintes et les relations sont explicitement définis.

-«**Consensuelle**» : L'ontologie définie reflète un point de vue général accepté par une communauté. Elle pourra être partagée et communément utilisée.

[Sowa, 1995] définit le principe de construction d'ontologie comme étant la catégorisation de concepts liés au domaine d'intérêt, de manière à obtenir un catalogue de types de choses existant définit dans un langage formel.

Bien que les définitions liées à la notion d'ontologie se diversifient, elles représentent des points de vue complémentaires. Ainsi, on peut généraliser la définition d'une ontologie comme une organisation hiérarchique des concepts de domaine d'intérêt associés à des propriétés sémantiques dans un langage de modélisation formel. Le but d'une ontologie est la spécification des connaissances de domaine.

2.2 Composants d'une ontologie

Afin de pouvoir spécifier les connaissances du domaine d'étude, une ontologie est constituée de certains composants à savoir : les concepts, les relations, les axiomes et les individus [Sowa, 1995]. La Figure I.1 présente des concepts de l'ontologie «Famille » et leurs interrelations.

2.2.1 Les concepts

Les concepts désignent les entités/objets portant les connaissances à spécifier. Dans la Figure I.1, Humain, Homme et Femme représentent les concepts dans l'ontologie. On distingue trois types de concepts : une intension, une extension et une terminologie.

-Une intension désigne formellement un concept.
-Une extension désigne l'ensemble des entités que le concept englobe. Ces entités se caractérisent par les propriétés définies par l'intention.
-Une terminologie désigne les labels (tel que les synonymes) qui lexicalisent le terme du concept. Elle est utilisée lorsque le sens du terme employé est ambigu.
Deux concepts peuvent avoir les propriétés suivantes :
-Equivalence : Deux concepts ayant la même sémantique sont dits équivalents.
-Disjonction : Si deux concepts sont sémantiquement inverses, alors ils sont dits disjoints.

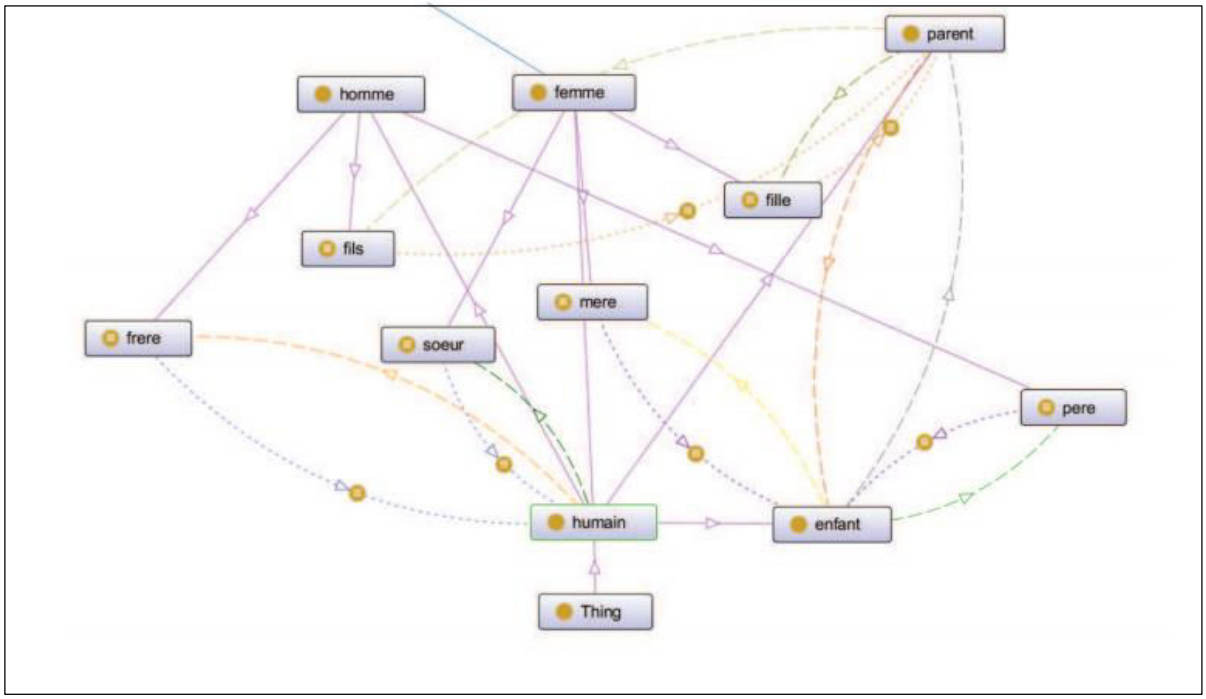

Figure I.1: Exemple de taxonomie d'une ontologie [Sowa, 1995]

2.2.2 Les relations

Les relations constituent l'ensemble des liens établis entre les concepts afin de définir des interactions binaires entre eux [Sowa, 1995]. Les liens familiaux, dans Figure I.1 sont exprimés par des relations. Par exemple, la relation *a_parent* relie les concepts enfant vers parent. On distingue deux types de relations : Les relations taxonomiques et les relations associatives.

- **Les relations taxonomiques** : Elles désignent les relations de subsomption entre les concepts de la structure taxonomique. Par exemple, un concept C_i subsumé par un autre C_j possède les mêmes propriétés que C_j.

- **Les relations associatives** : Ce sont des relations non taxonomiques, appelées « Object-property» et désignent des interactions binaires entre les concepts telles que : « co-occure_avec», « lié_à», etc.

2.2.3 Les axiomes

Les axiomes permettent de spécifier les interprétations sémantiques des composants de l'ontologie. Elles représentent des assertions ou des propriétés qui sont toujours vraies définies sur les concepts et/ou les relations. Les axiomes désignent les connaissances acceptées par défaut (élémentaire) qui ne peuvent être explicitement définies par les concepts et/ou relations. Elles peuvent également modéliser des contraintes sur les attributs (leurs valeurs). Dans la Figure I.1, les concepts Homme et Femme sont définis comme disjoints, grâce à un axiome, ce qui implique qu'aucune instance ne peut appartenir à la fois aux classes Femme et Homme.

2.2.4 Les individus

Les individus sont utilisés pour représenter les concepts spécifiques / singuliers des concepts génériques. On peut également modéliser les instances de relations entre les instances de concepts.

2.3 Les types d'ontologies

Les ontologies sont catégorisées en différentes classes selon certains critères : les connaissances définies dans l'ontologie, et le niveau de détail.

2.3.1 Catégorisation selon le contenu des connaissances

Les ontologies sont catégorisées par [Guarino, 1997], selon les objets de conceptualisations définis dans l'ontologie (Figure I.2).

-Les ontologies de haut niveau : Ces ontologies sont appelées aussi des ontologies supérieures ou abstraites. Ce type d'ontologie comporte les concepts généraux subsumant des concepts plus spécifiques existants dans les sous domaines. Ce type d'ontologie sert à restreindre les incohérences des entités les plus spécifiques à travers une structure de représentation des concepts génériques.

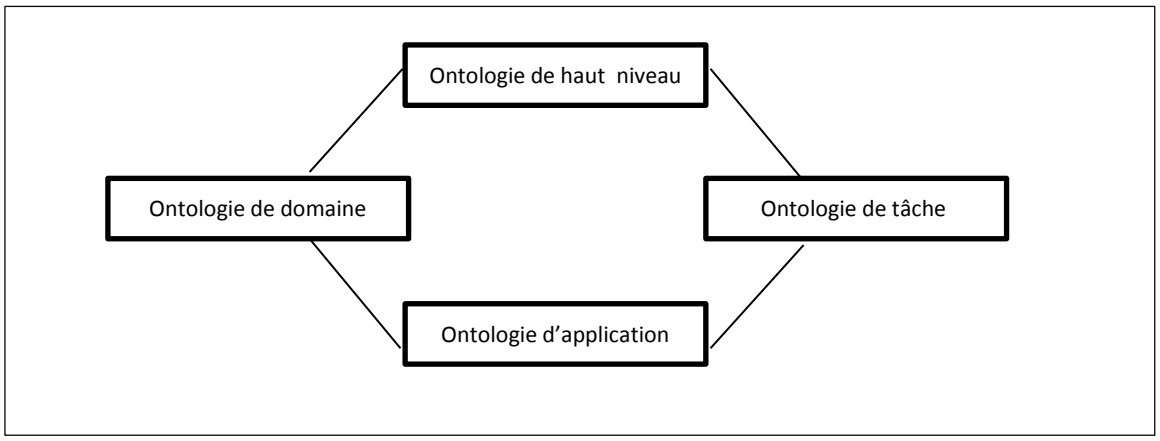

Figure I.2: Classification des ontologies selon Guarino

- **Les ontologies de domaine** : Contrairement aux ontologies supérieures, ce type d'ontologies porte des connaissances sur un domaine spécifique en définissant les concepts ainsi que les relations primitives régissant ce domaine. Les concepts (respectivement des relations) déclarés dans l'ontologie de domaine, représentent des spécifications des concepts (respectivement des relations) de l'ontologie de haut niveau.

-**Les ontologies de tâche** : Ces ontologies sont appelées Domain-Task Ontologies. Ces ontologies sont encore plus spécifiques que les ontologies de haut niveau ainsi que les ontologies de domaine. Elles modélisent le vocabulaire systématique nécessaire à des tâches/activités spécifiques ou des résolutions à des problèmes particulières indépendamment de domaine.

-Les ontologies d'application : Le vocabulaire défini dans l'ontologie est entièrement dépendant du domaine d'application ainsi qu'une tache spécifique. L'ontologie d'application est à la fois une spécialisation des ontologies de domaine et des ontologies de tâches. Ce type d'ontologie est le plus spécifique.

2.3.2 Catégorisation selon le niveau de granularité

[Guarino, 1997] définit deux catégories des ontologies selon le niveau de détail ou de granularité.

- **Les ontologies de degré de détails fins** : Ce type d'ontologies modélise des descriptions des connaissances de domaine très détaillées/ spécifiques à travers la définition des concepts et des relations pertinents du domaine d'étude.

-**Les ontologies de degré de détails faibles** : Ce type d'ontologies dispose de niveaux de granularité moins détaillés tels que les ontologies de haut niveau dont les concepts et les relations régissant le domaine sont génériques et abstraits.

3. Les méthodes de construction d'ontologies

Dans la littérature, trois types de conception d'ontologies ont été proposés: la conception entièrement manuelle ('From scratch'), la conception basée sur l'apprentissage d'ontologie et la construction d'ontologies basée sur la fusion ou l'enrichissement d'ontologies (voir Figure I.3).

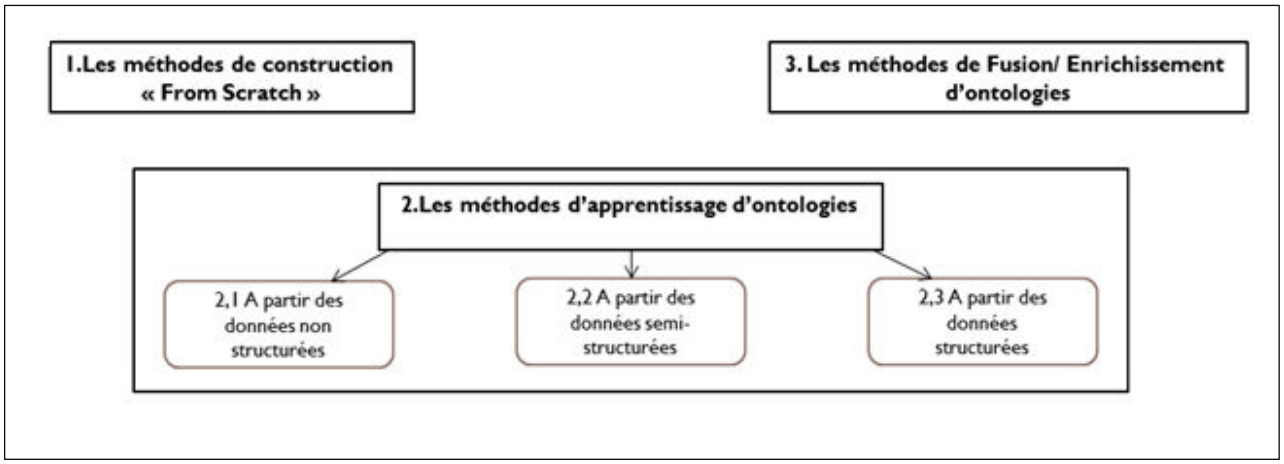

Figure I.3: Classification des méthodes de construction d'ontologies

Bien que plusieurs principes et méthodes aient été définis pour faciliter la construction manuelle des ontologies, ce moyen reste très coûteux en termes de temps et pose ultérieurement de nombreux problèmes tels que les problèmes de maintenance, de validation et de mise à jour. Quant aux méthodes d'apprentissage d'ontologie, elles se basent sur l'exploitation des données existantes (textes, images, documents). Le troisième type respectif aux méthodes de fusion/enrichissement est fondé sur l'exploitation des ressources ontologiques existantes liées au domaine d'intérêt. Chacun de ces types de conception dispose

d'un certain nombre de méthodologies. Dans ce qui suit, nous allons définir ces différents types en illustrant les méthodes appropriées proposées dans la littérature.

3.1 Les méthodes de conception « From Scatch »

Plusieurs méthodes pour la construction manuelle de l'ontologie ont été proposées dans la littérature, telles que : "Enterprise Ontology", On-To-Knowledge (OTK), STANFORD, Kactus, TOVE, SENSUS. Dans ce qui suit, nous allons décrire les méthodes les plus citées dans la littérature.

3.1.1 La méthodologie « Entreprise Ontology »

Les auteurs Ushold et al. [Uschold, et al., 1995] ont été les premiers à proposer une méthode de construction d'ontologies basée sur l'expérience du développent de l'ontologie d'Entreprise. Cette méthodologie se base sur les étapes suivantes :

-Spécification du contexte de l'ontologie et le but de sa construction.
-Formalisation des concepts et relations pertinents au domaine.
-Evaluation de l'ontologie proposée.

3.1.2 La méthodologie On-To-Knowledge (OTK)

Dans [Maedche, et al., 2001], les auteurs ont proposé une méthode itérative de développement d'ontologie. Elle commence par l'acquisition des connaissances sous une forme informelle jusqu'à la production d'une ontologie répondant aux exigences. Cette méthodologie se base sur les étapes suivantes :

- Spécialisation du contexte, sources de connaissances, parties prenantes, objectifs et directives de l'ontologie.
- Spécialisation d'une ontologie générique.
- Raffinement de l'ontologie jusqu'à atteindre l'objectif souhaité.
- Evaluation et maintenance afin de contrôler de la qualité de l'ontologie obtenue.

3.1.3 La méthodologie SENSUS

Dans [Swartout, et al., 1997], les auteurs ont proposé une nouvelle méthode de construction d'ontologies partant d'une plus grande appelée l'ontologie SENSUS par l'extraction des termes pertinents qui se rapportent au domaine et éliminant les termes inutiles. Cette méthodologie se base sur les étapes suivantes :

- Identification des termes pertinents au domaine.
- Mapping des termes clés à SENSUS.
- Enrichissement par de nouveau concepts.
- Validation de l'ontologie.

3.1.4 La méthodologie TOVE

Dans [Grüninger, et al., 1995], les auteurs ont proposé une méthode de construction d'un modèle logique de connaissances du projet TOVE[1] (TOrento Virtual Enterprise) en se basant sur les étapes suivantes :

- Evocation des problèmes liés à l'application en question.
- Formulation des questions auxquelles l'ontologie à construire doit répondre.
- Identification d'une terminologie à partir des objets figurant dans les questions.
- Modélisation formelle des définitions des termes de la terminologie identifiée.
- Evaluation de l'exhaustivité de l'ontologie.

3.2 Les méthodes de conception d'ontologie basées sur l'apprentissage

De nombreuses méthodes basées sur l'apprentissage ont été proposées dans la littérature. Ces méthodes visent à automatiser la tâche de construction d'ontologie par l'utilisation de certains supports sémantiques afin de réduire les temps alloués aux différentes phases de développement et améliorer la qualité de l'ontologie obtenue. Pour ce fait, dans cette section, nous proposons de catégoriser ces méthodes selon le support d'apprentissage utilisé pour la conception, à savoir: Les données non structurées, les données semi-structurées et les données structurées. Nous présentons par la suite, chacune de ces méthodes en utilisant quelques approches proposées dans la littérature.

3.2.1 Construction d'ontologies à partir des données non structurées

Etant donné la richesse d'informations existantes dans les données non structurées telle que les textes, plusieurs méthodes s'en sont servies pour la construction d'ontologie. Partant d'un corpus de documents, [Bendaoud, et al., 2007] a proposé une nouvelle méthode de construction d'ontologie liée au domaine astronomique. Cette méthode se base sur les techniques d'Analyse Formelle du Concept (AFC) d'Analyse Relationnelle des Concepts (ARC). Elle repose sur cinq étapes :

- L'identification des objets qui sont en relation avec le domaine d'intérêt à partir du corpus,
- La classification de ces objets selon les propriétés qu'ils partagent pour la construction d'un treillis de concept,
- La construction de noyau de l'ontologie à partir des résultats de la deuxième étape,
- La détection des relations transversales par la technique d'ARC,
- L'enrichissement de l'ontologie.

Cette méthode est limitée juste à l'extraction de triplets (Sujet, Verbe, Complément) et ne garantit pas la génération d'une structuration hiérarchique des relations entre les concepts. Dans [Wong, et al., 2012], les auteurs présentent une technique pour la construction automatique d'une ontologie en se basant sur la classification hiérarchique du document. Le

[1] http://www.eil.utoronto.ca/tove/ontoTOC.html

procédé s'applique sur un ensemble de textes formant un corpus de documents de domaine et crée une structure hiérarchique (arbre), où à chaque nœud est associé un ensemble de termes dérivés des vecteurs caractéristiques de document. Chaque valeur du vecteur caractéristique représente un poids pour mesurer l'importance de la propriété dans la description du document cible. A partir de ces vecteurs, une matrice de similarité basée sur la distance entre les paires de documents est construite. Cette matrice est utilisée pour construire les clusters. Le modèle de cluster résultant est un arbre binaire hiérarchique dans lequel les nœuds représentent les clusters obtenus et les feuilles sont les événements classés. Cet arbre binaire sera converti ensuite en une taxonomie de concepts.

La qualité de l'ontologie produite dépend étroitement de la qualité des documents textuels utilisés. Si ces derniers n'assurent pas la couverture du domaine en terme de concepts pertinents, l'ontologie générée peut être significative.

3.2.2 Construction des ontologies à partir des données semi-structurées

Les données semi-structurées constituent également une source de connaissances à exploiter. Plusieurs auteurs se sont basés sur ce type de données pour la construction d'ontologies à savoir : HTML, XML, RDF, bases de données, etc.

Dans [Touma, et al., 2015], les auteurs présentent deux méthodes génériques de construction d'ontologies à partir de documents semi-structurés XML. La technique la plus simple consiste à créer un mapping direct entre les balises XML et les éléments OWL mais cette technique est très limitée dans les formats de schéma XML en raison de la nature primitive des correspondances. Une deuxième méthode plus sophistiquée consiste à extraire les concepts de l'ontologie à partir de balises XML et de créer, ensuite une hiérarchie des concepts extraits en utilisant les techniques syntaxiques.

Dans [Timon, et al., 2009], les auteurs exploitent les données semi structurées HTML pour la construction d'ontologies. Le processus proposé pour l'extraction des connaissances comporte six étapes, à savoir :

- La préparation des pages web,
- La transformation de ces pages en éliminant les non pertinentes,
- Le regroupement des données en clusters de termes similaires,
- La reconnaissance des modèles des pages web clustérisés,
- Le raffinement de l'ontologie en annotant les clusters obtenus,
- La révision de l'ontologie par des ingénieurs de connaissances pour la validation de son contenu.

L'ontologie extraite fournit des informations structurées et pertinentes pour des applications telles que le commerce électronique et la gestion des connaissances.

3.2.3 Construction des ontologies à partir des données structurées

Plusieurs méthodes ont été basées sur les données structurées pour la construction d'ontologie. Parmi les données structurées, on cite les bases de données relationnelles qui assurent l'accessibilité et l'évolutivité des informations utiles enregistrées.

Dans [Pasha, et al., 2012], les auteurs proposent une méthode pour l'acquisition des ontologies à partir des bases de données relationnelles. L'approche proposée commence par l'extraction des informations de schéma de base de données relationnelle, telles que les noms de relation, les noms d'attributs, les clés primaires, les clés étrangères et les contraintes d'intégrité. Puis, on procède à l'analyse des données extraites afin d'identifier les concepts de l'ontologie. En troisième lieu, les tuples de base de données relationnelle sont récupérés pour aboutir enfin aux alignements entre ces triplets et les instances ontologiques. L'auteur se sert, alors des règles pour extraire la structure ontologique et transformer les triplets de base de données de relations en des instances ontologiques.

3.3 Les méthodes de Fusion/ intégration d'Ontologies

Les méthodes de fusion/intégration ou enrichissement d'ontologies constituent des processus qui se basent sur l'exploitation des ressources ontologiques sources pour l'extraction ou l'obtention d'une ontologie plus exhaustive. Partant d'un ensemble des ontologies locales indépendamment développées par différentes communautés (portant sur un domaine d'intérêt), une ontologie globale est produite [Catarci, et al., 1993]. A ce stade, on distingue dans la littérature deux méthodes d'ingénierie ontologique : *les méthodes d'intégration* et *les méthodes d'enrichissement*. Les méthodes d'intégration différencient entre les ontologies sources et l'ontologie générée. Dans ce cas, une nouvelle ontologie sera produite en assemblant, adaptant d'autres ontologies disponibles. Pour les méthodes d'enrichissement [Fareh, et al., 2013], on considère une ontologie locale comme étant une ontologie cible qui sera enrichie par de nouvelles connaissances extraites à partir des ontologies locales sources.

Parmi les méthodes de fusion d'ontologies existantes on cite 'OntoDNA' [Kiu, et al., 2007], DKP-AOM [Fahad, et al., 2007], Prompt [Noy, et al., 2000], Chimaera [McGuinness, et al., 2000], [Raunich, et al., 2011], [Maiz, et al., 2010], HCONE [Kotis, et al., 2006].

Les deux méthodes existantes requirent l'établissement des mappings/alignements sémantiques entre les ontologies locales, par le biais d'un processus appelé processus d'alignement d'ontologies. Les mappings (les interopérabilités sémantique) réfèrent à des mises en correspondances ou liaisons entre les entités (concepts, relations, instances) d'ontologies en jeux. Partant de deux ontologies locales, l'alignement produit un ensemble de correspondances, chacune liant deux entités par une relation (équivalence, subsomption, incompatibilité, etc.), à laquelle on associe un degré de similitude [Farah, et al., 2008].

4. Alignement d'ontologies

La réutilisation des ressources ontologiques est limitée à la diversité des ressources ainsi que l'hétérogénéité de leur contenu ce qui nécessite une réflexion approfondie sur la manière

de les réutiliser. A ce stade l'alignement des ontologies intervient comme solution d'interopérabilité des ressources sémantiques indépendantes (Figure I.4). La recherche de correspondances entre les ontologies a fait l'objet de nombreux travaux. Dans cette section, nous définissons les notions d'alignement ainsi que ses techniques standards, et nous présentons, par la suite les méthodes d'alignement proposées dans la littérature.

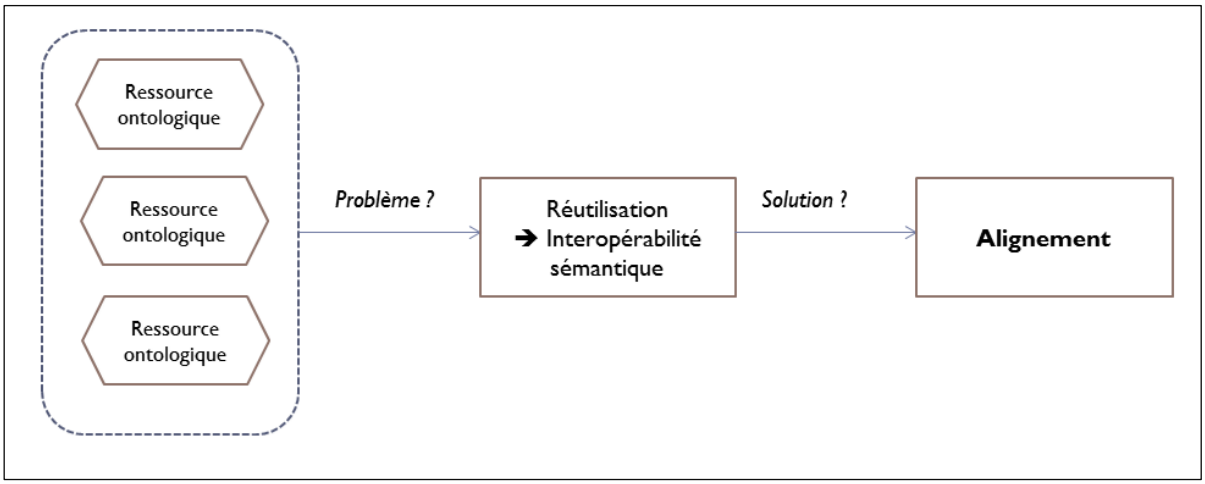

Figure I.4: Alignement comme solution pour l'interopérabilité sémantique

4.1 Principe et défis

Plusieurs définitions ont été proposées pour l'alignement d'ontologies. La définition la plus partagée est celle de [Sampson, 2007]: «*Ontology alignment is the process where for each entity in one ontology O_1we try to find a corresponding entity in the second ontology $O_2$with the same or the closest meaning*» (Figure I.5).

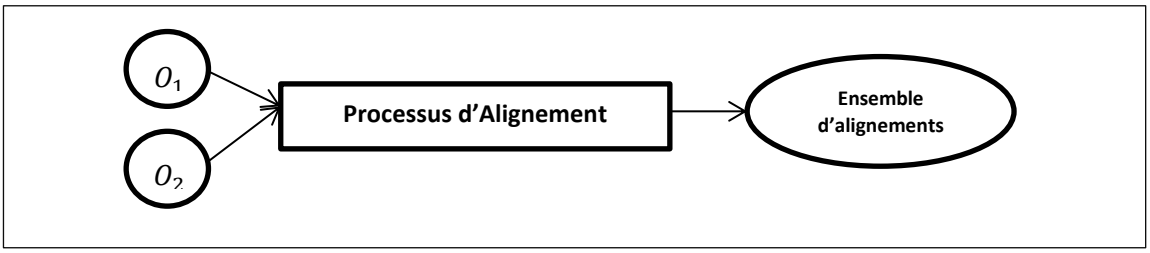

Figure I.5: Processus d'Alignement

L'étude d'alignement d'ontologies en tant que discipline d'ingénierie ontologiques a été intensément abordée par les chercheurs cette dernière décennie [Kandpal, et al., 2014]. En effet, le développement de plusieurs ontologies portant sur le même domaine de manière indépendante et avec différentes perspectives et points de vue conduit à l'apparition de plusieurs formes d'hétérogénéité ente les différents modèles conceptuels ce qu'on appelle le problème d'interopérabilité. Pour cela, le processus d'alignement vise à pallier aux différentes formes d'hétérogénéité classées selon [Bouquet, et al., 2004] comme suit:

-Hétérogénéité syntaxiques: Ce type d'hétérogénéité concerne les formats de représentation des modèles conceptuels. En effet, il existe plusieurs langages de représentation d'ontologie tels que : XML, RDF, KIF, OWL qui se caractérisent par différentes syntaxes.

-**Hétérogénéité terminologiques :** Ce type d'hétérogénéité est dû à la nomination des entités dans les ontologies, par exemple : la synonymie, le langage utilisé, les abréviations, etc.

-**Hétérogénéité conceptuelle :** Chaque ontologie est développée dans un contexte spécifique, pour un objectif particulier, ce qui induit à la création des ontologies dont les contenus sont différents en termes de couverture, de granularité et de perspectives de modélisation.

-**Hétérogénéité sémiotique/ pragmatique:** Dans différentes conditions, les communautés peuvent interpréter et analyser différemment une même ontologie.

Le processus d'alignement se compose de différentes étapes partagées par la majorité des méthodes proposées dans la littérature (Figure I.6).

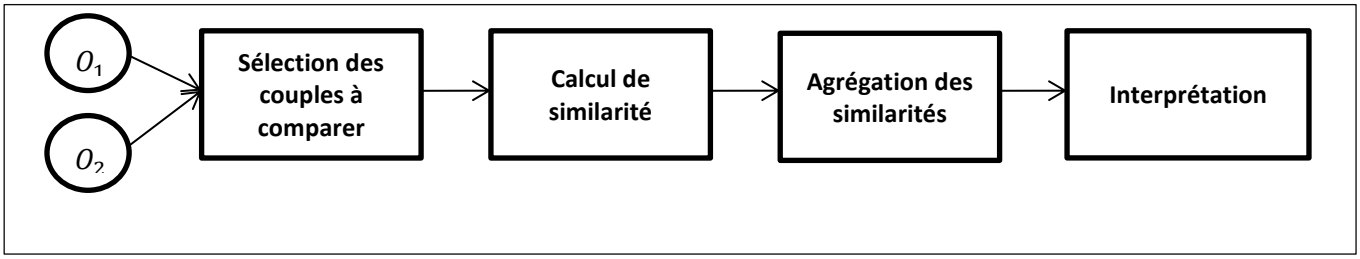

Figure I.6: Les étapes du processus d'alignement

La première étape consiste à unifier les formats des ontologies (source et cible) d'intérêt et définir les entrées du processus. La sélection des éléments en entrée peut considérer toutes les entités ontologiques (concept, relation, etc.) ou certains d'entre elles. La comparaison de ces couples dans la 2ème étape se base sur l'utilisation des techniques de similarité renvoyant des valeurs de similarité indiquant à quel degré un couple d'entités est similaire. Dans le cas d'utilisation de plusieurs techniques pour un même couple d'entités, les valeurs générées sont agrégées afin de ne renvoyer qu'une seule valeur qui sera interprétée à la fin du processus. En sortie, le processus génère, un ensemble d'alignements générés sous la forme $<id, e, é, n, Rel>$ où :

-id : l'identifiant de l'alignement généré.
-e et é : représentent les entités mises en correspondance et provenant de l'ontologie source et cible.
-n : le degré de similitude dans l'intervalle [0,1].
-Rel : la relation entre les entités mises en correspondance.

Dans le but de rapprocher les entités des ontologies, des techniques d'alignement ou de similarité ont été proposées dans la littérature. Les techniques d'alignement peuvent être catégorisées comme suit :

-**Techniques syntaxiques** : Ces techniques consistent à comparer les chaines de caractères des noms/ labels des entités ontologiques (telles que : la distance de Levenshtein [Levenshtein, 1966], mesure de Lin [Chin-Yew, et al., 2000], ou autres.).

-**Techniques structurelles** : Ces techniques consistent à comparer les définitions des entités par rapport à leurs hiérarchies. On distingue les techniques structurelles internes et les techniques structurelles externes (telle que : la distance de Wu& Palmer [Wu, et al., 1994] et la distance de généralisations/spécialisalisations de concepts). Les techniques structurelles internes comparent les propriétés des entités (telles que les cardinalités d'attributs, domaine, portée, etc.). Les techniques structurelles externes comparent le positionnement des entités dans les taxonomies en calculant la distance qui les sépare.

-T**echniques sémantiques** : Ces techniques consistent à comparer les interprétations sémantiques des entités en se basant sur des ressources externes telles que : les dictionnaires, les thesaurus, etc. Cette technique est très utile lorsque les couples des entités à comparer sont synonymes.

4.2 Méthodes d'alignement d'ontologies

Dans la littérature, plusieurs méthodes d'alignement d'ontologies ont été proposées. LeTableau I.1 illustre quelques méthodes d'alignement d'ontologies proposées dans la littérature.

Certaines méthodes réalisent de bonnes performances dans certains cas et moins bonnes dans d'autres [Hamdi, 2012]. Ceci est dû à plusieurs facteurs agissant sur la qualité des alignements produits à savoir : la stratégie d'alignement adoptée, le choix des techniques, la nature des ontologies à aligner, le volume de l'ontologie, et la complexité des connaissances modélisées. En effet, lorsqu'il s'agit d'ontologies d'un domaine réel tels que : l'agriculture, la médecine, etc. la tâche de mise en correspondance des ontologies s'avère compliquée et inefficace, où le nombre d'opérations requis pour l'alignement $O(m \cdot n)$ avec $m = |O_1|$ et $n = |O_2|$. Le nombre de concepts dans chaque ontologie doit être comparé à des concepts dans l'autre ontologie en utilisant des techniques de similarité renvoyant chacune un degré de similarité. Ces degrés sont ensuite combinés afin de générer une seule valeur indiquant les couples de concepts du similaires.

Dans le but de diminuer la complexité du problème d'alignement, la limitation des tailles des ontologies est nécessaire. Dans ce contexte, la notion de clustering a récemment été introduite dans le but de décomposer en des sous parties autonomes des entités ontologiques [Ben Abbes, 2013]. Plusieurs critères ont été définis, pour le partitionnement d'ontologies, à savoir, la taille des modules/clusters générés, la distance séparant les clusters, et/ou la distance séparant les entités au sein du cluster [Ben Abbès, et al., 2012].

Ceci permet, ainsi, de transformer le problème d'alignement d'ontologies entières en des blocs d'entités ontologiques de tailles réduites. Les processus d'alignement basés sur la modularité des ontologies se caractérisent par une étape supplémentaire au début du processus qui permet de partitionner ou clusteriser les entités d'intérêt.

	Mesure de Similarité	Support de connaissances utilisé	Techniques	Automatisation
Méthode basée sur WordNet [Kong, et al., 2005]	-Similarité structurelle externe -Similarité Syntaxique	WordNet	-Les fonctionnalités du WordNet : SynSet, Les champs PTR	automatique
H-Match [Castano, et al., 2006]	-similarité linguistique -similarité structurelle	Thesaurus	- Trois niveaux d'alignement : Linguistique Linguistique et structurelle Linguistique et structurelle	-automatique
S-Match [Euzenat, et al., 2008]	-syntaxique -sémantique	WordNEt	-Exécution séquentielle des techniques de similarité	-automatique
Méthode basée sur l'ACF [Guan-yu, et al., 2010]	-similarité syntaxique -similarité linguistique	WordNet	-Treillis de Galois -Jena : outil d'analyse d'ontologie	Semi-automatique
H-Cone [Kotis, et al., 2006]	-similarité syntaxique -similarité sémantique	WordNEt	-Morphisme sémantique -LSI /DL	Semi-automatique
Méthode basée sur la classification [Maiz, et al., 2010]	-similarité sémantique -similarité terminologique	aucun	-inférence logique	Semi-automatique
Méthode d'enrichissement sémantique [Fareh, et al., 2013].	-similarité structurelle - similarité terminologique -similarité sémantique	WordNEt	-Thesaurus	automatique
Prompt [Noy, et al., 2000]	-similarité syntaxique -similarité structurelle	aucun	- Exécution séquentielle (syntaxique puis structurelle)	Semi-automatique
OLA [GiunchiGlia, et al., 2006]	-similarité sémantique -similarité structurelle interne -similarité structurelle externe		-Agrégation des similarités	automatique

Tableau I.1: Quelques méthodes d'alignement proposées dans la littérature

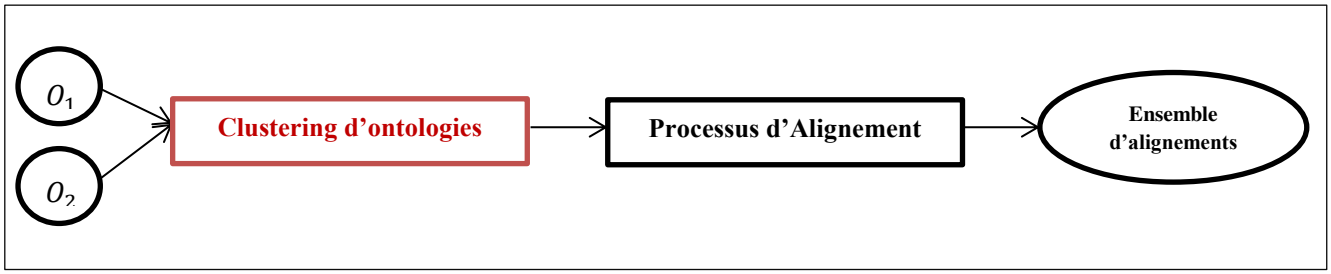

Figure I.7: Processus d'alignement basé sur le clustering d'ontologies

Dans ce qui suit, nous présentons une revue des méthodes basées sur l'emploi de techniques de clustering (supervisé ou non supervisé) dans l'objectif d'améliorer le processus d'alignement d'ontologies.

4.3 Méthodes d'alignement d'ontologies basées sur le clustering d'ontologies

Nous présentons dans ce qui suit les méthodes d'alignement les plus connues dans la littérature basées sur le clustering d'ontologies [Otero-Cerdeira, et al., 2015] telles que : Falcon-AO, TaxoMap, COMA++, etc.

4.3.1 Falcon-AO

Une méthode basée sur le clustering d'ontologies, appelée Falcon-AO a été proposée dans [Hu, et al., 2006]. La première étape du processus se base sur le partitionnement d'ontologie en un ensemble de blocs (Figure I.8) en utilisant la notion des liens pondérés à l'aide de deux techniques de similarité : la technique syntaxique (basée sur la comparaison des labels des éléments des ontologies) et la technique structurelle (basée sur la comparaison de similarités structurelles qui exploite les alignements générés par la technique syntaxique et/ou fournis en entrée de la technique). Le calcul des liens pondérés $link(c_i; c_j)$ entre deux concepts c_i, c_j s'opère comme suit :

$$\boldsymbol{poids}(\boldsymbol{c_i};\ \boldsymbol{c_j}) = \begin{cases} \text{ʃ} = \alpha * \boldsymbol{sim_{struct}}(\boldsymbol{c_i}, \boldsymbol{c_j}) + (\mathbf{1} - \alpha)\boldsymbol{sim_{syn}}(\boldsymbol{c_i}, \boldsymbol{c_j}) \ \boldsymbol{si}\ \text{ʃ} > \varepsilon \\ \mathbf{0} \end{cases} \quad \textbf{(I.1)}$$

où : $sim_{struct}(c_i, c_j)$ est la similarité structurelle entre $c_i;\ c_j$ basée sur la mesure de Wu et Palmer [Wu, et al., 1994]; sim_{syn} est la similarité syntaxique ; ε et α sont deux valeurs fixées par l'utilisateur pour la variation de poids attribués aux liens de pondération.

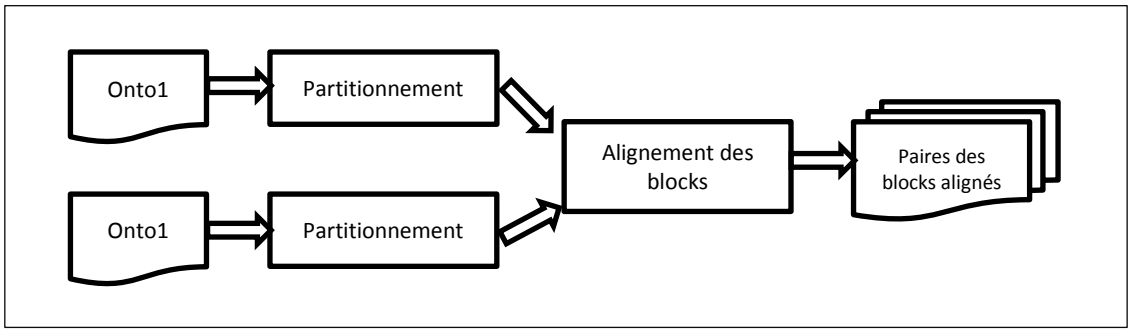

Figure I.8: Architecture de la méthode Falcon-OA

L'algorithme commence avec n blocs comportant chacun un seul concept, ensuite une méthode itérative est appliquée. A chaque itération, l'algorithme choisit le bloc qui a la cohésion maximale (la cohésion désigne la somme des poids des liens reliant les concepts du même bloc) et le bloc qui a la valeur de couplage maximale (Le couplage désigne la somme des poids des liens reliant les concepts appartenant à deux blocs distincts) avec ce premier bloc.

$$Cohésion(B_i) = goodness(B_i, B_i)$$

$$Couplage\ (B_i, B_j)\ ;\ B_i \neq B_j$$

$$\boldsymbol{goodness(B_i, B_j) = \frac{\sum_{c_i \epsilon B_i,\, c_j \epsilon B_j} poids(c_i, c_j)}{Taille\ (B_i) * Taille(B_j)}} \qquad \textbf{(I.2)}$$

L'algorithme fusionne ces deux blocs, les remplace par le cluster résultat et met à jour les nouvelles valeurs de couplage et de cohésion. Le critère d'arrêt est satisfait, lorsque les blocs construits ont atteint une taille maximale fixée au préalable par l'utilisateur.

Dans la phase d'alignement, on sélectionne les paires des blocs contenant plus d'entités en commun, la valeur de proximité entre deux blocs doit être supérieure à un seuil. La similarité des entités est calculée en utilisant une technique terminologique. La dernière étape consiste à aligner les entités des clusters similaires ou proches en utilisant les techniques structurelles et linguistiques.

Cette méthode réussit à clusteriser les ontologies de grande taille. Cependant la phase de traitement des blocs similaires est très couteuse en terme de temps vu que le calcul de proximité des blocs qui est basé sur les éléments partagés par les deux ontologies, doit être procédé sur toutes les paires de blocs possibles des deux ontologies.

4.3.2 TaxoMap

Deux méthodes ont été proposées dans [Hamdi, et al., 2008]. La première méthode commence par partitionner l'ontologie cible en utilisant le même algorithme de partitionnement adopté par Falcon-OA. Par conséquent, le partitionnement de l'ontologie source est réalisé en suivant celui de l'ontologie cible. En effet, la méthode détermine pour chaque bloc dans l'ontologie cible, l'ensemble des ancres (les entités syntaxiquement similaires partagées par les deux ontologies) qui constitueront le futur cluster dans l'ontologie source. Finalement, les entités des clusters similaires sont mises en correspondance.

La deuxième méthode se distingue par le co-clustering des ontologies, c.à.d. le clustering des deux ontologies en même temps. Pour simuler le parallélisme, l'ontologie cible est partitionnée en favorisant la fusion des blocs partageant des ancres avec la source, et on partitionne la source en favorisant la fusion des blocs partageant des ancres avec un même bloc généré pour la cible. L'algorithme de partitionnement est celui proposé par Falcon-OA.

4.3.3 COMA++

Un système dédié à l'alignement des graphes de schémas larges a été proposé dans [Massmann, et al., 2011]. La méthode proposée se base sur l'idée 'diviser pour régner'. Une ontologie est décomposée en des clusters (appelés fragments) qui seront, par la suite, mis en correspondance. Le clustering de l'ontologie est basé sur un ensemble de règles heuristiques prédéfinies pour la production d'un ensemble de clusters de taille réduite. Chaque cluster est identifié par un concept racine qui sera utilisé pour la comparaison des clusters dans la phase d'alignement de clusters. En effet, deux clusters dont les deux concepts racine sont sémantiquement similaires, ceci implique la similarité des deux clusters. L'alignement sera ensuite mené sur les entités appartenant aux clusters jugés similaires en utilisant la technique structurelle et terminologique.

4.3.4 Autres méthodes

Dans [Kachroudi, et al., 2013], les auteurs proposent également une méthode initiée par les concepts terminologiquement équivalents dans les deux ontologies à aligner. Le concept initié dans l'ontologie cible est considéré comme étant le centre d'un nouveau block. Ainsi, les concepts sémantiquement et structurellement similaires au centre seront rajoutés dans le cluster. A cette fin, les auteurs proposent d'utiliser la ressource externe WordNet [Miller, 1995] pour examiner le voisinage de chaque concept. Pour chaque concept, on identifie la liste des termes qui lui sont sémantiquement équivalents à partir de WordNet. De ce fait, chaque concept similaire sera assigné au cluster d'intérêt. Ceci permet d'assurer la cohésion des clusters produits. Le même traitement sera répété pour chaque concept rajouté au cluster. Le processus s'arrête lorsqu'il n'y aura aucun lien sémantique entre les concepts de cluster et les concepts restants ou lorsque la taille du cluster atteint une valeur maximale. Une fois que le partitionnement de l'ontologie cible est achevé, on crée le même nombre de clusters dans l'ontologie source. La phase d'alignement finale consiste à aligner les entités des clusters correspondants aux ancres (les concepts des ontologies source et cible).

Dans [Algergawy, et al., 2014], le système d'alignement proposé est initié par la phase de clustering d'ontologies où chaque ontologie est clusterisée indépendamment en un ensemble des clusters disjoints. La tâche de clustering se base sur une mesure de similarité structurelle afin d'identifier les éléments structurellement similaires. Chaque concept sera assigné à un cluster, ensuite un algorithme itératif sera déclenché. Ce dernier consiste à fusionner les clusters structurellemnt similaires. Le processus s'arrête lorsque la taille du cluster atteint une valeur maximale. La deuxième phase, consiste à identifier les clusters source et cible similaires. A cette fin, chaque cluster est converti en un vecteur de termes, la similarité est ainsi calculée en utilisant les techniques SVM et TF/IDF. Une fois les clusters similaires sont identifiés, on procède à l'alignement des éléments correspondants.

4.3.5 Motivations

Le Tableau I.2 , révèle une étude comparative des méthodes d'alignement basées sur la modularisation des ontologies introduites ci-haut. Ces travaux ont été adressés dans le but de décomposer les entités ontologiques en des modules autonomes. Bien que les approches proposées consistent à clusteriser les ontologies en vue de réduire le temps d'exécution du processus d'alignement à travers la réduction des espaces de recherche des entités correspondantes, elles souffrent de quelques limitations qui peuvent être résumées comme suit:

- Pour l'initiation de la tâche de clustering, la plupart des méthodes proposées parcourent les entités des ontologies (source et cible) pour la détermination des ancres (les entités syntaxiquement partagées). Ceci requiert énormément de temps surtout lorsqu'il s'agit de manipuler des ontologies assez volumineuses.

- Dans la phase d'alignement des clusters (l'identification des clusters similaires ou sémantiquement proches), la plupart des méthodes comparent les ensembles des blocs (provenant des ontologies source et cible), ce qui nécessite également un temps d'exécution important.

- Les techniques de clusters adoptés dans la plupart des méthodes proposées conduisent à un très grand nombre ou très petit nombre de clusters, or les clusters de très grand nombre d'entités peuvent révéler des problèmes des recherches de correspondances, et les clusters de tailles très réduites peuvent induire à une perte d'information sémantique.

- Tous les clusters produits par les méthodes citées ci-haut sont 'crisp' ou 'rigide', or les concepts déclarés dans l'ontologie disposent de plusieurs propriétés et sont multi-attributs, de ce fait, un concept peut appartenir sémantiquement à plusieurs clusters.

Méthode d'alignement basé sur le clusetring d'ontologies	**Entrée du système**	**Nature de Clustering**	**Technique de Clustering**	**Mesures de similarités utilisées**	**Utilisation d'autres techniques**	**Relations déduites**	**Limitations**
Falcon-AO [Hu, et al., 2006]	-Graphes RDF	-Agglomératif	-Clustering basé sur les liens pondérés.	-Syntaxique -Structurelle		≡	-Limitée à un langage ontologique spécifique. -Parcours de tous les clusters pour l'identification des clusters similaires.
TaxoMap [Hamdi, 2012]	- OWL -Taxonomie	-Agglomératif	-la méthode de clustering proposé dans Falcon.	-Syntaxique -Structurelle		≡ ⊆	-Dépend la qualité structurelle de l'ontologie
COMA++[Massmann, et al., 2011]	-XSD -Schéma graphe	-Utilisation des règles heuristiques	-Clustering basé sur la structure.	-Syntaxique -Structurelle (descendants, ascendants)	-Règles heuristiques	≡	-Utilisation des informations limitées pour la détermination des clusters limités.
[Kachroudi, et al., 2013]	-Réseau sémantique	-Agglomératif	-Clustering basé sur WordNet	-Sémantique -Syntaxique	-WordNEt	≡	-Evaluation entière des ontologies d'entrée la détermination des ancres.
[Algergawy, et al., 2014]	-Schéma graphe	-Agglomératif	-Clustering basé sur la structure.	-Structurelle -Syntaxique	-SVM -TF/IDF	≡	-Parcours de tous les éléments ontologiques pour la détermination des clusters similaires.

Tableau I.2: Tableau comparatif des méthodes d'alignement basées sur le clustering d'ontologies.

4.4 Les ontologies du domaine mammographique

Une ontologie mammographique est une ontologie qui prend en compte et gère toute information liée à la mammographie [Taylor, et al., 2012]. Ces informations portent sur les tumeurs, les formes d'anomalies, les diagnostics, les évaluations, les facteurs de risques, etc. Dans ce qui suit nous rappelons les principales ontologies mammographiques proposées dans la littérature.

4.4.1 L'ontologie 'MammOnto'

Cette ontologie a été conçue en 2003 [Bo, et al., 2003] afin de fournir un vocabulaire communément admis et des définitions formelles qui peuvent être utilisées afin décrire les images des seins radiologiques, les résultats anormaux des examens et les évaluations médicales pour faciliter le partage des connaissances et par la suite la réutilisation. L'ontologie a été développée en utilisant BI-RADS qui est un lexique dédié à la description des constations radiologiques contenant les descripteurs d'images (la forme des lésions, la texture), les types de lésions (calcification, masse), les types du cancer du sein (carcinome canalaire in situ) et les stades du cancer (stade I). Le langage utilisé pour le développement est DAML+OIL qui est un ancien langage par rapport aux langages courants (OWL, OWL-DL, OWL1, OWL2) pour le développement des ontologies, et dont les capacités de la représentation: Logique de Description (DL) de cette ontologie sont très limitées. Cette ontologie n'est pas accessible pour les travaux de recherche donc ses métriques sont inconnues [Idoudi, et al., 2014].

4.4.2 L'ontologie 'Breast Cancer Grading Ontology (BCGO) '

Cette ontologie [Adina, 2010] a été proposée par le projet 'European Virtual Physiological Human' en 2012 et développée avec le langage OWL-DL. Initialement, elle contenait 129 classes, 19 sous-classes et 68 propriétés.

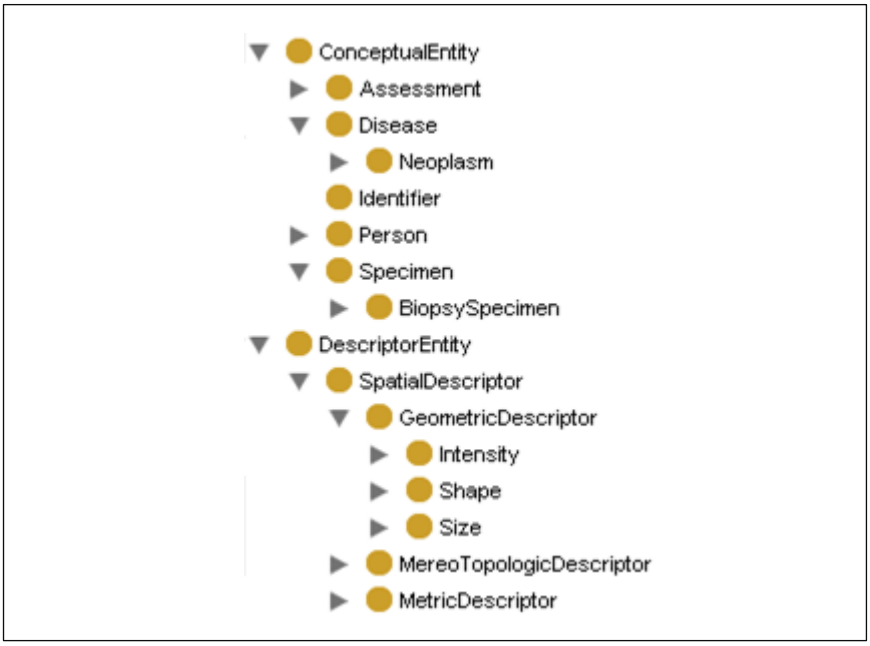

Figure I.9: Extrait de l'ontologie BCGO

Ensuite, elle a été étendue dans [Idoudi, et al., 2014] pour comporter 531 concepts et 100 propriétés. Le système de vision sur lequel se base cette ontologie est conçu pour améliorer les performances globales du processus de classement des mammographies. Cette ontologie (Figure I.9) est également associée à des règles écrites en langage SWRL pour la partie raisonnement.

4.4.3 L'ontologie 'GIMI Mammography Ontology'

Cette ontologie a été développée en 2012 [Taylor, et al., 2012], elle a été utilisée pour décrire la richesse et la complexité du domaine et a été mis en œuvre avec OWL 2, le plus récent des langages de développement d'ontologies, avec l'outil de développement *protégé_4*. Le but de cette ontologie est de l'intégrer dans un outil d'apprentissage pour comparer les annotations des stagiaires avec celles des experts.

Elle est constituée principalement de deux ontologies : « Core Mammographic Ontology » et « Mammography Learning Ontology ».

4.4.4 L'ontologie Core Mammographic Ontology (mammo.owl)

L'ontologie *mammo.owl* contient les classes pertinentes liées à la mammographie et ses composantes. Elle contient 692 classes et 135 propriétés. La structure hiérarchique conceptuelle de cette ontologie s'articule principalement autour des grandes familles: les entités anatomiques, les entités conceptuelles, les anomalies, les recommandations, et les diagnostics (Figure I.10).

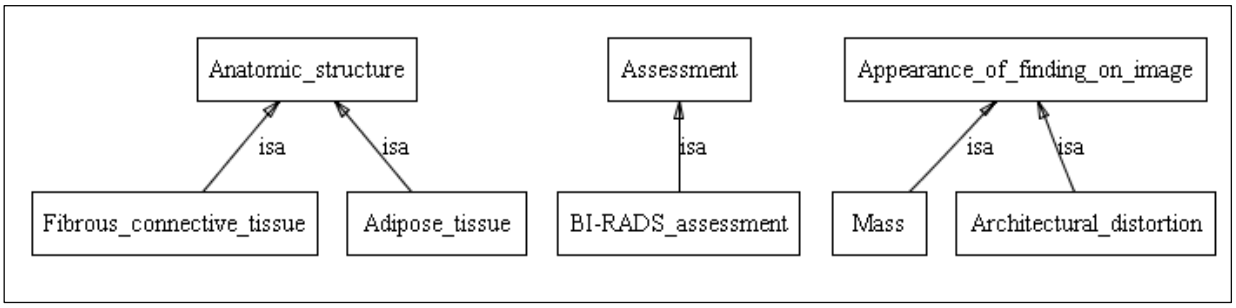

Figure I.10: Extrait de l'ontologie mammo.owl

Le processus de construction a été manuel suite à plusieurs interviews avec les experts.

4.4.5 L'ontologie 'Mammography Learning Ontology '

Cette ontologie définit deux hiérarchies de classes: La première désigne les classes de correspondance entre les deux ensembles d'annotations. Une correspondance d'apparence annotée est définie comme une paire d'annotations (relatives à l'expert et au stagiaire). L'apparence annotée correspond à une constation sur la mammographie. La deuxième classe représente les concepts qui sont utilisés dans l'enseignement. Ils décrivent les diverses situations caractéristiques. Cette ontologie comporte 740 concepts et 142 propriétés.

Figure I.11 : Extrait de l'ontologie mammo_learning

4.4.6 L'ontologie 'Mammographic Ontology'

Cette ontologie a été développée dans le cadre du projet intitulé 'Extraction et annotation des mammographies digitales basées sur les ontologies' supportées par la fondation de science nationale turque qui vise à associer des concepts de haut-niveau et de la sémantique à des données d'image médicale et exploiter ces médias sémantiques pour la recherche d'information et l'extraction.

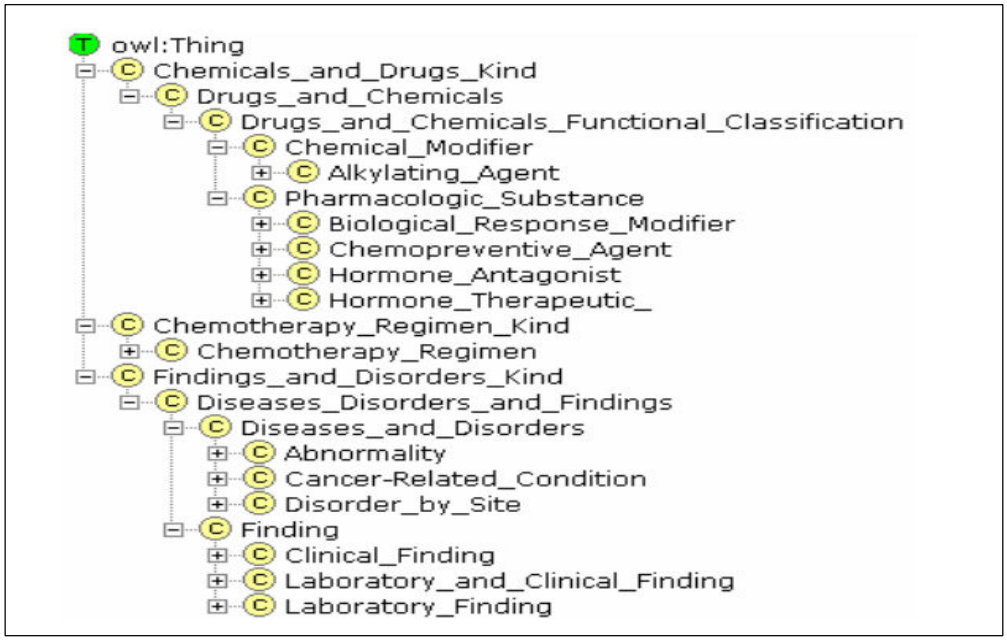

Figure I.12: Vue hiérarchique des concepts de l'ontologie 'Mammographic Ontology'

Ainsi, une ontologie a été développée pour combler le fossé sémantique entre les fonctions de bas niveau extraites par des méthodes de traitement d'image et des concepts de haut niveau Figure I.12. Elle comporte 48 top-classes et a été développée en OWL-DL avec une méthode itérative.

4.4.7 L'ontologie 'Mammography Annotation Ontology (MAO)'

Cette ontologie a été développée afin d'être intégrée dans un système d'annotation d'anomalies observées dans une mammographie [Bulu, et al., 2012]. Elle fournit un vocabulaire commun et des connaissances pour rendre les annotations compréhensibles et calculables par l'ordinateur. Elle permet au système de récupérer des ressources pertinentes, par extraction de connaissances implicites à partir de données explicites (Figure I.13). Elle définit également les concepts et les relations qu'ils entretiennent entre eux. L'application basée sur cette ontologie, permet à l'utilisateur de récupérer les cas similaires pertinents pour une requête particulière sur la base de différentes caractéristiques. Cette ontologie n'est pas accessible, ses métriques sont donc inconnues [Bulu, et al., 2012].

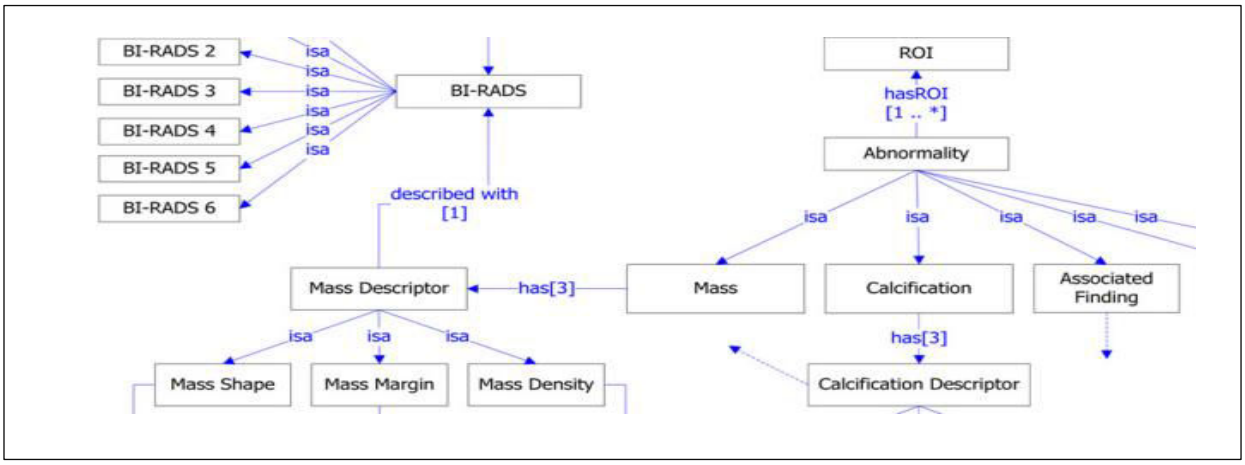

Figure I.13: Vue extraite de l'ontologie MAO

4.4.8 L'ontologie 'Breast Cancer Ontology'

Cette ontologie a été développée afin d'être intégrée dans une application destinée aux usagers de santé [Messai, 2009]. La construction de cette ontologie est basée sur des corpus textuels liés au domaine. Elle a été utilisée dans la reformulation de requêtes des usagers de santé (à savoir les patients atteints du cancer du sein et leurs entourages) (Figure I.14). Des poids numériques différents sont assignés aux relations entre les concepts selon leur internement dans le corpus. Ces valeurs ont été ensuite utilisées dans l'algorithme de propagation d'activation pour le calcul des concepts les plus caractéristiques d'une requête. Les métriques de cette ontologie ne sont pas disponibles [Messai, 2009].

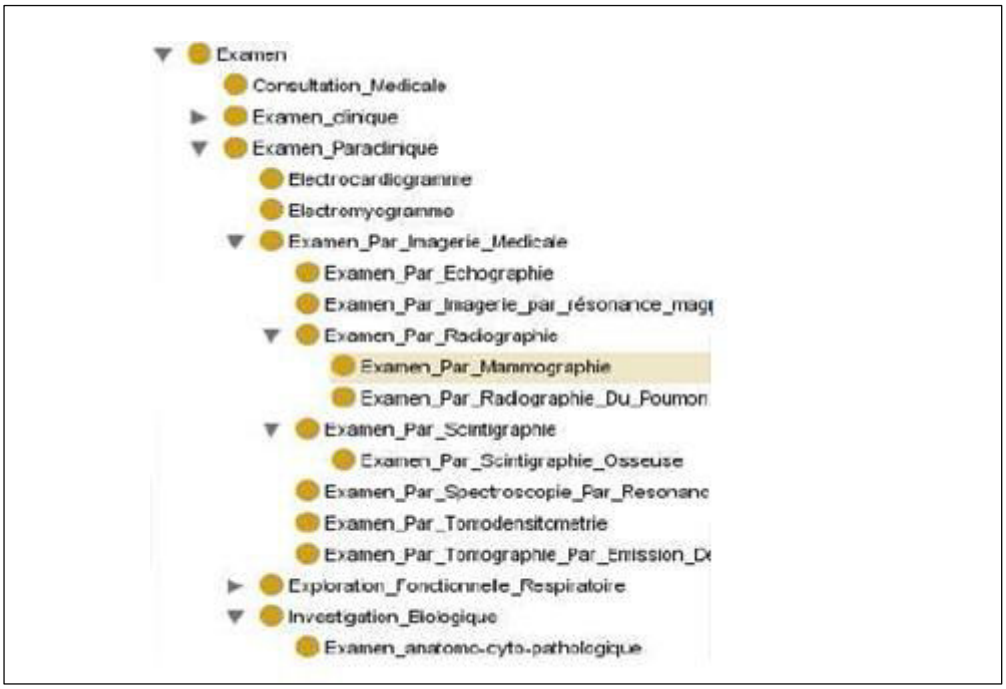

Figure I.14: Extrait de l'ontologie 'Breast Cancer Ontology'

4.4.9 Motivations

En analysant les ontologies mammographiques existantes (Tableau I.3), nous faisons ressortir les remarques suivantes :

- Le domaine mamographique est caractérisé par un ensemble de classes qui sont fortement liées et indispensables à la modélisation des connaissances mammographiques. Toutefois, les ontologies proposées dans la littérature mettent l'accent sur des classes spécifiques distinctes du domaine.
- L'ontologie *MAMMO* modélise un très grand nombre de concepts liés aux lésions décrites en terme de forme, taille, densité, etc. En outre, certaines classes évoquées dans cette ontologie sont d'aspect académique et n'interviennent pas dans la prise de décision. Ceci est dû à la nature de l'application dans laquelle cette ontologie est intégrée (apprentissage des stagiaires).
- En vue de sa conception (intégration dans un système de classification de mammographie), l'ontologie *BCGO*, modélise un très grand nombre de concepts liés à l'évaluation et la classification des mammographies (présence ou pas de cancer, stade d'évolution).
- En raison de sa conception (intégration dans un système d'annotation des anomalies dans une mammographie), l'ontologie *MAO* modélise une couverture des descripteurs morphologiques des observations radiologiques.
- L'ontologie Breast-Cancer Ontology (dont le but de conception est l'intégration dans une application destinée aux usagers non professionnels), met l'accent sur les concepts de nature informative permettant au public de formuler leurs problèmes de santé ainsi qu'une meilleure compréhension des termes techniques. Une limitation de cette

ontologie réside dans le fait qu'elle ne définit pas les éléments pertinents au domaine de la mammographie.

- L'ontologie MammOnto modélise un très grand nombre de concepts liés aux facteurs de risques.
- La majorité des ontologies mammographiques proposées sont pauvres en termes d'axiomes et de relations modélisant les corrélations a priori du domaine. En effet, la modélisation des associations entre les concepts pertinents du domaine est nécessaire afin de fournir une meilleure interprétation sémantique du domaine mammographique.

Ontologie	Métriques	But de conception	Langage de développement	Caractéristiques
MammOnto[Bo, et al., 2003]	-non disponible	- fournir un vocabulaire communément admis et des définitions formelles afin de décrire les mammographies.	- DAML+OIL	- Définition des résultats anormaux des examens et des évaluations médicales. -Utilisation du lexique Bi-Rads.
BCGO [Adina, 2010]	-531 concepts -100 propriétés	- Intégration dans un système de classification des mammographies.	- OWL-DL	-Elle est associée à des règles écrites en langage SWRL pour le raisonnement inférentiel. -Identification des concepts liés à l'évaluation et la catégorisation des mammographies (ACR).
GIMI Mammography Ontology'[Taylor, et al., 2012]	-Core Mammographic Ontology -Mammography Learning Ontology	- Intégration dans un système d'apprentissage pour comparer les annotations des stagiaires avec celles des experts	- OWL 2	-Elle liste différents concepts liés aux facteurs de risque. -Bien structurée et contient un grand nombre de concepts.
Core Mammographic Ontology'Mammo'[Taylor, et al., 2012]	-692 concepts -135 propriétés,	-Description des concepts pertinents au domaine mammographique.	- OWL 2	- Les concepts s'articulent principalement autour de ces grandes familles: les entités anatomiques, les entités conceptuelles, les anomalies, les recommandations, et les diagnostics.
Mammography e-Learning Ontology [Taylor, et al., 2012]	-740concepts -142 propriétés	-Comparaison des annotations des stagiaires par rapport à celles de l'expert.	- OWL 2	-Cette ontologie ne peut être fonctionnelle. Elle est limitée au contexte de l'application dont laquelle elle est intégrée.
Mammographic Ontology'	-48 concepts	- La recherche et l'extraction d'information	- OWL-DL	- Combler le fossé sémantique entre les fonctions de bas niveau extraites par les méthodes de traitement d'image et les concepts de haut niveau.
MAO [Bulu, et al., 2012].	-non disponible	- Intégration dans un système d'annotation des anomalies observées dans une mammographie.	- OWL-DL	-Les concepts relatifs aux anomalies évoquées dans l'ontologie sont très restreints à savoir les masses et les calcifications. -Elle met l'accent sur les descriptions des masses en termes de forme, taille, texture, etc.
Breast-Cancer Ontology [Messai, 2009].	-non disponible	-Intégration dans une application destinée aux usagers non professionnels	- OWL-DL	-Cette ontologie met l'accent sur les concepts de nature informative aidant les usagers de santé à formuler leurs problèmes.

Tableau I.3: Synthèse des ontologies mammographiques existantes

5. Conclusion

Dans ce chapitre, nous avons présenté un aperçu des différents axes de recherche sur lesquels s'appuient nos travaux de thèse. Nous avons présenté les méthodes de construction d'ontologies, particulièrement le processus d'alignement qui constitue une étape préliminaire pour les méthodes de fusion ou d'enrichissement d'ontologies. En effet, l'alignement permet de résoudre les problèmes d'interopérabilité sémantique des ressources sémantiques et il occupe une place centrale dans la gestion des ontologies. Une attention particulière, a été portée, aux méthodes d'alignement basées sur le clustering d'ontologies comme solution au problème de passage à l'échelle où nous avons dressé leurs avantages et limitations.

Nous avons, également, constaté que de nombreuses ressources ontologiques couvrant le domaine mammographique sont créés indépendamment les unes des autres. L'analyse de ces ontologies a conduit au fait que bien qu'elles représentent différentes perspectives du domaine mammographique, elles sont plus ou moins complètes. Il est illusoire de chercher à concevoir une ontologie globale qui couvrirait au mieux le domaine mammographique et serait adaptée à différentes applications. A cet intérêt, nous nous intéresserons, dans ce qui suit à adapter l'alignement et l'enrichissement au domaine visé.

II. Chapitre 2 : Processus d'Extraction des connaissances et Règles d'association

Sommaire :

1. Introduction

L'avènement des technologies d'acquisition de données dans les différents domaines (la mammographie, l'échographie, les images satellitaires, etc.) induit à un cumul des masses de données volumineuses comportant diverses expériences antérieurement menées. Ces dernières disposent d'un fort potentiel pour fournir de nouvelles connaissances. Une manière de généraliser les expériences passées est la production de règles expertes à partir des données recueillies à travers un processus d'extraction de connaissances à partir des données (ECD). Dans ce chapitre, nous abordons, dans un premier temps, la notion de fouille de données et l'extraction des connaissances à partir des données(ECD). Une attention particulière sera apportée aux règles d'association qui sert de base à nos travaux de thèse. En deuxième temps, nous dressons un panorama des différentes approches proposées, dans l'état de l'art, portant sur le post-traitement des connaissances générées.

2. Le processus d'extraction de connaissances à partir des bases de données

Au-delà des méthodes d'analyse de données traditionnelles, Les méthodes d'extraction de connaissances à partir des données (ECD) s'intéressent à l'exploitation des bases de données disponibles afin de découvrir de nouveaux motifs ou modèles utiles à la compréhension du domaine [Ruiz, et al., 2014]. Selon [Fayyad, et al., 1996] l'extraction de connaissances est définie par : « *Un processus non trivial de découverte des modèles valides, nouveaux, potentiellement utiles, compréhensibles à partir d'une base de données* ».

2.1 Les taches d'extraction des connaissances à partir des bases de données

Dans ce contexte, cette discipline a été largement étudiée en milieu médical et ses travaux touchent à de nombreux axes : l'imagerie, l'aide à la décision médicale, etc. Jusqu'aujourd'hui, plusieurs méthodes et techniques ont été proposées et appliquées selon l'objectif appelé aussi tache de fouille de données. Ces taches sont, ainsi, classées comme suit :

- **Classification** : Elle cherche à apprendre un modèle qui correspond aux données antérieurement abordées et propose des solutions pour des données présentes. Les classes se présentent en tant que champs particuliers avec des valeurs discrètes.

- **Régression** : elle permet d'analyser la relation d'une variable par rapport à une ou plusieurs autres variables.

- **Description** : Elle permet de décrire ce qui se passe dans une base de données complexe sans forcément prouver une hypothèse ou répondre à une question. Elle aide plutôt à comprendre le comportement des éléments dans la base de données et fournir une vision synthétique. Ceci est réalisé avec les outils d'analyse de données.

- **Clustering :** Elle permet d'identifier des classes ayant les mêmes caractéristiques selon la similarité de leurs attributs à partir d'un ensemble d'objets non classifiés.

- **Estimation** : Cette tâche consiste à examiner les caractéristiques d'un objet afin d'estimer la valeur d'un champ dont les valeurs sont continues. Alors que la classification traite des résultats discrets, l'estimation traite des résultats à valeurs continues.

- **Prédiction** : La prédiction est différente de la classification et l'estimation de fait que les objets sont classés en estimant la valeur future d'un champ ou en prédisant son comportement. Ceci est basé sur l'apprentissage à partir des cas antérieurs dont la variable à prédire est connue. Les données historiques, par exemple, sont utilisées pour construire des modèles pour prédire le comportement futur.

2.2 Les méthodes d'extraction des connaissances à partir des bases de données

Le choix de la méthode de fouille dépend du jeu de données, du problème à résoudre, la finalité du modèle ainsi que le domaine d'application. La littérature propose de nombreuses méthodes d'analyse qui peuvent être divisées en des méthodes classiques et complexes. Les méthodes classiques connues aussi par les méthodes de visualisation, sont utilisées pour faciliter l'interprétation des résultats avec des outils généralistes de l'informatique tels que les requêtes d'extraction dans les bases de données, les graphes de représentation, les histogrammes, l'analyse des statistiques. Quant aux méthodes complexes, elles ont pour but la réalisation des taches précises. Elles sont présentées dans ce qui suit.

2.2.1 Les réseaux de neurones

Les réseaux de neurones se présentent comme un ensemble de nœuds connectés entre eux. L'architecture d'un réseau de neurone est semblable à celle du cerveau organisée en neurones (les unités élémentaires) et synapses. Ils se présentent comme des modèles informatiques composés de trois couches : couche d'entrée, couche intermédiaire, couche de sortie (Figure II.1) dont chacune est constituée d'un groupe de nœuds de traitement disposés dans un schéma similaire à un réseau de neurones biologiques.

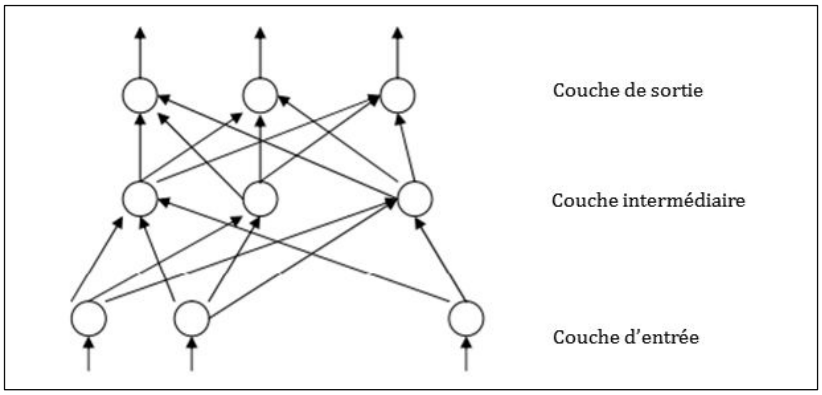

Figure II.1: Réseau de neurone

Une valeur d'activation est associée à chaque nœud et une valeur de poids est associée à chaque connexion. Une fonction d'activation régit le déclenchement des nœuds et la propagation des données via des connexions réseau. Le modèle construit permet de prédire un

résultat à partir des variables prédicatrices. La variable de résultat est discrète dans le cas d'une classification et elle est continue dans le cas d'estimation. Les réseaux de neurones peuvent être utilisés pour la classification, l'estimation, la segmentation et la prédiction.

Dans [Heydari, et al., 2012], les réseaux de neurones ont été employés pour la prédiction de l'obésité infantile afin d'éviter la maladie de l'obésité, les maladies dégénératives et chroniques.

2.2.2 Les arbres de décision

Un arbre de décision est constitué d'une racine, des nœuds qui partitionnent les individus selon leurs valeurs en deux ou plusieurs catégories et les feuilles qui désignent les classes prédéfinies. Ce modèle de prédiction se construit d'une manière récursive en divisant progressivement les données en des sous-groupes de plus en plus homogènes (voir Figure II.2). Ceci dépend du choix de la variable de partitionnement. En effet, un attribut qui se présente avec peu de valeurs constitue une bonne variable de partitionnement. Finalement, pour classifier un élément on le place dans la racine et selon la condition qu'il satisfait, il passe d'un nœud père à un nœud fils. Le nœud final constitue sa classe. Le modèle obtenu peut être aperçu comme étant un ensemble de règles.

Dans [Thangaraju, et al., 2015] les auteurs ont utilisé les arbres de décision pour prédire l'existence du cancer du foie à un stade précoce, en se basant sur une base de données contenant trois types de cancer.

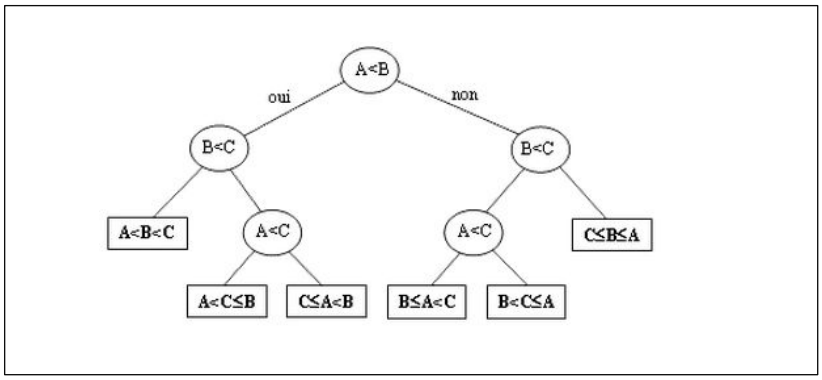

Figure II.2: Arbre de décision

2.2.3 Machines à vecteur de support (SVM)

Cette technique permet de représenter des instances, de sorte que les exemples de différentes catégories sont distants, elle consiste à découvrir s'il est possible de séparer ces instances avec un hyperplan de dimension (p - 1). Une instance est considérée comme un vecteur de dimension p. En général, il existe plusieurs hyperplans qui peuvent séparer les données ; tel que l'hyperplan de marge maximale. Dans la Figure II.3, on observe trois hyperplans utilisés.

Dans [Stoean, et al., 2013] les auteurs ont utilisé la technique des machines à vecteur de support (SVM) pour une haute précision de prédiction des systèmes d'aide à la décision médicale où le modèle d'apprentissage est effectué grâce à la technique SVM.

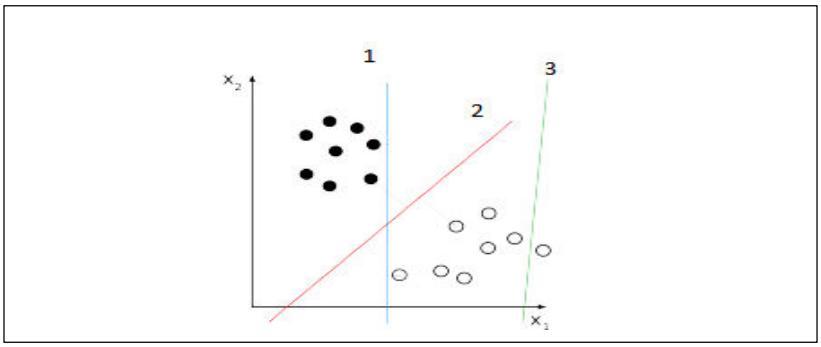

Figure II.3: Hyperplans de séparation de données

2.2.4 Les réseaux bayésiens

Le réseau bayésien se présente comme un graphe acyclique constitué de plusieurs nœuds représentant les différentes variables continues ou discrètes et des arcs constituant les dépendances probabilistiques (voir Figure II.4). La relation P(X|Y) est exprimée par un arc reliant Y à X, où X est le parent de Y. Le but de l'utilisation du réseau bayésien est de découvrir l'hypothèse la plus probable à travers le réseau de dépendance. Des connaissances *apriori* permettent d'approuver les dépendances sur les données d'apprentissage.

Dans [Baviskar, et al., 2013], les réseaux bayésiens ont été utilisés pour la construction d'un modèle de risque basé sur les données épidémiologiques, où l'objectif consiste à trouver les interdépendances entre les divers attributs de données et de déterminer la valeur seuil de dose de rayonnement pour laquelle le nombre de décès est négligeable.

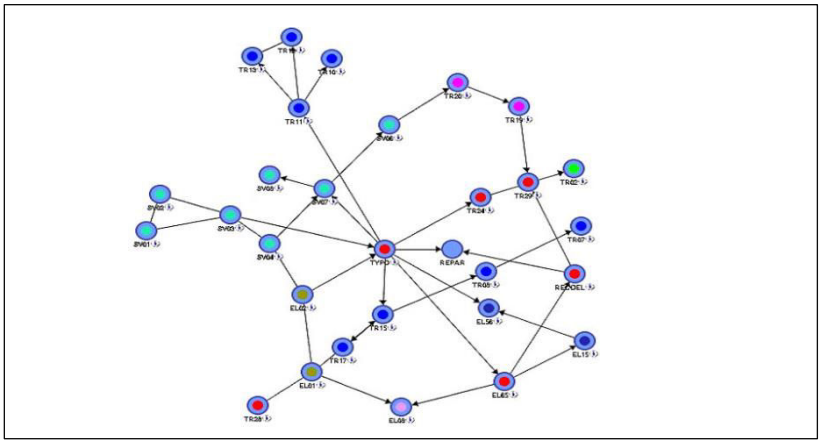

Figure II.4: Réseau bayésien

2.2.5 K-plus proches voisins

Le k-plus proches voisins est un algorithme de raisonnement à partir des cas dédié à la classification et l'estimation. L'algorithme stocke les données résolues en mémoire et récupère

les K cas les plus similaires lorsqu'une classification est requise en décidant la classe à laquelle appartient le nouveau cas. La classe est celle qui se produit le plus dans les K cas retirés. L'échantillon de test (cercle vert) dans la Figure II.5 doit être classé soit à la première classe des carrés ou à la deuxième classe de triangles). Ainsi, les éléments clés sont les suivants : (i) La fonction de distance pour mesurer la similarité entre l'instance à classifier et les plus proches voisins. (ii)Le nombre K de cas à extraire. (iii) la fonction de choix de la classe en fonction des classes des voisins les plus proches.

Dans [Raikwal, et al., 2012], les auteurs ont appliqué la technique des k-plus proches voisins sur des données médicales des patients atteints de diabète afin de classer et découvrir un modèle de données pour prédire des maladies future.

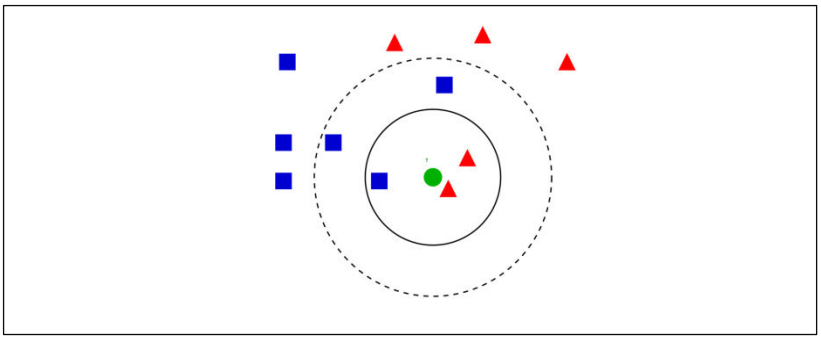

Figure II.5: Classification des K plus proches voisins

2.2.6 Les règles d'association (RA)

Cette technique non supervisée de fouille de données cherche à identifier des groupes d'items (appelés itemsets) qui se coproduisent selon certaines métriques. Le principal avantage de l'utilisation des RA, c'est qu'elles sont compréhensibles et faciles à interpréter par l'utilisateur.

Dans [Markandey, 2015], l'auteur utilise les règles d'association pour analyser une grande base de dossiers médicaux, où l'objectif était d'identifier les relations entre les procédures effectuées pour un patient et les diagnostics reportés. Elles ont été utilisées aussi pour la recherche des associations inhabituelles de médicaments et montrer comment certains symptômes se situent les uns par rapport aux autres.

2.2.7 Synthèse

Les outils et les taches décrits ci-haut font de la fouille de donnée une science riche. Ceci est dû à la complexité des données et du domaine étudié qui ne sont pas souvent standards ou faciles à manipuler. De plus, dans un seul domaine, les travaux de fouille peuvent être assez variés, traitant des problématiques particulières. On établit une étude comparative de ces méthodes dans le Tableau II.1 en considérant ces critères : l'objectif, le type d'apprentissage, les différents algorithmes d'induction, les avantages ainsi que les inconvénients.

Néanmoins, une précision des objectifs et de la finalité de l'application permet de sélectionner les outils adéquats à utiliser. Notre travail porte sur la fouille de données dans le domaine mammographique où les connaissances médicales s'accroissent de façon spectaculaire ainsi que les données qui sont de plus en plus nombreux. Ainsi, l'adaptation des techniques de fouille, à savoir les RA, dans notre contexte peut être d'un grand avantage. En effet, l'extraction de RAs concerne, principalement, un data mining descriptif. Elles sont bien connues pour aider la compréhension des nouvelles connaissances nécessaires à la pratique par l'utilisateur, puisque des relations sous la forme SI, ALORS sont souvent considérées comme proches du raisonnement humain [Koskinen, 2012].

	Objectif	Apprentissage	Algorithmes d'induction	Avantages	Inconvénient
Arbre de décision	Classification/ Prédiction	supervisé	-CART -ID3 -C4.5 -OCM -SLIQ -SPRINT	-Tolère les valeurs manquantes. -Sélection des attributs pertinents. -Interprétation faciles des résultats. -La classification des cas est efficace.	-Les classes se limitent aux nombre de classes prédéfinies. -n'est pas incrémental.
Règles d'association	Classification/ Recherche d'association	Non supervisé	-APRIORI -ECLAT -SSDM -KDCI -FP-Growth	-Elles sont compréhensibles et faciles à interpréter. -Le nombre d'itemsets n'est pas défini.	-Certaines règles peuvent être inutiles. -Nécessite un post traitement ou filtrage.
K-plus proches voisins	Classification/ Clustering	Non supervisé	K-means	-Ne nécessite pas des informations sur les données. -Peut manipuler des attributs de différentes natures. -Peut être combiné avec autres outils de fouille -Peut découvrir de structures cachées	-Les résultats sont difficiles à interpréter. -Les résultats sont sensibles au choix des paramètres initiaux. -Pour les données de type mixte, le choix de la fonction de distance est difficile.
Réseau de neurone	Estimation/ Prédiction/ Classification/ Clustering	supervisé	-AdaBoost -Learn++	-Robustesse par rapport au bruit. -Calcul rapide - Peut être combiné avec autres outils -Possibilité de traiter de descripteurs numériques, discrets ou mixtes	-Temps d'apprentissage long. - L'algorithme d'apprentissage n'est incrémental
Réseau bayésien	Classification/ Recherche d'association	supervisé		-La classification des cas est efficace. -Résultats claires.	-Les données doivent être volumineuses pour un résultat précis
SVM	Classification/ Prédiction	supervisé		-Robustesse par rapport au bruit.	-Le taux d'erreur peut être considérable

Tableau II.1: Synthèse des méthodes de fouilles de données

2.3 Les étapes de processus d'extraction de connaissances

Le processus d'extraction de connaissances à partir des bases de données (ECD) comporte trois principales étapes (Figure II.6), à savoir : (i) la préparation de données appelée aussi l'étape de prétraitement des données, (ii) la fouille des données et (iii) l'évaluation des connaissances générées appelée aussi l'étape de post-traitement.

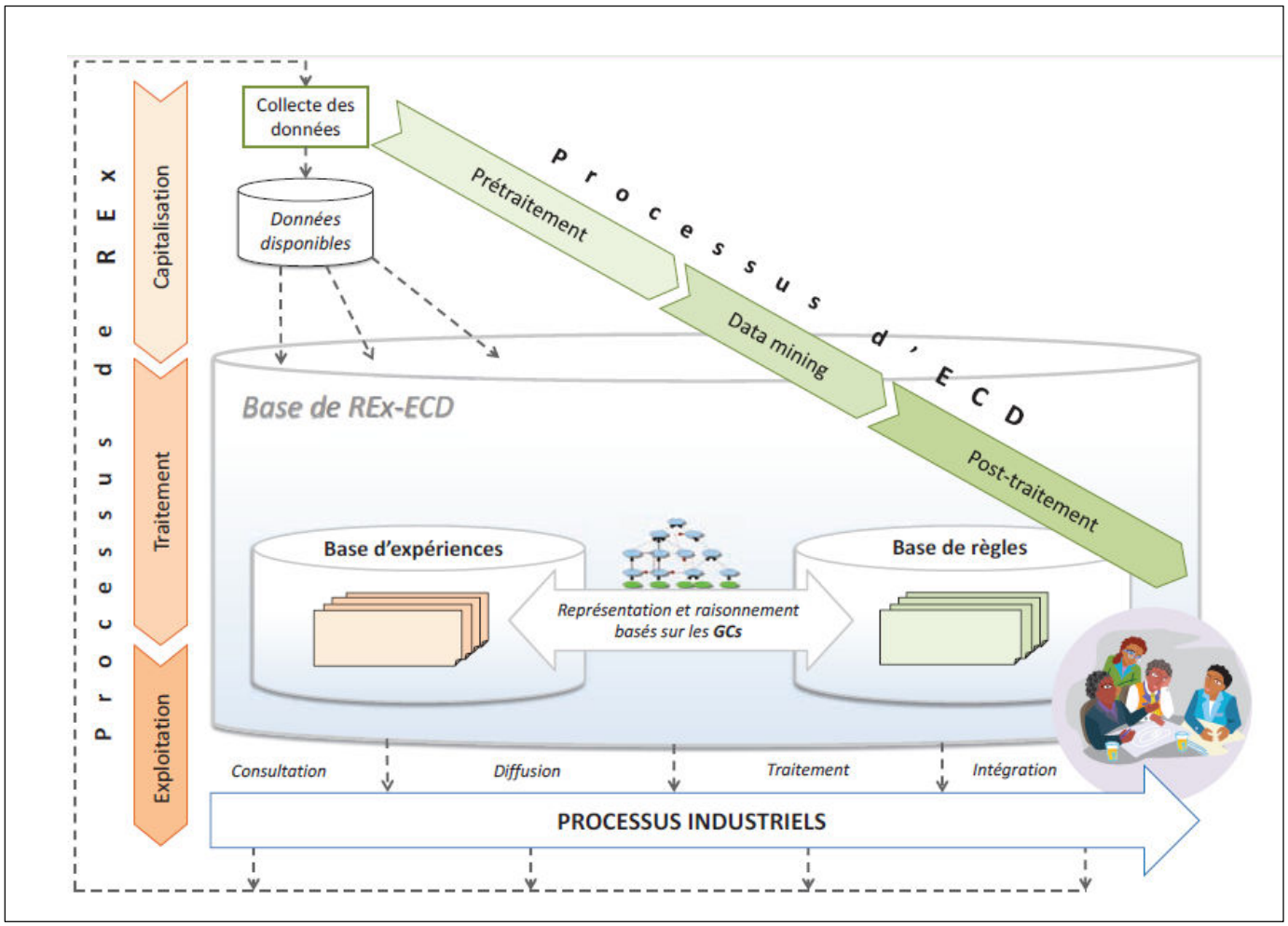

Figure II.6: Processus d'extraction de connaissances à partir de données (ECD) [Ruiz, et al., 2014]

2.3.1 Prétraitement des données

Cette étape est d'une grande importance dans l'ECD et demande une attention particulière afin de pouvoir éliminer les données inconsistantes et non fiables garantissant ainsi une meilleure qualité des connaissances résultantes [Ruiz, et al., 2014]. Cette étape est étroitement liée à la qualité, le type, le format des données, l'objectif d'extraction des connaissances, la nature du résultat attendu, etc. D'une manière générale, la préparation des données inclut quatre sous-étapes, à savoir :

- Le nettoyage des données,
- L'intégration, la réduction,
- La transformation des données.

Ces sous-étapes ne sont pas toujours ordonnées et différentiable car elles peuvent être dans certains cas, entrelacées.

Le nettoyage des données vise à corriger et/ou éliminer et/ou remplacer les erreurs et/ou incohérences et/ou absence des données en entrée en fonction du problème à résoudre. Cette étape est importante car elle permet de guider l'utilisateur dans le traitement des données pertinentes au contexte. Pour l'intégration de données, cette sous-étape est indispensable lorsque les données dont on dispose proviennent de différentes sources hétérogènes. Ainsi, l'objectif est d'intégrer les données hétérogènes dans une nouvelle base globale [Han, et al., 2006].

Pour les données contenant des attributs ou transactions inutiles et insignificatifs au problème, l'étape de nettoyage de données, consiste à réduire le volume de données et considérer seulement celles qui sont pertinentes. Cette étape permet également d'optimiser la mémoire et le temps des algorithmes de fouille de données.

Quant à l'étape de transformation de données, elle permet de transformer les données brutes (éventuellement inexploitable par la machine) en une représentation formelle. Ce format dépend de l'algorithme de fouille de données qui sera utilisé.

2.3.2 La fouille de données

Cette étape constitue le cœur du processus d'extraction des connaissances. Les techniques de fouille de données dépendent étroitement de certains critères tels que : le type de données, l'objectif d'extraction, le problème soulevé, le domaine d'étude [Ruiz, et al., 2014]. Dans le contexte médical, l'exploitation de la fouille de données dépend de la nature des données d'entrée (par exemple : image mammographique, dossier médical, etc.) et les connaissances qu'on désire avoir (par exemple : prédiction, classification, etc.).

Il s'agit de manière générale, de choisir la tache adéquate de fouille de données [Wadii, 2012] qui peut être : la classification, la prédiction, la régression, etc. (présentés dans la section 2.1), et par conséquent sélectionner les méthodes correspondantes (par exemple : les arbres de décision pour la tache de classification).

L'issue de cette étape désigne des modèles ou des relations portant de nouvelles connaissances de domaine d'étude. Le comportement des modèles générés dépend des techniques exploitées. Par exemple dans le cas de prédiction, il s'agit de déterminer des valeurs inconnues des variables à partir de la base de données en jeu [Choudhary, et al., 2009].

2.3.3 Post-traitement des connaissances générées

Cette étape est appelée aussi post-mining [Baesens, et al., 2000]. Elle inclut : la visualisation, l'évaluation et la validation des connaissances résultantes. En effet, les modèles ou relations générées ne sont pas forcément fiables ou significatifs et ne doivent pas être directement utilisés. Ainsi, l'interprétation de ces connaissances (généralement réalisée par un

expert de domaine) est cruciale afin de pouvoir les qualifier comme étant de nouvelles connaissances prêtes à être utilisées sur le terrain.

A cette fin, certains critères d'évaluation ont été définis dans la littérature. Ces critères incluent : les mesures subjectives (nécessitant l'intégration de l'expert) et objectives (basées sur les statistiques), le classement (ranking), la nouveauté, l'importance, etc.

3. Les règles d'association pour l'extraction des connaissances

Nous abordons dans cette section, le principe de la technique des Règles d'Association (RA) qui sera adopté dans ce travail de thèse pour l'extraction des connaissances dans le domaine mammographique.

3.1 Motivations

L'étude des corrélations entre les symptômes, les observations radiologiques, les facteurs de risque et les diagnostics a été toujours un défi majeur pour les médecins. En effet, dans les grands volumes d'information, on constate de plus en plus l'apparition de nouvelles formes radiologiques, de nouvelles tumeurs, de ce fait, le besoin d'exprimer et formaliser des corrélations entre ces attributs augmente de jour en jour. Ce problème a été initialement résolu par les techniques d'analyse «traditionnelles » qui ne sont plus utiles en raison de la complexité croissante des bases de données.

A l'origine, le besoin d'extraire des associations entre les attributs dans une base de données est apparu dans le contexte d'analyse des transactions de vente dans les supermarchés [Agrawal, et al., 1993]. L'objectif était d'étudier le comportement des clients dans l'achat des produits du supermarché en vue de réformer et améliorer les stratégies du marketing du magasin.

Les RAs représentent une technique de fouille de données descriptive visant à identifier des corrélations et/ou associations entre les valeurs des attributs dans un domaine donné. Cette technique se caractérise par sa simplicité de lecture et de compréhension [Koskinen, 2012]. Elles se présentent sous la forme: $Si(hypothèse)\ alors(conclusion)$.

3.2 Principe des règles d'association

Le principe d'extraction des RAs consiste à trouver des ensembles de valeurs d'attributs (définis dans la base de données) fréquents. Soit une base de données contenant n transactions (cas) $T = \{t_1, t_2, .., t_n\}$, où chaque transaction se présente avec un sous ensemble d'items de I contenant m items (désignant une valeur d'attribut) $I = \{i_1, i_2, .., i_m\}$.

Une règle d'association représente une association entre deux ensembles d'items X et Y appelés itemsets à partir d'une base de données $\{T, I\}$. La relation orientée entre X et Y est définie comme étant : $X \rightarrow Y$, avec X, $Y \epsilon\ I$; X et Y sont mutuellement exclusifs $X \cap Y = \varnothing$, ainsi X définit la partie hypothèse (appelée aussi antécédent ou prémisse) et Y définit la partie conclusion.

Sémantiquement, une RA propose une relation appropriée entre les itemsets X et Y qui a été détectée dans la base de données. Ceci dit que si les items présents dans X existent alors, il est fort probable que les items Y soient présents dans la même transaction . Autrement, la présence de X permet de conclure la coprésence de Y. Ceci est admissible conformément à deux mesures statistiques à savoir : le support et la confiance. Ces deux mesures sont étroitement liées à la base de données en jeu. Ces mesures seront définies en détail dans ce qui suit.

3.3 Mesures d'évaluation des règles d'association

Le nombre des RAs générées par les algorithmes d'extraction est étroitement lié à certaines mesures statistiques ayant pour objectif l'évaluation des règles. Nous présentons ci-dessous quelques exemples les plus utilisés à savoir : le support, la confiance, Lift, conviction.

3.3.1 Support

La mesure de support d'une RA: $X \rightarrow Y$ est une valeur numérique dans [0,1]. Elle définit la proportion des transactions comportant X et Y à la fois dans la base de données [Agrawal, et al., 1993]. Elle désigne la fréquence d'occurrence de la RA sans tenir compte de l'ordre d'apparition d'itemsets identifiés X et Y (c.à.d. $Support(X \rightarrow Y) \equiv Support(Y \rightarrow X)$). Un itemset ayant la valeur de support supérieure à un seuil fixé par l'utilisateur, dit $minSup$ est appelé itemset fréquent. Cette mesure dépend de la taille de la base de données.

3.3.2 Confiance

La mesure de confiance d'une RA: $X \rightarrow Y$ est une valeur numérique dans [0,1]. Elle définit la proportion de transactions comportant Y par rapport aux transactions totales comportant X [Agrawal, et al., 1993]. Autrement, elle désigne la probabilité conditionnelle de Y sachant X. Contrairement à la mesure de support, la confiance considère l'ordre d'apparition des itemsets dans la RA.

3.3.3 Lift

La mesure de Lift d'une RA: $X \rightarrow Y$ relie la confiance d'une RA au support de sa conclusion. Elle estime l'opportunité des transactions dans la base de données ayant X d'avoir Y. Cette mesure se base sur le calcul de l'indépendance entre les valeurs d'attributs contenus dans X et Y [Giudici, 2003].

3.3.4 Conviction

La mesure de conviction se base sur le calcul (pour chaque règle d'association RA: $X \rightarrow Y$ la déviation de la dépendance entre la probabilité d'occurrence de l'hypothèse et la probabilité de non occurrence de la conclusion dans les transactions [Brin, et al., 1997].

3.4 Algorithmes de génération des règles d'association

Parmi les algorithmes d'extraction des RAs proposés dans la littérature, l'algorithme Apriori constitue l'algorithme le plus couramment utilisé et constitue le point de départ de plusieurs algorithmes [Marinica, 2010].

3.4.1 Algorithme Apriori

L'algorithme Apriori se base sur le contrôle itératif de croissance des ensembles d'items dans la base de données. Cet algorithme admet que si un ensemble d'items est fréquent, alors ses items sont aussi fréquents [Agrawal, et al., 1993]. Généralement, le problème d'extraction des RAs à partir de la base de données se base sur deux principales étapes :

- Génération des itemsets fréquents dans la base de données,

- A partir des itemsets fréquents, les RAs dont la confiance est supérieure à un seuil donné sont générées.

3.4.1.1 Génération des itemsets fréquents

L'algorithme Apriori est un algorithme itératif (voir Figure II.7) où la découverte des itemsets les plus fréquents est réalisée graduellement, initialement, l'algorithme cherche (en utilisant la fonction *apriorigen*) les 1-itemsets (contenant un seul item) les plus fréquents, ensuite, les 2-itemsets (contenant deux items), etc. L'ensemble L_{k-1}de $(k-1)$ itemsets fréquents trouvés dans la $k-1^{ème}$ itération sont utilisés pour générer le nouvel ensemble C_k de k-itemsets candidats potentiellement fréquents.

Figure II.7: Algorithme Apriori [Agrawal, et al., 1993]

Entrées: Bases de données BD
Sorties L : Ensemble des itemsets
Debut

$L_1 = \{1-itemsets\}$
Pour $(k = 2 ; L_{k-1} \neq \varnothing; k++)$ **faire**
  C_k=apriorigen (L_{k-1})
  Pour toute transaction t ϵ BD **faire**
    C_t= subset (C_k, t)
    Pour tout candidat c dans C_t **faire**
      c.count++
  FinPour
  $L_k = \{c \epsilon C_k | \, c.count \geq minSup\}$

apriorigen (L_{k-1})
Pour tout itemset cϵ C_k **faire**
    Pour tout $(k-1)$ sous ensemble s de c **faire**
    Si ($s \epsilon L_{k-1}$)
      Supprimer c de C_k
  FinPour
FinPour

Ces candidats sont évalués lors d'un nouveau passage sur les données lorsque le support de chaque candidat est calculé. L'itemset candidat dont $(k-1)$ sous-ensembles ne figurent pas dans C_k est supprimé de C_k. Ainsi, les transactions de la base de données, sont ensuite, balayés pour calculer les supports de chaque itemset dans C_k à l'aide de la fonction $sousEnsemble$ qui reçoit en entrée une transaction et l'ensemble C_k et retourne l'ensemble des itemsets dans C_k satisfaisant la transaction en question.

Une fois les itemsets L les plus fréquents sont déterminés, il s'agit de générer les RAs dont les mesures de confiance sont supérieures à un seuil prédéfini par l'utilisateur (voir la Figure II.8). Pour chaque ensemble d'item l_k, l'algorithme trouve les sous ensemble a_{m-1} de l_k et propose un ensemble de règles sous la forme $a_{m-1} \rightarrow (l_k - a_{m-1})$ avec une valeur de confiance supérieur au minConf. L'union et l'antécédent des règles générées à partir de l_k représente l'itemset l_k. La fonction *gen-règles* rajoute les règles valides dans l'ensemble *Règles*.

3.4.1.2 . Extraction des règles d'association

Figure II.8: Génération des RAs [Agrawal, et al., 1993]

Entrées: L : Ensemble des itemsets
Sorties : Ensembles des RAs

Début
Pour tout itemset l_k ($k \geq 2$) **faire**
    gen-règles (l_k, l_k)

gen-règles (l_k, l_k) (l_k: $K-$ itemset, a_m: $m-$ itemset)
A={(m-1)itemsets a_{m-1}| $a_{m-1} \subset a_m$}
Pour tout $a_{m-1} \in$ A **faire**
  Conf=support(l_k)/support(a_{m-1})
  Si (conf $\geq$ linConf) **alors**
    gen-règles (l_k, a_{m-1})
    Règles= Règles $\cup$ R
Retourber *Règles*

4. Méthodes de post-traitement des règles d'association

Vu le grand nombre des RAs générées (dont beaucoup peuvent ne pas être intéressantes), qui ne sont pas éventuellement exploitables par l'être humain, la quatrième étape du processus d'ECD, concernant le traitement des règles, est notamment cruciale pour aider l'utilisateur à explorer les RAs les plus efficaces. Le problème est ainsi passé, de l'étape de fouille de données pour l'extraction des associations implicites à la fouille des règles pour l'extraction des connaissances pertinentes à la compréhension du domaine.

Les approches de post-traitement des RAs peuvent être résumées comme suit : La réduction de nombre des RAs produites, l'évaluation, le classement/ordonnancement et le filtrage. Les travaux de post-traitement peuvent être assez multiples, et le choix dépend des

attentes de l'utilisateur. La plupart des approches proposées se basent sur la mesure d'intérêt d'une RA encodant l'utilité ou l'importance du modèle.

Les premiers travaux pour l'évaluation des règles d'association ont été limités à l'application de deux mesures d'intérêt : le support et la confiance. Ensuite, plusieurs mesures complémentaires ont été proposées afin d'assurer la sélection des meilleures règles [Fabrice, 2010]. Le but de ces mesures est de pouvoir produire des règles utiles facilement interprétables par l'expert de domaine. Ce dernier valide ensuite les meilleurs RAs pour des éventuelles utilisations. Ces mesures peuvent être classées en des mesures objectives et mesures subjectives. Les mesures objectives dirigées par les données permettent d'évaluer principalement l'éventualité des règles, tandis que les mesures subjectives, consistent à intégrer les connaissances de l'utilisateur.

4.1 Analyse objective des règles d'association

Initialement, l'évaluation objective de la force d'une RA est réalisée au moyen du support et confiance. Le Tableau II.2 montre les différentes interprétations de la combinaison de ces deux paramètres.

Tableau II.2: Compromis entre Support et Confiance d'une règle [Hajlaoui, 2009]

	Confiance faible	***Confiance élevé***
Support élevé	La RA est rarement valide mais peut être utilisée fréquemment	La règle est souvent valide et peut être utilisée fréquemment
Support faible	La règle est rarement valide et ne peut être utilisée que rarement	La règle est souvent valide mais ne peut être utilisée que rarement

Actuellement, plusieurs méthodes d'analyse objectives ont été proposées dans la littérature. On illustre dans ce qui suit quelques exemples de ces méthodes en définissant leurs principes.

Dans [Freitas, 1998] l'auteur propose une mesure d'intérêt qui permet de quantifier la surprise d'une RA. L'auteur admet qu'une RA est considérée surprenante si sa généralisation minimale constitue une règle dont la conclusion est différente de la RA. Soit une RA: $A_1, A_2, \dots A_n \rightarrow C$; la R_i représente une généralisation minimale obtenue suite à la suppression d'un attribut A_i de l'antécédent. La mesure de la surprise de RA consiste à compter combien de fois la conclusion des $R_{i;1\leq i\leq n}$. Plus cette mesure est importante, plus la RA est surprenante pour l'utilisateur. La mesure est calculée comme suit :

$$\boldsymbol{disjSurp(RA)} = \frac{\sum_i \boldsymbol{DiffConclusion_i(RA)}}{\boldsymbol{m}} \quad \textbf{(II.1)}$$

Avec $DiffConclusion_i(RA) = \begin{cases} 1\ si\ Conclusion(R_i) \neq Conclusion(RA) \\ 0\ sinon \end{cases}$

Dans [Bottcher, et al., 2009], les auteurs proposent une nouvelle méthode pour l'analyse des RAs en exploitant deux bases de données liées au même domaine à des instants différents afin de détecter des motifs stables dans le temps. Les deux ensembles des RAs générées sont ensuite comparés afin de pouvoir suivre le changement historique. Chaque nouvelle RA est comparé par rapport à son historique en comparant le changement au niveau de support ainsi que la confiance :

$$\boldsymbol{\Delta(Support(R_i)) = Support_2(R_i) - Support_1(R_i)} \quad \textbf{(II.2)}$$

$$\boldsymbol{\Delta(Confiance((R_i))) = Confiance_2(R_i) - Confiance_1(R_i)} \quad \textbf{(II.3)}$$

Finalement, selon les résultats obtenus, les RAs sont classées en des catégories :

- Stabilité,

- Changement non rapide,

- Changement homogènes.

Dans [Mohd, et al., 2011], les auteurs présentent une approche systématique pour l'évaluation des règles découvertes. La stratégie proposée combine l'exploration de données et les techniques de mesure statistique, y compris l'analyse de redondance, l'échantillonnage et l'analyse statistique multi variée, pour ignorer les règles non significatives. L'approche proposée se base sur un processus d'échantillonnage, développement d'hypothèse, construction de modèles et enfin une mesure basée sur les techniques d'analyse statistique pour vérifier l'utilité et la qualité des règles découvertes. Cela permet de filtrer les règles redondantes, trompeuses, aléatoires et se produisant par hasard.

4.2 Analyse Subjectives des règles d'association

L'analyse subjective des RAs sollicite l'intervention de l'expert pour guider l'extraction des RAs les plus pertinentes. Plusieurs méthodes d'analyse subjective ont été proposées dans la littérature, avec l'utilisation de différents supports sémantiques pour la modélisation des connaissances. On illustre dans ce qui suit quelques exemples de ces méthodes en définissant leurs principes.

Dans [Marinica, 2010], l'auteur propose de modéliser les connaissances de l'utilisateur en utilisant le formalisme proche de celui de la RA, appelé Schéma de Règles. Ces derniers permettent de définir, les attentes de l'utilisateur concernant les RAs produites. Pour la modélisation des connaissances de domaine, l'auteur propose d'utiliser une ontologie de domaine. Ainsi, l'utilisateur peut utiliser un ensemble d'opérateurs de traitement interactif

appliqué sur les schémas de règles soit pour l'élagage, la recherche de conformité, la recherche des contradictions, etc.

Pour mettre en correspondance les connaissances de l'utilisateur et les RAs découvertes, une nouvelle technique a été proposée dans [Padmanabhan, et al., 1998], et révisée plus tard dans [Padmanabhan, et al., 2000]. Ces méthodes permettent l'extraction des RAs correspondantes aux connaissances modélisées sous la forme des règles. L'extraction des RAs inattendues est proposée dans [Natarajan, et al., 2004]. Cette méthode consiste à extraire uniquement les RAs qui contredisent logiquement le conséquent ou l'antécédent des connaissances de l'utilisateur.

Dans [Ana, et al., 2009], les auteur proposent une nouvelle approche pour l'évaluation des RAs . Cette approche est basée sur deux composantes: l'ontologie de domaine, utilisée pour le calcul de la distance sémantique entre deux items, et les préférences de l'utilisateur modélisant ses points de vue par rapport au domaine. Cette approche permet d'évaluer la pertinence des paires d'items. Ainsi, (i) la distance sémantique indiquant à quel point deux items sont sémantiquement proches, chaque type de relation étant pondéré différemment et (ii) les connaissances de l'expert représentent les deux composantes utilisées pour guider le processus durant la phase de sélection des ensembles d'items.

Dans [Ruiz, et al., 2014], la formalisation des connaissances et des expériences du domaine est assurée en utilisant un formalisme de représentation des connaissances : les graphes conceptuels. Ainsi, l'extraction de nouvelles connaissances dans les RAs, est basée sur une approche interactive innovante en utilisant les opérations de raisonnement conceptuel graphiques.

Dans [Razan, et al., 2014] les auteurs ont proposé une mesure d'intérêt basée sur une ontologie pour encoder l'utilité d'une règle afin de pouvoir sélectionner et classer les règles en fonction de leur importance. La mesure proposée s'appuie sur le calcul sémantique de la similarité entre deux items i_1, i_2 en utilisant sur la mesure de Wu&Palmer qui se base sur l'emplacement des concepts (qui leur correspondent) ainsi que leur ancêtre commun dans la hiérarchie ontologique.

Dans [Mansingh, et al., 2011], les auteurs ont proposé de créer des partitions significatives dans l'ensemble des règles d'association extraites. L'approche est basée sur la combinaison des connaissances provenant d'une ontologie avec la mesure objective de fiabilité. Les auteurs distinguent cinq partitions qui intéressent les experts du domaine à savoir : (1) Des règles d'une importance élevée qui peuvent conduire à la découverte de nouvelles croyances; (2) Des règles connues avec une importance élevée qui correspondent aux connaissances de l'expert du domaine ; (3) Des règles d' une importance moins importante qui ne sont pas soutenues par les données ; (4) Des règles manquantes qui n'ont pas été extraites par induction de règles d'association (La présence de ces règles est due aux valeurs de seuils erronées) et ; (5) Des règles contradictoires permettant l'expression de la négation.

Dans [Hamania, et al., 2014], l'auteur propose une nouvelle distance sémantique des règles d'association basée sur une ontologie floue. L'auteur évalue l'intérêt des RAs en termes

d'imprévisibilité : Les règles sont intéressantes si elles ne sont pas connues à l'utilisateur ou en contradiction avec les connaissances existantes (ou les attentes) de l'utilisateur, et *Actionabilité* : les règles sont intéressantes si l'utilisateur peut faire quelque chose avec elles à sa/son avantage. La distance proposée est basée sur le poids associé au concept et la pertinence de la relation entre un concept donné le concept le subsumant.

4.3 Synthèse

Le Tableau II.3 synthétise les travaux de post traitement, les plus citées dans la littérature, basés sur la mesure subjective des RAs. Nous pouvons constater que les supports de modélisation de connaissances de l'utilisateur sont multiples à savoir : les Schémas de Règles, les réseaux bayésiens, les graphes conceptuels, les ontologies, etc. En ce qui concerne les travaux basés sur les ontologies, ces méthodes n'exploitent pas réellement les connaissances modélisées et se contentent de mesurer la distance entre les items dans la RA. Pareillement, l'utilisateur ne peut réaliser qu'une seule action sur les modèles de connaissances (Elagage, classement, etc.), ce qui ne permet pas de valider définitivement la qualité des règles découvertes. Or, l'ontologie représente un support sémantique puissant utile pour l'exécution de plusieurs actions de post-traitement des RAs.

5. Ontologies et Règles d'association

Plusieurs approches se sont intéressées au couplage des règles d'association et ontologies. Selon la manière dont l'ontologie a été employée, ces travaux peuvent être classés en deux classes (Figure II.9):

- L'ontologie dans le processus de fouille de données,
- L'ontologie comme objectif de fouille de RA.

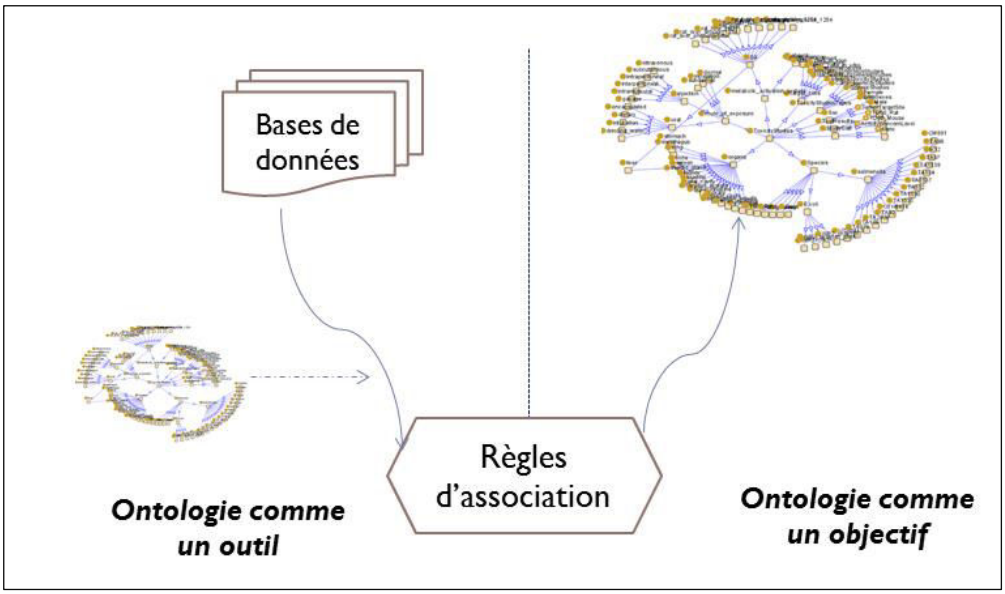

Figure II.9: Ontologie et Règles d'association

5.1 L'intégration de l'ontologie dans le processus d'extraction des règles d'association

Ces travaux visent à codifier la description du processus de fouille de données en fonction des connaissances a priori du domaine modélisées par des ontologies du domaine. Ces derniers peuvent intervenir dans plusieurs étapes du processus d'ECD telles que l'étape initiale de prétraitement des données ou l'étape finale de post traitement des connaissances extraites :

- *Utilisation de l'ontologie dans l'étape de prétraitement des données* :

L'utilisation des ontologies dans cette étape permet d'extraire les attributs les plus appropriés et générer le modèle le plus intéressant pour la connaissance découverte. Dans ce contexte, [Zeman, et al., 2009] utilisent une ontologie dans l'étape de préparation des données pour une méthode de récupération des connaissances. Cette méthode, appelée Ferda, repose sur des calculs d'observations et des statistiques. Dans cette méthode, nous identifions deux moyens d'utilisation des ontologies à savoir la construction d'une catégorisation d'attributs et l'inspection/ l'exploitation des attributs liés sémantiquement.

Dans [Ferraz, et al., 2013], les auteurs proposent un module d'enrichissement de transactions en se basant sur une ontologie de domaine (comportant les des éléments des transactions). Ayant comme entrée une base de données des transactions, ce module génère un ensemble de transaction dont les attributs sont enrichis. Ce dernier est soumis au processus d'ECD afin de générer des RAs enrichies.

Les approches proposées dans [Bellandi, et al., 2008] [Hou, et al., 2005] utilisent une ontologie pendant la phase de prétraitement où les données dans la base de données sont levées aux concepts les plus généralisés. Ceci permet de faciliter l'interprétation des règles générées du fait qu'elles contiennent des concepts de haut niveaux qui représentent des renseignements plus riches que les termes spécifiques.

- *Utilisation de l'ontologie dans l'étape de post traitement des connaissances extraites* :

L'utilisation des ontologies dans cette étape permet de découvrir et/ou valider les règles informatives et sémantiques à partir des règles extraites [Marinica, 2010], [Mansingh, et al., 2011]. La section 4.2 présente un aperçu détaillé.

5.2 L'utilisation des RAs pour l'enrichissement d'ontologie

Bien que, beaucoup de travaux [Mahmoodi, et al., 2016] [Gim, et al., 2015] [Dou, et al., 2015] [Nebot, et al., 2012] se basent sur l'utilisation des ontologies comme un guide d'extraction de règles ou de motifs, il s'avère, à l'inverse, que peu de travaux d'enrichissement d'ontologie à partir des RAs se sont dressés dans la littérature [Idoudi, et al., 2016].

Les approches d'enrichissement d'ontologies à partir des RAs proposées dans la littérature se basent sur deux étapes :

- La détermination des concepts et relations candidats,

- Le placement de ces termes au sein de l'ontologie cible.

Dans [Maedche, et al., 2000], [Bendaoud, 2006], [Stumme, et al., 2006], les auteurs ont proposé d'utiliser les corrélations fréquentes qui existent entre les termes d'un corpus. A l'issu du processus d'ECD, chaque règle exprime l'existence d'une relation entre deux concepts du domaine. Ce processus d'enrichissement reste semi-automatique car d'une part, le nombre de règles associatives découlées est très important et d'autre part, une intervention humaine est nécessaire pour labéliser sémantiquement les relations découvertes.

Dans [Di Jorio, et al., 2007], les auteurs se sont intéressés aux motifs séquentiels afin d'extraire les termes candidats à l'enrichissement, et de les corréler à la structure ontologique. La démarche proposée se compose de trois étapes : Il s'agit de fouiller, dans un premier temps, un corpus de texte afin extraire les motifs séquentiels qui comportent les items candidats à l'enrichissement. Ensuite, on procède à la recherche de l'emplacement adéquat. Ce rapprochement est réalisé en utilisant une mesure de proximité. Finalement, les nouveaux éléments sont placés au sein de l'ontologie existante.

Tableau II.3: Comparaison des méthodes de post-traitement basées sur l'analyse subjectives des RAs

Auteurs	Principe	Objectif	Support de représentation des connaissances de l'utilisateur
[Fauré, et al., 2006]	-Comparer les RAs par rapport aux modèles de connaissances dans un réseau bayésien	-Recherche des RAs surprenantes	-Réseau bayésien
[Ana, et al., 2009]	-Calcul de la distance sémantique basée sur l'ontologie -intérêt de l'utilisateur	-Evaluation des RAs	-Ontologies -Relation pondérées
[Marinica, 2010]	-Application des opérateurs pour la comparaison des schémas de règles et les RAs.	-Elagage, -Recherche de conformité -Recherche des RAs surprenantes	- Schéma de Règles -Ontologie
[Narayana, et al., 2013]	-Calculer le classement des RAs en se basant sur le contexte des items dans l'ontologie.	-Classement des RAs	-Ontologie
[Razan, et al., 2014]	-Calcul de la distance sémantique basée sur l'ontologie.	- Classement des RAs.	-Ontologie
[Ruiz, et al., 2014]	-Comparer les RAs par rapport aux connaissances modélisées dans un graphe conceptuel	-Elagage	-Graphe conceptuel

6. Conclusion

Dans ce chapitre, nous avons introduit le processus d'extraction des connaissances à partir des entrepôts de données. Nous avons porté une attention distinctive aux RAs grâce à leur intérêt particulier d'analyser les relations entre les valeurs des attributs d'une base de données. Toutefois, l'utilité des RAs est fortement limitée par la quantité énorme générée, ce qui complique la tâche d'analyse et de validation de ces connaissances. De ce fait, l'étape de post-traitement s'avère indispensable pour la réduction, le ranking, le filtrage des RAs. L'étude des méthodes de post-mining conduite a démontré deux types de post traitement, à savoir : les méthodes objectives guidées par la structure des données, et les méthodes subjectives sollicitant l'intervention des utilisateurs.

Grace aux utilités diverses des RAs, nous proposons dans ce travail de les exploiter pour l'analyse des relations entre les entités pertinentes liées au domaine de la mammographie pour une meilleure compréhension de domaine. Dans ce contexte, nous présentons une démarche originale qui prend en compte deux manières d'évaluer et d'analyser les règles d'association : une évaluation sémantique subjective basée sur un mécanisme de comparaison de ces règles par rapport aux connaissances existantes en vue de permettre à l'utilisateur de filtrer les nouvelles règles, et une évaluation objective sémantique afin de mesurer l'importance de chaque règle en vue de simplifier et faciliter à l'expert la tache de l'analyse.

III. Chapitre 3 : Vers une Approche de Fouille de Connaissances pour l'Enrichissement d'ontologies

Sommaire :

1. Introduction

Ce chapitre présente notre contribution concernant la proposition d'un processus d'enrichissement d'ontologie. Nous nous intéressons dans cette partie du travail à l'enrichissement conceptuel et relationnel de l'ontologie en exploitant de manière séparée deux courants scientifiques considérés comme étant deux sources de connaissances possibles à notre problématique, à savoir les ressources ontologiques existantes et les nouvelles connaissances extraites à partir des expériences antérieurement menées. D'une manière générale, la tâche d'enrichissement est souvent composée de trois phases : l'une consiste à identifier les éléments pertinents utiles à l'enrichissement, la fouille des résultats obtenus, et finalement leur placement au sein de l'ontologie de référence. Ces trois phases sont dressées dans les deux contextes d'enrichissement employés.
L'enrichissement conceptuel se base sur l'introduction de nouveaux concepts. Le processus proposé s'appuie sur trois étapes (1) L'étape de pré-alignement d'ontologies (2) L'étape d'alignement des ontologies pour l'identification des concepts pertinents à introduire et (3) L'étape d'enrichissement de l'ontologie cible avec placement adéquat de nouvelles connaissances extraites de l'ontologie source. Pour l'enrichissement relationnel, la méthode proposée se base sur l'extraction des associations entre les termes d'une base de données de domaine. L'approche proposée s'appuie sur trois étapes (1) L'étape d'extraction des règles d'associations à partir d'un corpus de domaine (2) Etape de filtrage et classification de ces règles afin d'évaluer leur pertinence (3) L'étape d'enrichissement de l'ontologie cible par de nouvelles associations établies entre les concepts.
Nous détaillons dans ce qui suit les différentes phases du processus d'enrichissement conceptuel et relationnel proposé.

2. Fouille de connaissances

Dans nos travaux de thèse, nous exploitons la richesse du contenu informatif des sources hétérogènes de connaissances pour une meilleure compréhension du domaine. L'originalité de notre approche s'introduit à travers la définition d'un nouveau concept : *la fouille de connaissances*. Cette notion, a été employée dans le but d'extraire de nouvelles connaissances ciblées à partir des connaissances existantes.

Partant du constat que les connaissances peuvent être considérées comme des données à structure développée, notre perspective consiste à développer des algorithmes spécifiques de fouille dans des bases de connaissances (telle que les ontologies OWL, les règles) dans le but de découvrir de nouvelles connaissances orientées. Partant de cette notion, l'approche que nous proposons se base sur l'exploitation de deux principaux axes dans le but d'extraction des connaissances utiles à l'enrichissement de notre base de connaissances cible.

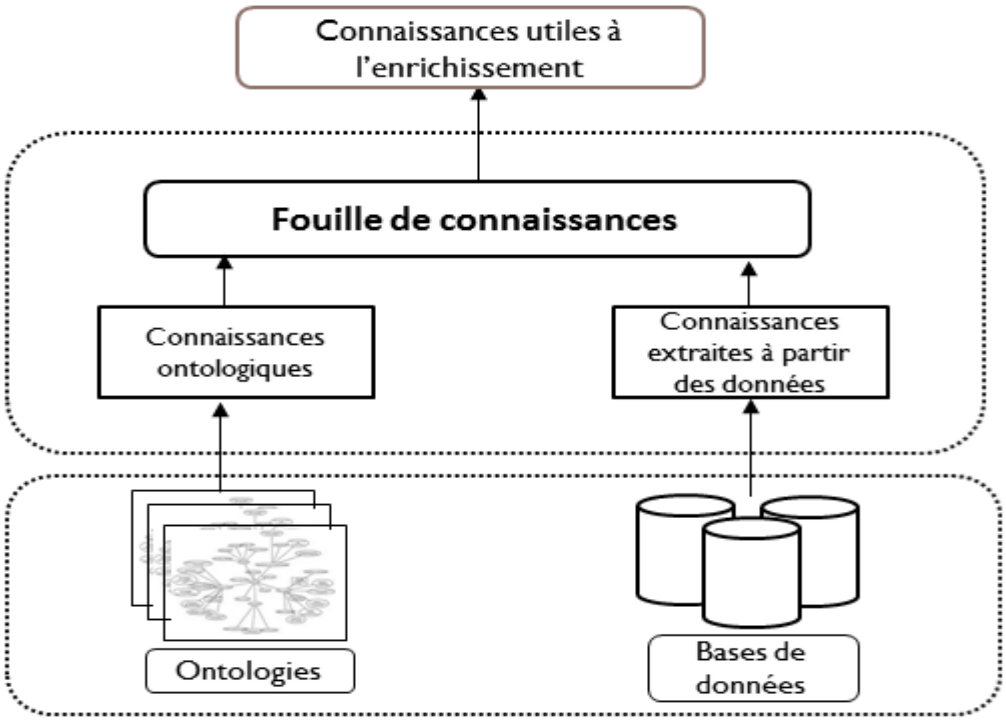

Figure III.1: La fouille de connaissances

3. Présentation de l'approche globale proposée

L'essence de notre travail s'articule autour des méthodes d'enrichissement d'une ontologie de référence permettant l'intégration de nouvelles connaissances provenant de différentes sources. Le processus d'enrichissement d'ontologies consiste à étendre une ontologie cible par l'ajout de nouveaux éléments (concepts et/ou relations) [Petasis, et al., 2011]. La contribution de ce travail s'articule autour de deux types d'enrichissement basés sur la fouille des connaissances: Enrichissement Conceptuel et Enrichissement Relationnel (voir la Figure III.2).

- *Première Contribution : La fouille de connaissances pour l'enrichissement conceptuel de l'ontologie*

La démarche proposée d'*enrichissement conceptuel* se base principalement sur l'exploitation de ressources ontologiques existantes. Dans ce contexte, notre problème consiste à délimiter les éléments de l'ontologie source utiles à l'enrichissement. De ce fait, l'alignement des ontologies en jeux (l'ontologie de référence et l'ontologie qui sert à l'enrichissement) constitue une étape nécessaire afin de mettre en correspondance ou de rapprocher les entités des différents modèles conceptuels. Néanmoins, l'étape de recherche des mappings est particulièrement laborieuse lorsqu'on se trouve face à des ontologies du monde réel, volumineuses et larges (avec une centaine de concepts). En effet, quand les ontologies sont de grande taille, l'efficacité des méthodes d'alignement diminue considérablement [Euzenat, et al., 2010].

Le principal défi à résoudre consiste à restreindre l'espace de recherche d'alignements sans perte d'information. Pour ce fait, l'approche d'alignement, que nous proposons, se caractérise par une étape de préparation d'ontologies, appelée aussi étape de pré-alignement basée sur la fouille de connaissances ontologiques.

Notre contribution consiste, à ce niveau, en un processus de fouille de connaissances ontologiques qui résulte en une nouvelle réorganisation plus flexible de concepts de l'ontologie. La solution que nous proposons, est de transformer la structure monolithique et rigide de l'ontologie en des groupements des concepts structurés hiérarchiquement. A cette fin, un algorithme de *clustering hiérarchique flou* des concepts ontologiques est mis en place.

Une telle contribution présente une vision hiérarchique des clusters conceptuels, générés à partir de l'ontologie, et regroupant les concepts les plus similaires. Ceci permet non seulement la visualisation des niveaux de granularité des connaissances entre les modèles ontologiques, mais encore de faciliter la mise en correspondance et l'alignement itératif des concepts des ontologies.

Une telle contribution permet également d'améliorer les phases de recherche et d'extraction d'information, surtout que les connaissances du domaine se développent de façon spectaculaire, en particulier, dans le domaine mammographique où les ontologies peuvent rapatrier de plus en plus de nouvelles connaissances en raison du développement des différents types de tumeurs ou des formes radiologiques.

L'utilisation de la technique de clustering a pour but de transformer le problème d'alignement de deux ontologies entières en alignement 'light-weight' des groupements de concepts issus chacun d'une des deux ontologies d'intérêt. Ceci permet de réduire considérablement l'espace de recherche des correspondances. Dans ce contexte, nous proposons une nouvelle distance sémantique qui servira de base pour l'algorithme de clustering. Cette notion de 'proximité' est formalisée à l'aide d'une mesure basée sur la comparaison des contextes relationnels des concepts considérés.

Une fois l'étape de pré-alignement des deux ontologies candidates est achevée, on passe à l'étape d'alignement. Nous proposons alors de choisir une ontologie, appelée ontologie cible ou de référence et une autre ontologie considérée comme l'ontologie source. Le processus d'alignement se base sur l'emploi des mesures de similarité afin de traiter les différentes formes d'hétérogénéités. Elle génère en sortie, un ensemble d'alignements entre les entités. Cette phase est scindé en deux principales étapes : (i) Etape d'ancrage, et (ii) Etape de dérivation.

L'objectif de nos travaux ne se résume pas à aligner des ontologies existantes mais plutôt de représenter l'information utile. Pour cela, nous nous sommes attachés à proposer les éventuels placements des éléments d'enrichissement dans l'ontologie cible, en spécifiant les types de relations d'intérêt (à savoir les relations de subsomption, de proximité). Dans ce contexte, peu de travaux dans la littérature, à notre connaissance, précisent le type d'alignement généré. Afin de réduire les incohérences et les redondances des résultats. La

validation des résultats est assurée par l'expert du domaine, qui lui seul peut juger de la pertinence des relations déduites.

Finalement, l'ontologie cible est enrichie par de nouvelles connaissances provenant de l'ontologie source. On distingue deux modes d'enrichissement : l'enrichissement par clusters conceptuel et l'enrichissement des concepts. Pour le premier type, il s'agit d'enrichir l'ontologie noyau par des groupements de concepts. Dans le deuxième cas, les concepts de l'ontologie source qui apparaissent dans l'ensemble des alignements produits sont (pour certains d'entre eux) considérés comme des concepts candidats pour l'enrichissement.

- *Deuxième Contribution : La fouille de connaissances pour l'enrichissement relationnel de l'ontologie*

Dans une deuxième contribution, nous nous intéressons à *l'enrichissement relationnel* permettant d'enrichir l'ontologie par des associations établies entre les concepts. En effet, ces relations jouent un rôle important dans l'interprétation des mammographies. Pour ce type d'enrichissement, nous proposons l'exploitation de bases de données issues des événements passés relatives aux patients antérieurement diagnostiqués. Le principal objectif dans l'analyse des bases de données existantes est de fournir aux acteurs du domaine une information concise et structurée (nouvelles connaissances) décrivant le contenu des données analysées [Marinica, 2010]. A cette fin, on propose d'employer les règles d'association en tant que technique de fouille de données. Le choix de cette technique est dû à la simplicité et la compréhensibilité du modèle des motifs extraites. En somme, le principal objectif des règles d'association dans notre travail est d'identifier les cooccurrences des items et les relations d'implications entre les observations dans une mammographie et la classification correspondante pour l'amélioration de l'ontologie de référence.

Un problème potentiel lié aux règles d'association étant le grand nombre des règles générées ce qui limite leur utilisation. Pour cette raison, nous pensons qu'il est intéressant de filtrer et classer les règles générées afin de faciliter leur lecture par l'expert (qui lui seul peut confirmer la pertinence ou non de ces associations). Pendant la phase de post processing, on s'est intéressé dans un premier temps à filtrer/éliminer les connaissances déjà modélisées dans la base cible. Dans un deuxième temps, nous proposons de classer les nouvelles connaissances en termes d'importance afin d'aider l'expert à se focaliser aux règles les plus importants. Cette importance est étroitement liée aux attributs mis en relation. A cette fin, une distance sémantique est proposée en se basant sur l'ontologie de domaine. La validation des règles entraine l'enrichissement de l'ontologie cible par de nouvelles relations.

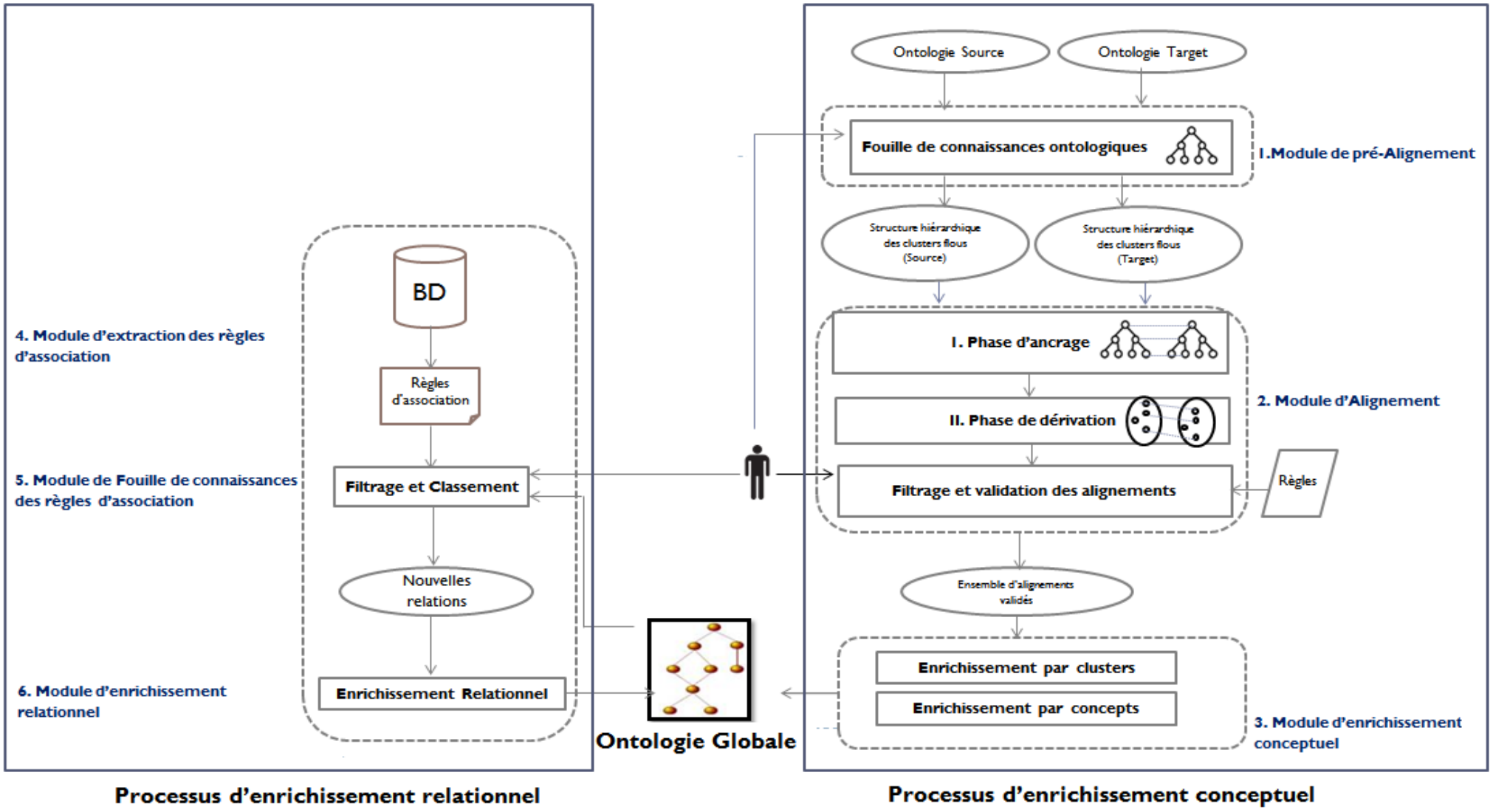

Figure III.2: Architecture de l'approche d'enrichissement d'ontologie basée sur la fouille de connaissances

3.1 Pré-Alignement d'ontologies basé sur la fouille des connaissances ontologiques

Afin de simplifier la tâche d'alignement des ontologies et réduire sa complexité, nous admettons qu'il est intéressant d'introduire une étape de préparation des ontologies basée sur la fouille de connaissances ontologiques. A travers cette étape, nous cherchons à découvrir une nouvelle répartition des concepts tout en cassant la structure monolithique et rigide de l'ontologie. Ceci est réalisé à travers la réorganisation de l'ontologie en une *structure hiérarchique des clusters flous des concepts*. En effet, l'utilisation des clusters permet de grouper de manière concise les concepts similaires, où ces derniers peuvent avoir différents degrés d'adhérence à plusieurs clusters simultanément, indiquant à quel degré un élément peut appartenir.

Partant de l'ensemble original de concepts ontologiques, l'algorithme de clustering divisif/partitif est appliqué d'une manière itérative. La brique de base de notre méthode est l'algorithme de clustering : C-medoides flous (FCMdd) qui sera appliqué sur les concepts de l'ontologie. A chaque niveau, chaque cluster est vérifié s'il représente véritablement un groupement des concepts dense. Dans le cas contraire, le cluster en question est subdivisé en des sous- clusters. Ainsi, la hiérarchie est élargie pour inclure de nouveaux clusters plus spécifiques. Chaque cluster est introduit par un point spécifique et descriptif permettant de la représenter.

3.1.1 FCMdd pour le clustering de l'ontologie

Fuzzy C-Medoid-FCMdd- [Bezdek, 1981] est une technique de clustering non supervisée floue qui représente une variante de l'algorithme Fuzzy C-Means [Bezdek, 1973] appliqué aux données relationnelles. Le FCMdd introduit la notion de sous-ensembles flous dans la définition des clusters: chaque élément appartient à chaque groupe avec un certain degré, et chaque cluster est caractérisé par un concept représentatif appelé medoid.

Cet algorithme nécessite la connaissance préalable du nombre de clusters et cherche l'espace de clustering possibles des concepts en optimisant d'une manière itérative la fonction objective J_{FCM}. Ainsi, il délivre un degré d'appartenance (compris entre 0 et 1) à chaque cluster pour chaque élément.

Nous avons adapté l'algorithme FCMdd de la manière suivante [Idoudi, et al., 2015]: Considérons l'ensemble des concepts X_n , l'ensemble de medoides de clusters (appelés aussi prototypes) v_i, le degré d'appartenance μ_{ik} d'un concept x_k par rapport à un cluster C_i est exprimé dans [0,1]. La fonction objective à minimiser est introduite comme suit:

$$\boldsymbol{J_{FCM}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^{m}\, d(x_k - v_i)^2} \qquad \textbf{(III.1)}$$

Où $x_k \in X_n$ désigne un concept, $d(.)$ est la mesure de distance sémantique entre deux concepts, $\boldsymbol{v_i}$ désigne le medoid du cluster C_i ; $1 \leq i \leq c$ (c est le nombre des clusters) et m est le degré de flou indiquant le degré de floue de la partition flou (m est strictement supérieur

à 1). Lorsque m est proche de 1, chaque concept est affecté intégralement au cluster le plus proche.

$U = [u_{ik}]$ représente la matrice des degrés d'appartenance des concepts aux différents clusters:

$$U = \begin{pmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{c1} & \cdots & u_{cn} \end{pmatrix} \quad \text{(III.2)}$$

$$\forall i \, \varepsilon\{1..C\}, \forall k \, \varepsilon\{1..N\}; u_{ik} \varepsilon [0,1] \quad et \; \forall k \, \varepsilon\{1..N\} \sum_{i=1}^{c} u_{ik} = 1$$

Où la $k^{ème}$ colonne U_k = (u_{1k}, u_{2k} ...u_{ck}) contient les c degrés d'appartenance du $j^{ème}$ élément aux c sous-ensembles flous.

Le degré d'appartenance d'un élément x_k est défini compte tenu de tous les autres clusters c_i . Il est calculé comme suit :

$$u_{ik} = \frac{\left(\frac{1}{d(x_k, v_i)}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\frac{1}{d(x_k, v_j)}\right)^{\frac{1}{m-1}}} \quad \text{(III.3)}$$

Le médoïde v_i de cluster C_i est formellement calculé comme suit :

$$v_i = \arg\min_{x_k \in C_i} \sum_{b \in C_i} d(x_k, b) u_{ik} \quad \text{(III.4)}$$

Le medoide d'un cluster [Fanizzi, et al., 2009] est représenté par le concept dont la distance moyenne par rapport aux autres éléments pondérés par leurs degrés d'appartenance est minimale (Figure III.3).

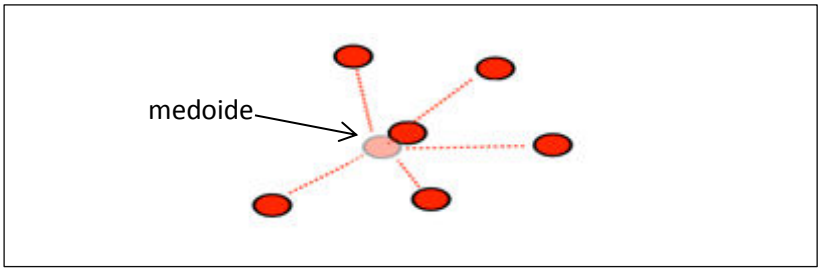

Figure III.3: Représentation de Medoid

Figure III.4: L'algorithme FCMdd pour le clustering d'ontologie

Entrées: $X = \{x_1, \dots, x_n\}$: L'ensemble des concepts ontologiques,

c : Nombre de clusters ,
m: Degré de flou,
$V = \{v_1, v_2, \dots, v_c\}$: set of medoids,
MaxIter : Nombre maximum d'itérations.

Sorties: C_c : Ensemble des clusters de concpets

Debut
Initialiser la matrice des degrés d'appartenance u_{ik} pour $i = 1, \dots, c, k = 1, \dots, n,$
Initialiser aléatoirement l'ensemble des medoides $V = \{v_1, v_2, \dots, v_c\}$,
Iter=0,
Repéter
Calculer les degrés d'appartenance u_{ik} for $i = 1 \dots c$ and $k = 1 \dots n$ According to (III.3);
Mise à jour de V v_i; i=1…c according to (III.4);
$V_{ancien} = V$;
Iter=iter+1 ;
Jusqu'à ($V_{ancien} = V$ // convergence ou iter=MaxIter)
Retourner C_c

La Figure III.4 présente l'algorithme FCMdd composé par les actions suivantes : premièrement, on initialise arbitrairement la matrice, le paramètre *m* et le nombre *c* de clusters qu'on souhaite avoir. Les équations **III.3** et **III.4** sont, ensuite calculées itérativement. L'algorithme alterne alors entre le calcul de la matrice d'appartenance U (en se basant sur les valeurs des medoids *V*), et le calcul de nouveaux medoids *V* (en se basant sur les valeurs de U), jusqu'à ce qu'à la convergence des medoides (les medoides deviennent stables).

3.1.2 Clustering Hiérarchique flou de l'ontologie

Dans cette partie, nous allons détailler notre algorithme de clustering hiérarchique flou de l'ontologie basé sur l'emploi de la technique FCMdd. Cette méthode, permet de construire une hiérarchie de clusters flous des concepts.

Partant de l'ensemble original de concepts ontologiques, la technique FCMdd est appliquée successivement. Le premier niveau de la hiérarchie, reflète les grands axes du domaine d'étude. Ce niveau sert à catégoriser les connaissances de domaine en des grandes classes, de thématiques ciblées. Ceci permet de faciliter les taches de visualisation et la recherche d'information. Par exemple, le domaine mammographique est caractérisé, selon l'expert de domaine, par quatre principales catégories : '*les entités anatomiques*', 'les e*ntités conceptuelles*', '*les anomalies apparentes'* et 'les d*iagnostics*'.

D'une manière itérative, chaque cluster est vérifié s'il représente véritablement un groupement cohérant des concepts homogènes. Dans ce cas, le cluster est subdivisé en deux nouveaux sous clusters. La décision de re-clustering dépend de deux paramètres. Le premier paramètre est lié à la cardinalité du cluster. Pour ce fait, un seuil N_{min} est fixé au préalable. Un

cluster ayant un nombre de concepts inférieur à N_{min} ne peut être re-clusterisé de nouveau. Ce paramètre permet de nous éviter la génération de clusters de petite taille qui peuvent induire une perte d'information. Le deuxième paramètre est lié à la mesure de la densité [Bordogna, et al., 2009] appelée aussi la qualité intra cluster. Cette dernière dépend principalement de deux facteurs : (i) la distance entre les concepts au sein du cluster (ii) les degrés d'appartenance des concepts par rapport à ce cluster. La mesure de la densité d'un cluster C est introduite est introduite comme suit :

$$\boldsymbol{\Delta(C) = 2.\left(\frac{\sum_{x_i \varepsilon C} \mu_i \, d(x_i, v_i)}{\sum_{x_i \varepsilon C} \mu_i}\right)} \quad \textbf{(III.5)}$$

Où $d(x_i, v_i)$ est la distance sémantique entre le concept x_i et le medoid du cluster C; $\mu_i$est le degré d'appartenance du concept x_i au cluster C.

Une faible valeur de $\Delta(C)$, par rapport à une valeur seuil prédéfinie α, signifie que les éléments au sein du cluster sont bien corrélées, et par conséquent, le clustering de ce cluster est inutile. Par contre, si la densité d'un cluster est élevée, cela nous amène à déduire que ce groupement de concepts ne représente en aucun cas une structure optimale, de ce fait, l'application de clustering s'impose.

Une fois qu'on a décidé le partitionnement du cluster en question, l'algorithme sélectionne les concepts dont les degrés d'appartenance sont supérieurs ou égaux à un seuil fixé par un expert $Degré_{seuil}$. En effet, une telle sélection est fondée sur l'idée qu'on ne peut re-clustériser que les concepts qui sont de bons représentants du cluster. Un concept dont le degré d'appartenance est inférieur à un $Degré_{seuil}$ ne mérite pas d'être re-clusterisé dans des sous clusters plus spécifiques. Ceci permet également, une convergence rapide de l'algorithme. Pour l'initialisation de l'algorithme, l'utilisateur sélectionne aléatoirement deux medoides respectifs aux nouveaux sous-clusters.

À ce stade, on suppose partir d'une partition 'rigide' de concept (une des conditions primordiales de l'algorithme de FCMdd standard), c.à.d. les degrés d'appartenance des éléments à clustériser sont égaux à 1. Une fois que le cluster p en question est clustérisé, les nouveaux sous-clusters p', p'' sont ajoutés à la hiérarchie, et le nouveau degré d'appartenance d'un concept x dans chaque sous-cluster est défini par le produit de l'ancienne valeur d'appartenance à p et la nouvelle valeur calculée. Formellement, le nouveau degré d'appartenance est calculé par l'équation (III.6) :

$$\boldsymbol{\mu'_i(x) * \mu(x)} \quad \textbf{(III.6)}$$

De ce fait, lorsqu'un élément x a une faible adhésion au cluster père, son nouveau degré d'appartenance au nouveau sous-cluster sera de même. Une conséquence importante de cette construction c'est que la structure obtenue définit en tout point de la hiérarchie un cluster flou, ceci dit, la somme des degrés d'appartenance d'un concept donné dans toutes les feuilles est égale à 1 (voir la Figure III.6).

Figure III.5: Clustering Hiérarchique Flou de l'Ontologie

Input: N_{min} , α, $Degré_{seuil}$
VecteursClusters C[niveau L] : Liste des vecteurs de niveau L.
Paramètre de l'algorithme FCMdd
Output: Structure hiérarchique des clusters

Début
Niveau←1
VecteursClusters C[niveau L] ← VecteursClusters C[niveau 1]
Répéter
  Pour tout C ε VecteursClusters C[niveau L]
    D← Densité du cluster(C) ;
    If D> α ;
      Niveau← Niveau +1 ;
    C′, C′′←F2Mdd (C);
    Ajouter C′, C′′ au VecteursClusters C[Niveau L+1]
  Fin pour

Jusqu'à |C| ≤ N_{min}
Fin répéter
Retourner VecteursClusters
Fin

L'algorithme de clustering hiérarchique flou présenté dans cette section est explicité dans la Figure III.5.

La Figure III.6 illustre une structure hiérarchique des clusters conceptuels flous des concepts représentés par leurs medoides. Cette structure permet une nouvelle visualisation hiérarchique des connaissances pour chaque base ontologique. Le niveau zéro comporte l'ensemble d'origine des concepts de l'ontologie. Les niveaux supérieurs de la hiérarchie se caractérisent par un niveau d'abstraction de connaissances plus élevé. Ces clusters sont de plus en plus spécifiques en descendant dans la hiérarchie jusqu'à arriver au niveau le plus spécifique (niveau n).

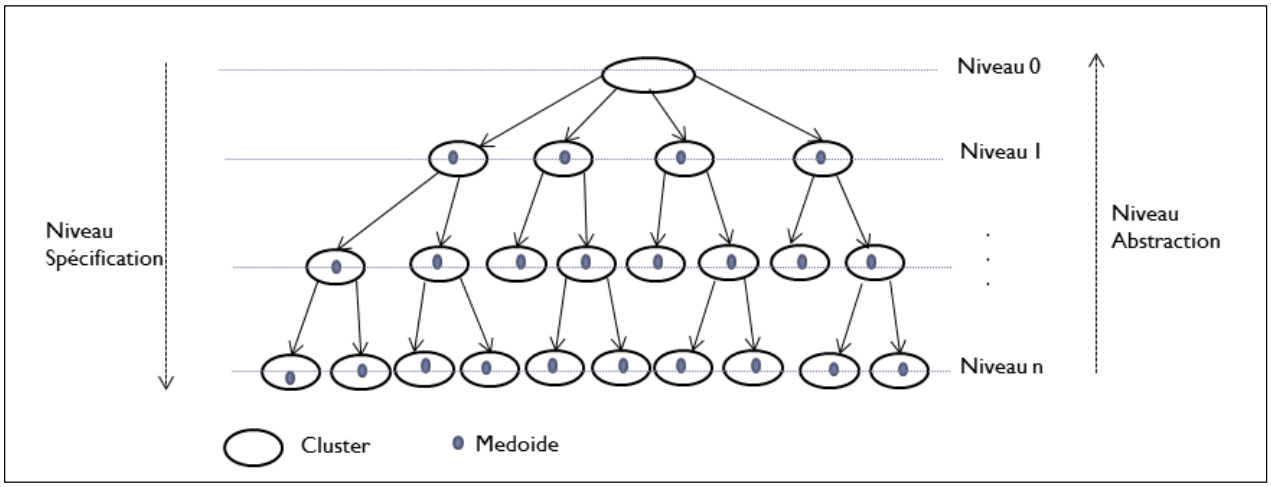

Figure III.6: Vision Hiérarchique des clusters conceptuels flous

3.1.3 Proposition d'une nouvelle distance sémantique

Différentes techniques de similarité d'ontologies ont été proposées dans la littérature telles que les distances de [Rada, et al., 1989], [Sussna, 1993] [Wu, et al., 1994] (voir Annexe 1 : Les mesures de similarité d'ontologie).

L'un des points fort de la distance sémantique proposée par rapport aux distances proposées dans la littérature est sa capacité à exploiter toutes les relations ontologiques notamment les relations sémantiques (c.-à-d. les liens entre les concepts autres que les liens hiérarchiques) dans l'évaluation de la similarité.

En effet, deux concepts sont d'autant plus similaires lorsque leurs contextes/ voisinages sont semblables. La notion de voisinage d'un concept a été inspirée de Maedche et staab dans [Maedche, et al., 2002], qui définissent la sémantique d'un concept au sein d'une ontologie par l'ensemble de ses généralisants et de ses spécialisants.

Ainsi, le contexte d'un concept donné est construite à partir des relations hiérarchiques 'is-a' ainsi que les relations de type 'object-property' dans l'ontologie. Le choix d'introduire ce type de relations se justifie par le fait que deux concepts qui se relient à une même entité peuvent partager en quelque sorte une partie de son interprétation sémantique.

De ce fait, si le voisinage d'un concept se limite à son père et/ou son fils, on obtient un voisinage assez pauvre, qui ne permettrait pas de tenir compte assez nettement du contexte d'interprétation du concept considéré au sein de l'ontologie.

Formellement, on définit le contexte d'un concept C(x) par ses ascendants, descendants, voisins, et les concepts qui sont en relation à travers la relation "object-property" :

$$C(x) = \{xi | (x, xi) \varepsilon E \cup (xi, x) \varepsilon E \cup x\}.$$

Avec E est l'ensemble des relations entre les concepts de l'ontologie.

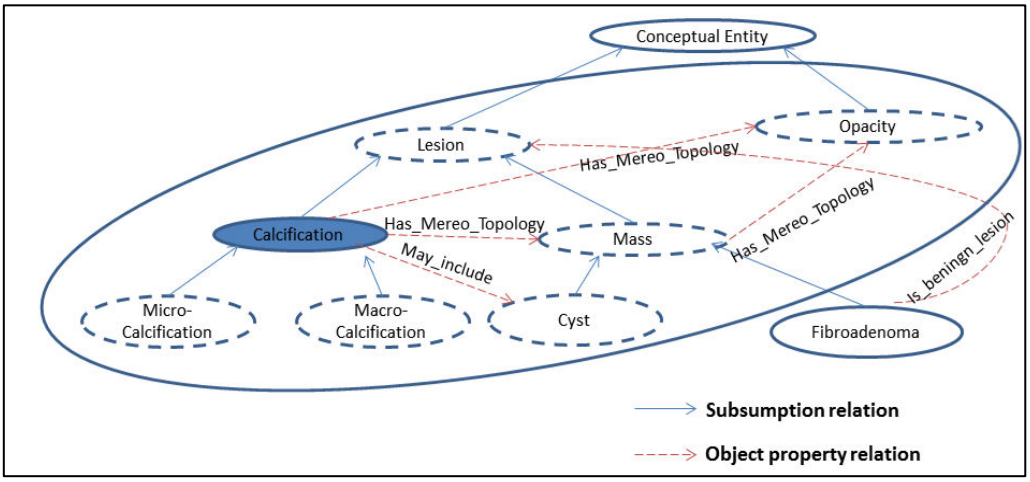

Figure III.7: Exemple du contexte du concept 'Calcification'

La notion de contexte examine le positionnement d'un concept par rapport aux autres concepts de l'ontologie (Figure III.7). Par conséquent, on définit la distance sémantique entre deux concepts donnés en se basant sur leurs contextes par l'équation suivante :

$$\boldsymbol{d(x_i,\ x_j)\ =\ 1-\ (2.\frac{|C(x_i)\cap C\ (x_j)\ |}{|C\ (x_i)|+|C\ (x_j)|})} \qquad \textbf{(III.7)}$$

$|C(x_i)\cap C\ (x_j)\ |$Représente le nombre des nœuds communs entre les deux contextes et $|C\ (x_i)| + |C\ (x_j)|$ sert à la normalisation de la distances. Cette distance exprime que plus les deux contextes se chevauchent, plus ils sont moins distants. L'équation (III.11) satisfait les propriétés de la distance à savoir :

- $d(x,y) \geq\ 0\ ; \forall\ x$ and y;
- $d(x,y)\ =\ 0 \Leftrightarrow x\ =\ y$;
- $d(x,x)\ =\ 0 \forall x$;
- $d(x,y)\ =\ d(\ y,x)\ \forall\ x\ and\ y$;

L'équation garantit ainsi que plus il y a des nœuds en commun entre les deux concepts en question, plus la similarité est élevée. En outre, l'équation montre que la distance dispose de plusieurs propriétés par exemple :

- La normalisation, $0 \leq d(x_i,\ x_j) \leq 1$,
- La symétrie $d(x_i,\ x_j)=d(x_j,x_i)$.

3.2 Alignement d'ontologies basé sur la fouille de connaissances ontologiques

La méthode d'alignement d'ontologies que nous proposons dans cette section tire profit de la nouvelle réorganisation des concepts dans les ontologies candidates. Deux ontologies candidates constituent l'entrée de cette étape. La première ontologie est appelée ontologie noyau, cible ou de référence, celle-ci est supposée offrir une couverture plus large et plus détaillée du domaine que celles fournies par les diverses ontologies d'applications. La deuxième ontologie, dite ontologie cible, sert de connaissances tout au long du processus d'alignement.

L'étape d'alignement des ontologies de domaine est réalisée en trois temps (Figure III.8). La première phase du processus, dite **phase d'ancrage**, consiste à aligner les clusters des deux structures hiérarchiques (cible et source) pour générer finalement les clusters sémantiquement proches des deux ontologies. Les résultats générés, qui se présentent sous la forme des couples de clusters alignés, sont utilisés dans la deuxième phase d'alignement appelée **phase de dérivation** qui détermine les alignements possibles entre les concepts des couples de clusters similaires. Finalement, on procède à l'étape de **filtrage et validation d'alignements** qui consiste à éliminer les alignements erronés.

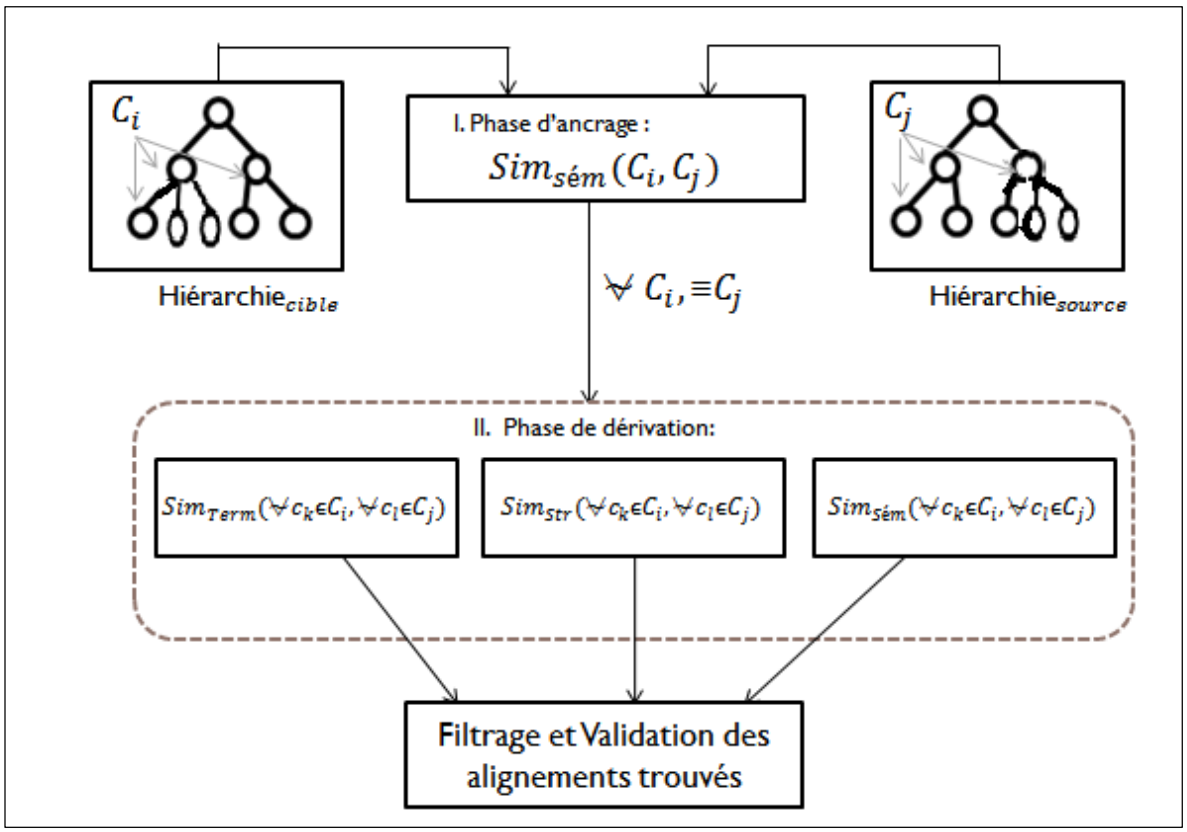

Figure III.8: Processus d'alignement d'ontologies source et candidat

3.2.1 Phase d'ancrage

Cette étape a pour but, l'identification des paires de clusters (1 à 1) sémantiquement équivalents (provenant des deux hiérarchies à aligner). Le principe d'alignement des hiérarchies est réalisée niveau par niveau. Commençant du premier niveau (voir Figure 3.2), les couples des clusters sémantiquement les plus proches sont identifiées, par la suite, nous procèdons itérativement, à la comparaison de leurs sous clusters respectifs jusqu'à atteindre le niveau le plus bas de la hiérarchie (niveau n). Chaque cluster de l'ontologie source est finalement aligné avec un seul cluster de l'ontologie cible. L'idée principale de cette méthode, est basée sur le fait que si deux clusters C_1^Tet C_1^S(voir la Figure III.9) sont sémantiquement similaires, alors leurs sous- clusters respectifs C_{1i}^T, C_{1j}^Tet C_{1i}^S, C_{1j}^S le seront aussi. Le calcul de la proximité sémantique sera mené alors sur chaque couple des sous-clusters (<C_{1i}^T, C_{1i}^S>, <C_{1i}^T, C_{1j}^S>, <C_{1j}^T, C_{1i}^S>, <C_{1j}^T, C_{1j}^S>).

Pour cela, nous proposons d'utiliser deux types de rapprochement. Le premier type se base sur la comparaison sémantique des deux medoides prédéfinis des clusters source et cible. Le deuxième type se base sur le nombre d'éléments partagés des clusters candidats.

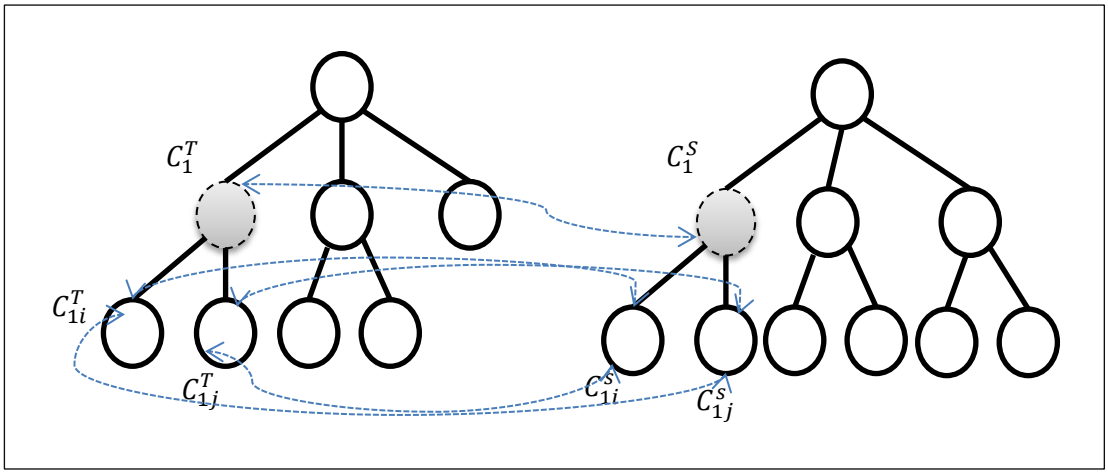

Figure III.9: Propagation de la similarité pour la comparaison des clusters des hiérarchies source et cible

a. Calcul de la proximité sémantique des médoïdes

Le premier type de comparaison basé sur les médoïdes, est justifié par le fait que si deux médoïdes v_i, v_j sont sémantiquement équivalents, il est fort probable que leurs clusters respectifs C_i, C_j le seront aussi. La similarité sémantique est calculée en utilisant la technique sémantique basée sur l'utilisation d'une ressource externe appelée «WordNet »[2].

Cette technique permet de chercher les relations sémantiques en s'appuyant sur la base lexicale WordNet[3]. Ce thesaurus de langue anglaise consiste à collectionner des termes (noms, verbes, etc) en des *'synsets'*. Ce dernier permet de regrouper l'ensemble de synonymes dénotant un terme donné. Les termes sont associés sous une forme lexicalisée (sans marque de féminin ni de pluriel). Les synsets sont reliés entre eux par des relations sémantiques : relation de généralisation/spécialisation et relation composant/composé.

L'alignement des clusters dans la phase d'ancrage consiste à trouver les correspondances entre les medoides en admettant l'hypothèse suivante : Deux concepts C_x et C_y sont sémantiquement similaires s'il existe un synset S de C_x qui soit aussi un synset de C_y ou qui soit lié au synset de C_y par une suite de relations de généralisation/spécialisation dans WordNet. Il s'agit donc de chercher pour chaque medoide du cluster source, les synsets qui correspondent à ses concepts associés. Plus on a des éléments en commun entre les synsets correspondants aux medoides, plus ces derniers sont sémantiquement proches.

Afin de calculer la similarité entre deux concepts c_x et c_y en exploitant les liens sémantiques existants dans WordNet, on utilise la fonction suivante proposée dans [Fareh, et al., 2013]:

$$\mathbf{sim_{sémantique}}(\mathbf{c_x}, \mathbf{c_y}) = \frac{|\mathbf{Syn}(\mathbf{c_x}) \cap \mathbf{Syn}(\mathbf{c_y})|}{\mathbf{min}(\mathbf{Syn}(\mathbf{c_x}), \mathbf{Syn}(\mathbf{c_y}))} \quad \textbf{(III.8)}$$

[2] http://wordnet.princeton.edu/
[3] http://wordnet.princeton.edu/

Où $|Syn(C_x) \cap Syn(C_y)|$ est la cardinalité des éléments en commun entre les synsets correspondants aux concepts c_x , c_y . $\min(Syn(c_x), Syn(c_y))$ est le minimum des deux ensembles de cardinalités de synsets $Syn(c_x), Syn(c_y)$.

b. Calcul du nombre d'éléments partagés

Le deuxième type de comparaison basé sur le nombre d'éléments partagés des clusters candidats consiste à comparer syntaxiquement les labels des éléments appartenant aux clusters candidats. Plus le nombre des éléments partagés est élevé, plus ces deux clusters sont sémantiquement proches. L'utilisation de la technique syntaxique est justifiée par sa simplicité d'implémentation, son cout réduit et son efficacité lorsqu'il s'agit d'utiliser des ontologies portant sur le même domaine. La comparaison des labels des concepts est définie comme suit :

$$\text{SimLex}\left(\text{Label}(c_x), \text{Label}(c_y)\right) = \begin{cases} 1 \;\; Si \; \text{Label}(c_x) \equiv \text{Label}(c_y) \\ 0 \; sinon \end{cases}$$

Le calcul de proximité des clusters basé sur le nombre d'éléments partagés est effectué comme suit :

$$\mathbf{proximity_{shared\_elemnts}}(C_i, C_j) = 2.\frac{|C_i \cap C_j|}{|C_i| + |C_j|} \qquad \textbf{(III.9)}$$

Où $|C_i \cap C_j|$ représente la cardinalité des éléments en commun entre les deux clusters $C_i \; et \; C_j$. $|C_i|$ représente le nombre des éléments dans$|C_i|$, $|C_j|$ représente le nombre des éléments dans$|C_j|$.

c. Combinaison

La décision des couples de clusters, sémantiquement les plus proches, est basée sur le calcul de la moyenne des proximités employées ci-haut. L'hypothèse de base est fondée sur le fait que la proximité sémantique globale de deux clusters est supérieure à celle entre deux autres clusters. La combinaison de ces derniers est calculée comme suit :

$$\mathbf{proximity}(C_i, C_j) = ß\mathbf{sim_{semantic}}(C_i, C_j) + (1 - ß)\mathbf{proximity_{shared\_elemnts}}(C_i, C_j) \qquad \textbf{(III.10)}$$

Avec ß**ϵ** [0,1].

d. Scénarios d'alignement des clusters

Dans cette étape, on distingue trois scénarios d'alignement possibles entre les clusters (source et cible) :

-'*Fully-aligned clusters' (*Figure III.10): les clusters source et cible sont mis en correspondance un par un, de même pour leurs fils directs. Ce scénario représente le cas le plus fréquent pour les ontologies du même domaine.

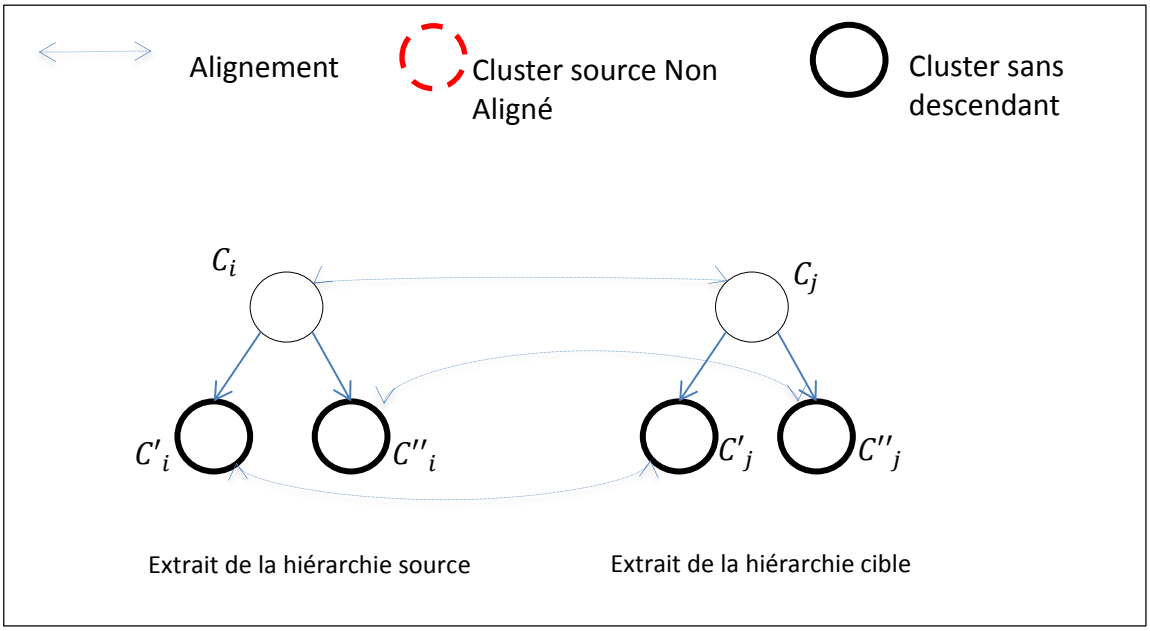

Figure III.10: Full-aligned clusters

-'*Non-aligned clusters*' (Figure III.11) : Dans ce scénario, un cluster source $\mathbf{C''_i}$ n'est aligné avec aucun des clusters cibles. Ceci est expliqué par la différence des perspectives modélisées au sein des ressources ontologiques. Le cluster source non-aligné peut appartenir au plus bas niveau de la hiérarchie (a) ou avoir des sous clusters descendants (b) (appartenant aux différents niveaux de granularité). Ce type de scénario est de grand intérêt pour l'utilisateur puisqu'il servira à l'enrichissement conceptuel de l'ontologie cible.

-'*Multi-aligned clusters*' (Figure III.12): Dans ce scénario, un cluster source C''_i sans descendants est aligné avec un cluster cible C''_i disposant de sous clusters. Ce cluster est re-comparé, ainsi, avec les sous clusters cibles pour retenir finalement le cluster cible qui lui est égale.

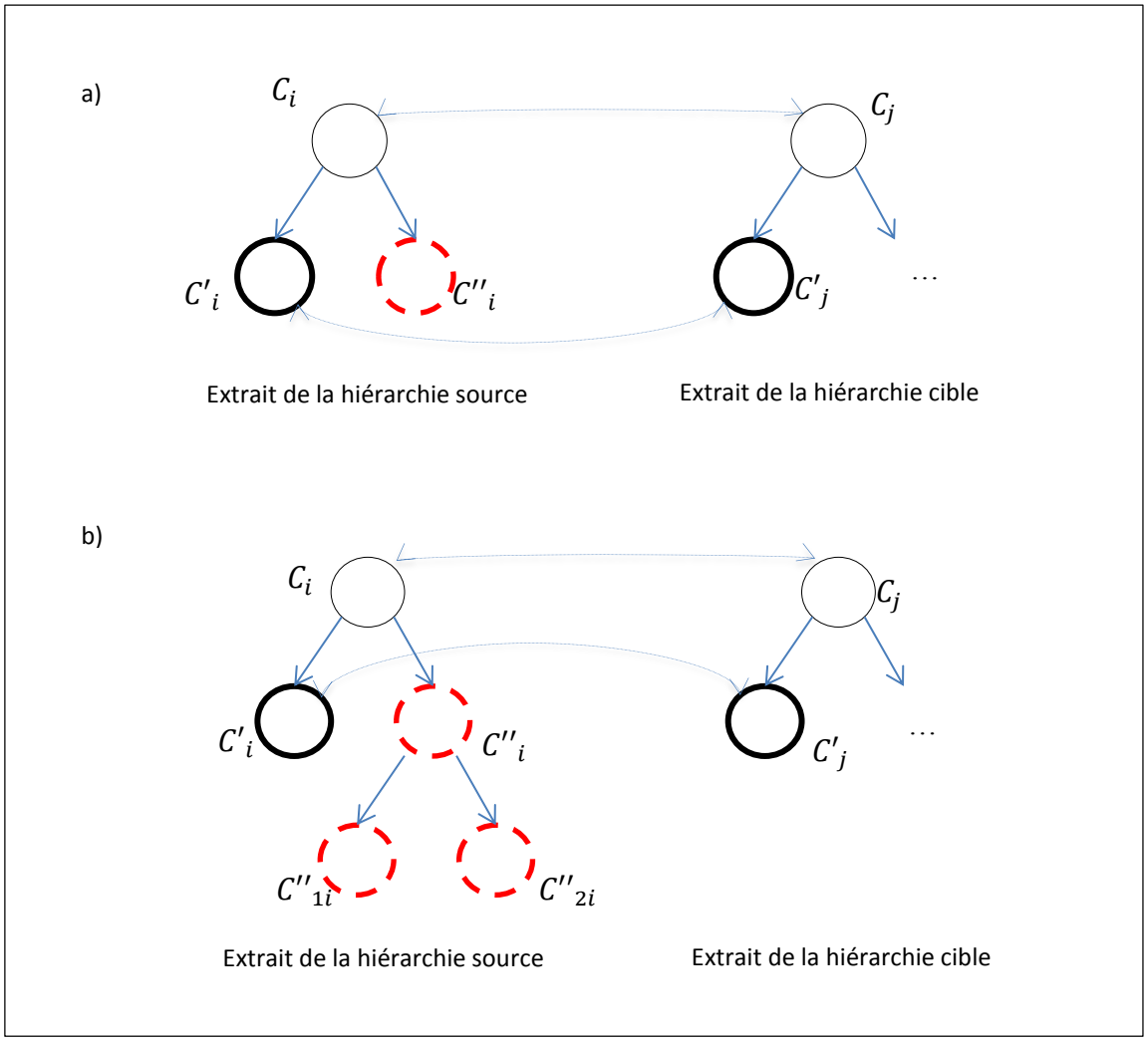

Figure III.11: Non-aligned clusters

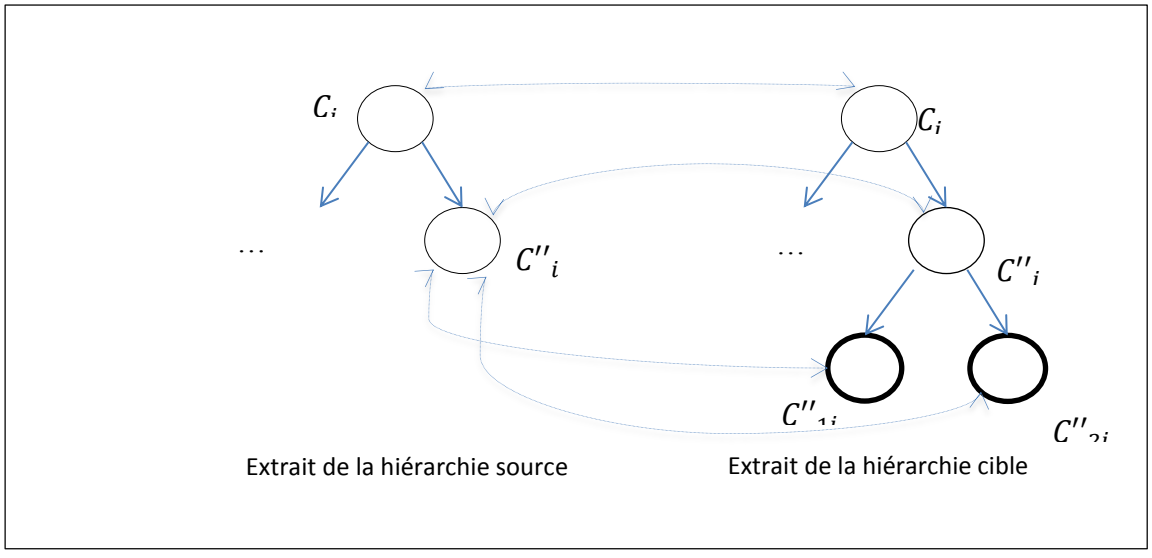

Figure III.12: Multi-aligned clusters

-*'Half-aligned clusters'* (Figure III.13): Dans ce scénario, un cluster source C''_j avec descendants est aligné avec un cluster cible sans descendants. Ceci est dû aux différents niveaux de granularité des deux bases de connaissances. Dans ce cas, on retient la paire des clusters (C''_j , C''_i).

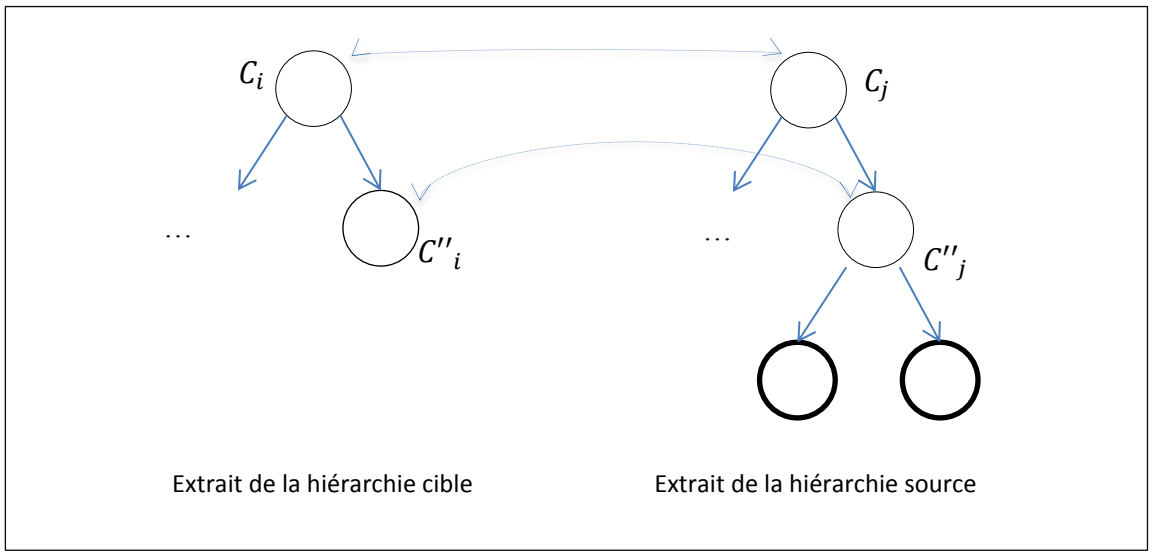

Figure III.13: Half-aligned clusters

3.2.2 Phase de dérivation

A ce stade, l'espace de recherche de correspondances entre les ontologies en entrée est réduit aux couples des clusters sémantiquement similaires. Le problème d'alignement est transformé, ainsi, en un alignement 'light-weight' des groupements de concepts.

Afin de pallier aux différents problèmes d'hétérogénéité, nous proposons d'utiliser une combinaison de techniques d'alignement : syntaxique, structurelle et sémantique. L'utilisation de cette diversité de techniques d'alignement aide à fouiller les concepts qui peuvent enrichir l'ontologie de référence. En effet, on s'intéresse dans cette étape, aux nouveaux concepts extraits de l'ontologie source. Ces derniers sont déterminés par l'extraction des relations de subsomption et disjonction par rapport aux connaissances de référence. Les relations d'équivalence trouvées ne sont pas d'un grand intérêt pour l'expert puisqu'elles n'apportent pas de nouvelles connaissances. Néanmoins, ce type de relation permet efficacement de filtrer les relations de subsomption pour pouvoir positionner les nouveaux concepts.

Dans cette étape, un ensemble de correspondances entre les concepts des ontologies sources et ceux de l'ontologie cible. Ces correspondances peuvent être des relations d'équivalence ou de subsomption.

Les alignements correspondants aux couples de clusters sont, ensuite, groupés dans un seul fichier. Ces résultats d'alignement nécessitent ainsi un filtrage pour pallier aux problèmes de redondance et garder finalement que les correspondances d'intérêt pour l'expert.

Nous détaillons par la suite les techniques d'alignement ainsi que l'utilité de chacune pour la détermination de la nature de correspondance entre les entités alignées.

a. **Mesure de similarité syntaxique**

Les techniques de similarité syntaxique consistent à effectuer des mesures de similarité en se basant sur les étiquettes et les labels correspondants aux couples de concepts à aligner en comparant caractère par caractère. Deux éléments sont considérés comme sémantiquement proches si les termes qui les désignent sont syntaxiquement proches.

Cette technique permet de révéler les relations de type équivalence « isEq» et subsomption « is-A». L'égalité stricte des chaines de caractères relatives aux deux concepts d'intérêt indique une relation d'équivalence. Dans le cas de l'inégalité, on procède à la vérification d'inclusion de labels. Dans ce cas, il s'agit d'une relation de subsomption entre ces concepts.

Dans la littérature, de nombreuses mesures de similarité syntaxiques ont été proposées jusqu'à présent. Parmi les mesures les plus utilisées dans le cadre de l'alignement d'ontologies : la distance de HAMMING, la similarité de JACCARD, la distance d'Edition, la distance de Levenshtein.

On a utilisé, dans ce travail, la mesure de Levenshtein [Levenshtein, 1966]en vue de son efficacité et sa simplicité dans l'implémentation. Cette mesure est calculée en se basant sur le nombre minimal d'opérations d'insertions, de suppressions et de substitutions de caractères nécessaires pour la transformation d'une chaine de caractère en une autre avec association des couts aux différentes éditions (par défaut 1). La fonction est définie comme suit :

$$\boldsymbol{Sim_{label}(c_1, c_2) = \frac{\delta(c_1, c_2)}{\max(|c_1|, |c_2|)}} \qquad \textbf{(III.11)}$$

avec $\delta(c_1, c_2)$ est la distance de Levenshtein. La fonction retourne une valeur dans [0,1].

Pour la vérification d'inclusion de labels, on vérifie si tous les mots de l'identificateur d'un concept A sont inclus dans les mots d'un concept B, alors B sera considéré comme plus spécialisé que A (B is-a A).

b. **Mesure de similarité structurelle**

La similarité structurelle entre deux éléments est calculée en fonction de leurs informations structurelles au sein de la hiérarchie ontologique. Dans la littérature, plusieurs mesures de similarité structurelle ont été proposées jusqu'à présent telles que les mesures de [Rada, et al., 1989], [Resnik, 1999].

Dans ce travail, nous proposons d'utiliser la mesure de Wu & Palmer [Wu, et al., 1994] en vue de son efficacité et sa simplicité. Cette technique se base sur le calcul de similarité entre deux concepts $c_1 \epsilon\, O_1$ et $c_2 \epsilon\, O_2$ en utilisant la similarité de leurs voisinages. Si le voisinage

$Sc(c_1, O_1)$ (à savoir les descendants et les ascendants) du concept c_1 est similaire au voisinage $Sc(c_2, O_2)$ du concept c_2 , alors c_1 et c_2 sont également similaires.

$$\boldsymbol{Sim_{str}(c_1, c_2)=\frac{\sum_{(rc_1,rc_2)\epsilon VR(c_1,c_2)} sim_{ling}(rc_1,rc_2)*sim_{WuPalmer}(c_1,c_2)*(1-|d_1-d_2|^2)}{|VR(|c_1,c_2|)|}} \quad \textbf{(III.12)}$$

Où :

- sim_{ling} est la similarité linguistique entre les concepts (rc_1, rc_2) qui désignent les voisinages des concepts(c_1, c_2).
- d_1=$sim_{WuPalmer}(c_1, rc_1)$.
- d_2=$sim_{WuPalmer}(c_2, rc_2)$.
- $VR(|c_1, c_2|) = \{(rc_1, rc_2)|\ rc_1\ \epsilon\ voisinage(c_1), rc_2\ voisingae(\ c_2)\}$, de telle sorte qu'on doit considérer le couple rc_1, rc_2, père-père ou fils-fils.

- $sim_{WuPalmer}(c_1, c_2) = \frac{depth(LCA(c_1,c_2))}{depth(c_1)+depth(c_2)}$ qui s'appuie sur la structure hiérarchique de WordNet.

- $LCA(c_1, c_2)$ désigne le petit ancêtre commun de c_1, c_2.
- $depth(LCA(c_1, c_2))$ désigne le nombre d'arcs séparant LCA(c_1, c_2) de la racine.
- $depth(LCA(c_1))$ désigne la longueur du chemin séparant c_1 de la racine.

c. **Mesures de similarité sémantique**

Cette mesure permet de valider les relations d'équivalence déduites par les techniques de similarité citées ci haut et déduire les relations de type subsomption« isEq» ou de proximité « isClose». Ces derniers peuvent ne pas être déterminés par les techniques structurelles ou terminologiques, ce qui justifie l'utilisation des techniques sémantiques à base d'une ressource lexicale tel que le WordNet. La mesure de similarité est calculée suivant la formule de l'équation (III.13).

d. **Validation et Filtrage des alignements**

Dans la phase de dérivation, l'utilisation de cette diversité de techniques d'alignement appliquées sur les couples de clusters sémantiquement similaires, permet de fouiller les concepts qui peuvent enrichir l'ontologie noyau. Ces alignements obtenus sont tous regroupés en un seul ensemble. A ce stade, on procède à la validation des résultats obtenus. Afin de catégoriser ces résultats selon le type de relation on propose l'utilisation d'un ensemble de règles inspirées du travail de [Messoudi, et al., 2013]. Pour tout couple de concepts $c_t\ \epsilon O_{cible}, c_s\ \epsilon O_{source}$ précédemment alignés, on procède ainsi (Figure III.14) :

- **Liste de relations d'équivalence « isEq»**:

Afin d'appuyer les relations d'équivalence déduites par les techniques terminologiques et structurelles, une vérification au moyen d'une technique sémantique semble intéressante.
Ces concepts sources sont reconnus comme redondant ou trop proche des concepts déjà existants et ne seront pas pris en compte dans l'étape d'enrichissement puisqu'ils ne permettent pas d'introduire de nouveaux concepts dans la cible.

- **Liste de relations de subsomption « isA»** :
  S'il existe une relation sémantique pour le couple de concepts vérifiant la propriété inclusion de label, on peut confirmer l'existence de la relation de subsomption.
  Ces alignements sont susceptibles de conduire à de nombreux enrichissements et augmentent le niveau de granularité dans l'ontologie.

- **Liste de relations de proximité** :
  Cette liste comporte les relations de type subsomption « isA» ou fraternité, elle est déduite pour les mappings qui sont seulement vérifiés par la technique sémantique $sim_{sémantique}(c_t, c_s) \geq \alpha_{sémantique}$. Réellement, deux concepts c_t, c_s ayant de labels différents ou peu proche ainsi qu'une similarité structurelle légèrement significative, peuvent révéler une relation sémantique implicite que seul l'expert pourra vérifier sa pertinence.

- **Liste de relations de disjonction :**
  Cette liste comporte tous les concepts c_s qui n'ont pas été mis en correspondance. En effet, ces concepts participent considérablement à l'extension de l'ontologie cible.

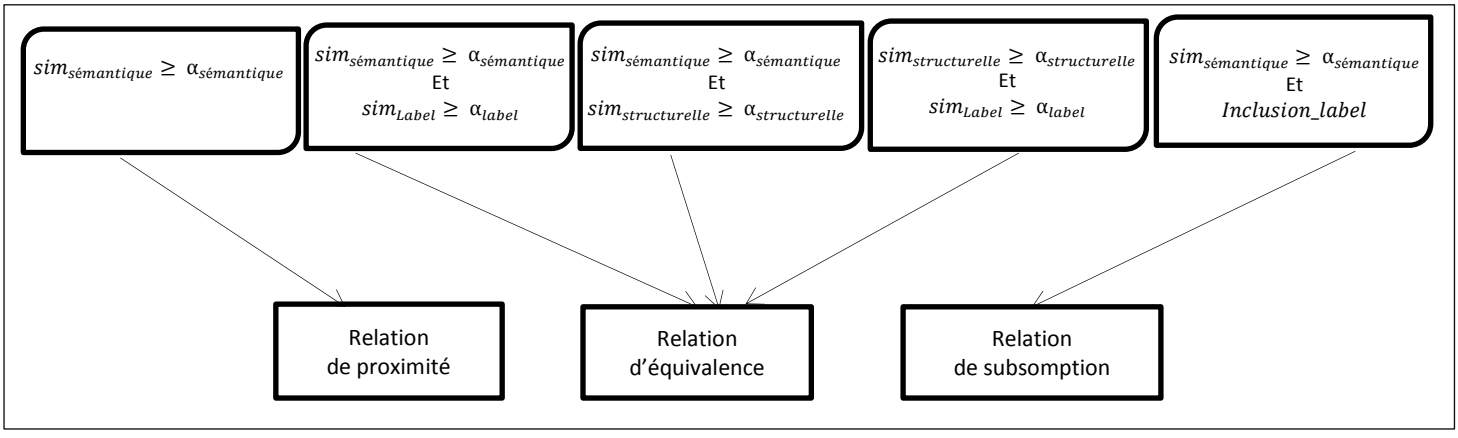

Figure III.14: Filtrage des alignements

Néanmoins, un concept c_t provenant de l'ontologie de référence peut figurer dans plusieurs alignements de même ou différents types de relations. Ceci est dû à la propriété floue des clusters en jeu ainsi que l'emploi de plusieurs techniques de similarité. Afin de

valider les résultats et résoudre les problèmes de conflits et/ou de redondances, une phase de filtrage s'impose. A cette fin, on propose d'employer un ensemble de règles:

- **Règle1** : Pour tout concept c_t qui apparait simultanément dans des relations d'équivalence (Figure III.15), on garde la relation qui maximise la mesure structurelle, l'autre relation sera rejetée. Ceci est dû au fait que deux concepts entièrement équivalents ne se présentent pas dans la même hiérarchie ontologique.

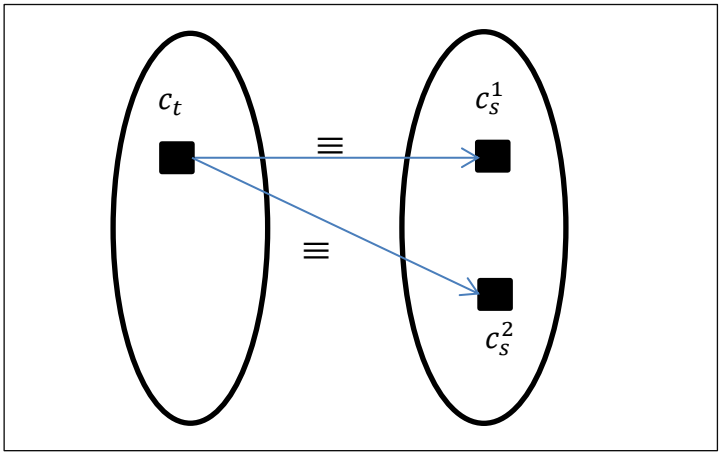

Figure III.15: Exemple d'alignement (Règle 1)

- **Règle2** : Un concept c_t peut également apparaitre simultanément dans des relations de différents types (subsomption et équivalence). Dans ce cas, on procède à la vérification, si le concept c_s^1 mis en correspondance d'équivalence avec c_t est le descendant ou ascendant du concept c_s^2 qui apparait dans la correspondance de subsomption avec c_t . Si c'est le cas, on garde seulement la relation d'équivalence, puisque la relation de subsomption ne nous apporte pas une nouvelle connaissance.

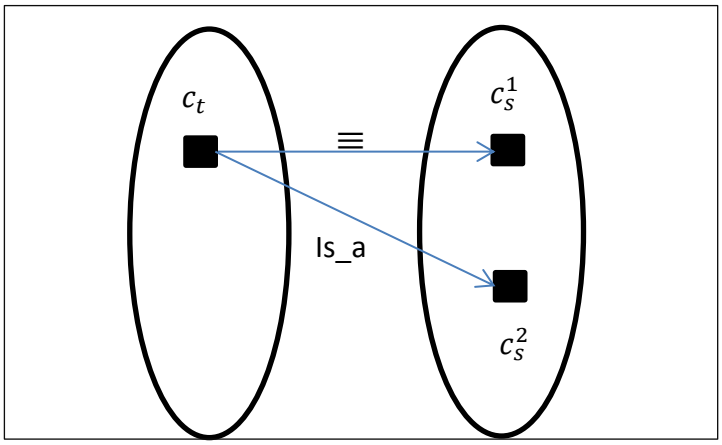

Figure III.16: Exemple d'alignement (Règle 2)

- **Règle3** : Un concept c_t peut également apparaitre simultanément dans des relations de différents types (de proximité et équivalence). Dans ce cas, on procède à la vérification, si le concept c_s^1 mis en correspondance d'équivalence avec c_t est

le descendant ou ascendant du concept c_s^2 qui apparait dans la correspondance de subsomption avec c_t . Si c'est le cas, on garde seulement la relation d'équivalence, puisque la relation de subsomption ne nous apporte pas une nouvelle connaissance.

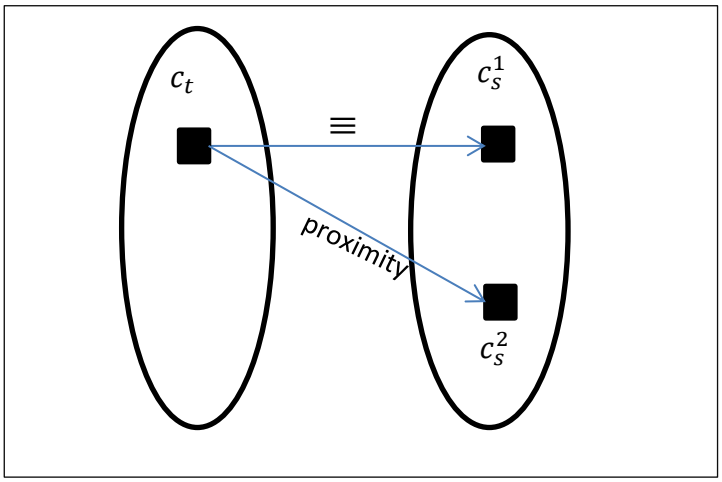

Figure III.17:Exemple d'alignement (Règle 3)

Ces introductions doivent être dirigées vers l'expert qui, seul, peut décider si une relation est bien pertinente pour un couple de concepts ou résulte d'une faute de frappe. Finalement, on garde seulement les alignements vérifiés et validés par l'expert de domaine.

3.3 Enrichissement conceptuel de l'ontologie

Nous nous intéressons, dans cette partie, à l'enrichissement par introduction de nouveaux concepts dans une ontologie dite cible. Dans ce cas, la tâche d'enrichissement est alors présentée comme composée de deux phases : l'une consiste en un enrichissement par des groupements de concepts qui n'ont pas été antérieurement alignés, le deuxième type consiste à utiliser les alignements jugés valides (Filtrées et reconnues par l'expert).

3.3.1 Enrichissement par clusters

Cette étape permet d'enrichir la hiérarchie de l'ontologie de référence par des clusters qui sont jugés disjoints par rapport aux clusters cibles. Un nouveau cluster se positionne comme descendant du cluster sémantiquement le plus proche de son cluster père. On distingue deux scénarios d'enrichissement possibles des clusters. Dans la Figure III.18_a, le cluster C_{i2}^s du niveau le plus spécifique (sans descendants) de la hiérarchie source et dont le cluster père C_i^s est sémantiquement proche de C_j^t, se place comme descendant du C_j^t. De même, pour le cas illustré dans la Figure III.18_b, où le cluster d'enrichissement dispose des sous clusters, l'enrichissement du groupe de clusters se fait en le plaçant comme descendant de C_j^t. Ceci signifie que si un cluster n'est pas aligné avec les clusters noyau, les sous clusters correspondants ne le sont pas. Cette étape d'enrichissement par groupement de concepts permet d'accélérer et assurer l'efficacité et la vérifiabilité du processus d'enrichissement. Ceci est justifié par le fait que l'emplacement des concepts est connu au préalable. En effet, la

topologie des concepts au sein du cluster dans l'ontologie source reste la même dans l'ontologie noyau.

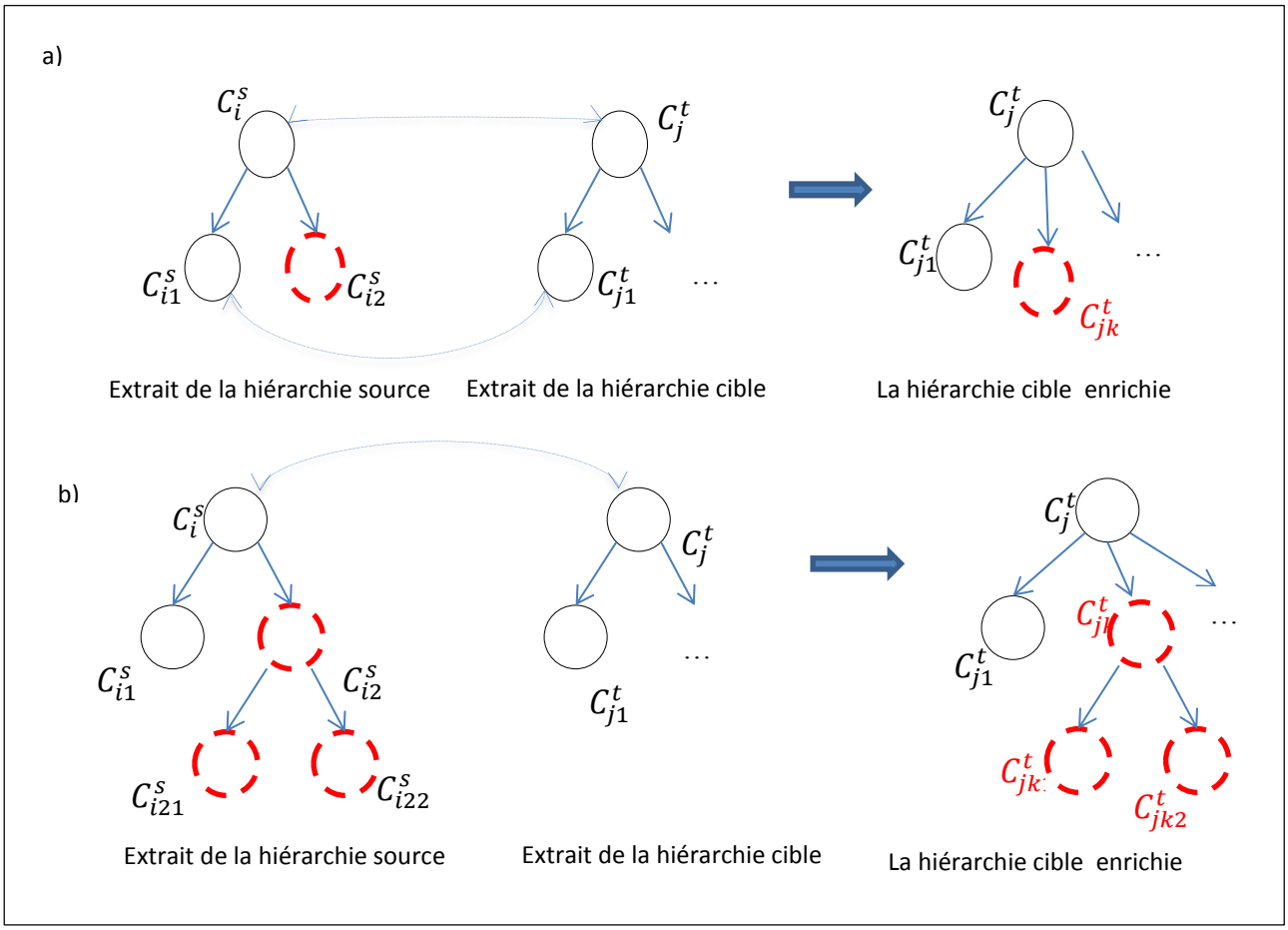

Figure III.18: Scénarios d'enrichissement d'ontologie de référence

Pour ce faire, il est requis de déterminer l'emplacement (dans l'ontologie noyau) du concept le plus subsumant dans le cluster source C_i^s. En d'autres termes, il s'agit de rechercher une relation de subsomption ou de proximité dans le cluster père C_j^t . Une fois le concept sémantiquement correspondant est déterminé, la sous hiérarchie des concepts dans C_i^s est rajoutée directement dans l'ontologie noyau.

3.3.2 Enrichissement de concepts

Cette étape permet d'ajouter les alignements jugés valides (Filtrés et reconnus par l'expert) dans l'ontologie noyau. L'enrichissement a pour but de rajouter de nouvelles connaissances qui n'existaient pas dans l'ontologie d'intérêt. De ce fait, d'ajouter les alignements jugés valides (Filtrés et reconnus par l'expert) dans l'ontologie noyau. Les concepts de l'ontologie source figurant dans les alignements produits (subsomptions, de proximité) sont considérés comme étant les concepts candidats dans le processus d'enrichissement.

En effet, les alignements d'équivalence ne vont pas permettre d'introduire de nouveaux concepts dans la cible. En revanche, ceux faisant intervenir les relations de subsomption (isA ou isMoreGnl) sont susceptibles de conduire à de nombreux enrichissements.

Ainsi, un concept c_s provenant de la source figurant dans un alignement de type $<c_s$ isA $c_t>$ est considéré, comme un nouveau concept intégrable à la cible et directement placé comme une spécialisation de c_t dans celle-ci. Pour les relations de proximité, un concept c_s provenant de la source figurant dans un alignement de type $<c_s$ NEAR $c_t>$ est considéré, comme un nouveau concept directement placé comme NEAR c_t dans l'ontologie noyau. L'enrichissement des concepts de l'ontologie source qui n'ont pas été mis en correspondance, est réalisé par l'expert, où la détermination de l'emplacement de ces nouvelles connaissances est réalisée manuellement.

3.4 Concept incrémental

3.4.1 Mise à jour de la hiérarchie

Dans l'éventualité de la réutilisation des structures hiérarchiques pour des applications ultérieures, il convient de pouvoir tenir compte les concepts rajoutés dans l'ontologie cible. En effet, cet enrichissement peut avoir une incidence sur des medoides prédéfinis des clusters ou la répartition initiale des concepts par rapport aux clusters prédéfinis de la hiérarchie. De ce fait, une étape de mise à jour de la hiérarchie s'impose. L'idée consiste à mettre à jour le plus grand cluster (du premier niveau) c_i^t subsumant le cluster c_{ij}^t (enrichie par le nouveau concept). La démarche que nous adoptons est basée sur le même algorithme présenté dans la section 3.1. La mise à jour de la hiérarchie peut créer de nouveaux clusters avec une répartition distincte des concepts ainsi qu'une nouvelle estimation des médoïdes, où on peut garder les mêmes clusters prédéfinis antérieurement. Dans les deux cas, le re-calcul de ces paramètres est indispensable afin de maintenir la cohérence et la cohésion de la structure hiérarchique.

L'enrichissement de l'ontologie par de nouveaux clusters peut, également, affecter la répartition des concepts par rapport aux clusters prédéfinis de la hiérarchie. De ce fait une étape de mise à jour de la hiérarchie s'impose. L'idée consiste à mettre à jour le plus grand cluster (du premier niveau) c_i^t subsumant le cluster c_{ij}^t (enrichie dans l'ontologie). La démarche que nous adoptons est basée sur le même algorithme présenté dans la section 3.1. Une fois que les nouveaux éléments ont été rajoutés, l'algorithme de clustering est appliqué itérativement depuis le premier niveau avec initiation basée sur les anciens medoides (pour la convergence rapide de l'algorithme). Néanmoins, dans l'algorithme présenté, l'estimation du nombre de clusters c par défaut, est égale à 2 avec l'enrichissement ; le nombre c s'incrémente pour le cluster père subsumant le nouveau.

3.4.2 Confrontation d'une « nouvelle connaissance » et l'ontologie globale

A l'issue de l'étape précédente, nous disposons d'une ontologie globale comportant les connaissances issues de la fusion de différentes bases de connaissances liées à la mammographie. Un contexte particulier est le rajout d'une nouvelle connaissance qui peut

être une nouvelle tumeur, un nouveau diagnostic, une nouvelle forme radiologique, etc. Dans ce cas, on procède, à réaliser une confrontation, ou une comparaison, entre la nouvelle connaissance et la base des connaissances. Cette confrontation consiste à décider l'emplacement adéquat du nouveau concept ou la nouvelle connaissance. L'objectif étant, bien évidement de déterminer le degré de compatibilité ou de ressemblance entre le concept cible et la base actuelle. Ceci, en utilisant le même formalisme que celui utilisé dans le processus d'enrichissement qui a pour objectif de comparer les degrés de ressemblance entre le nouveau concept et les clusters prédéfinis dans l'ontologie.

La démarche que nous proposons peut être résumée de la manière suivante. Dans un premier temps, le nouveau concept sera confronté aux différents medoides de la hiérarchie à travers le calcul de la similarité sémantique selon l'algorithme proposé dans la phase d'ancrage (section 3.2.1) du processus d'alignement. Une fois que le cluster le plus proche est retenu, dans un second temps, on procède au calcul de la mesure sémantique par rapport aux éléments du cluster correspondant. Le type de la relation avec le concept le plus proche est déterminé par l'expert.

3.5 Avantages de l'approche proposée

Contrairement aux approches du clustering 'rigide' induisant une représentation rigide des concepts [Hamdi, 2012] [Wei, et al., 2008], l'utilisation de l'approche floue permet d'augmenter la chance de trouver les alignements appropriés. En effet, la notion floue des clusters permet d'augmenter l'espace d'alignement tout en maintenant la consistance sémantique des concepts au sein du même cluster, ce qui permet d'améliorer la qualité d'alignement généré (Figure III.19).

L'utilisation de l'incertitude est due au fait qu'un concept dispose de différents attributs et propriétés permettant de l'assigner à différents clusters concurremment. Par exemple, pour la classification de la mammographie 'BI-RADS_3', ce concept indique la présence d'une anomalie éventuellement bénigne mais qui révèle une probabilité de malignité. De ce fait, ce concept peut être assigné simultanément aux clusters 'malignité' et 'bénignité' avec différents degrés d'appartenance.

La méthode que nous proposons se caractérise, également par la hiérarchisation des clusters dont l'utilité est démontrée dans l'étape de recherche des clusters similaires. Au lieu de comparer tous les blocs provenant des ontologies source et cible paire à paire, cette recherche devient plus restreinte en propageant la comparaison d'un niveau à un autre dans la hiérarchie sur des clusters spécifiques.

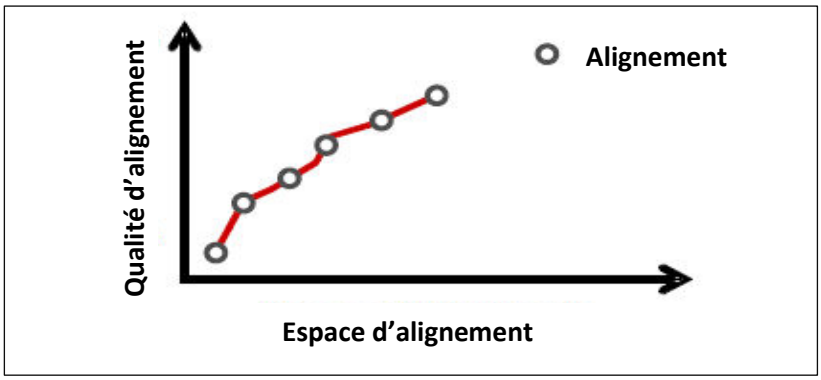

Figure III.19: Variation de la qualité d'alignement en fonction de l'espace de recherche de correspondances [Thayasivam & Doshi, 2014]

4. Processus d'enrichissement relationnel basé sur la fouille des connaissances des règles d'association

Après avoir enrichi l'ontologie cible par de nouveaux concepts provenant des bases de connaissances existantes, il est indispensable de rajouter les associations adéquates entre les concepts prédéfinis. Réaliser ces mises à jour manuellement demeure une tâche coûteuse et fastidieuse puisqu'elle mobilise un ou plusieurs experts du domaine pour identifier les concepts ainsi que leurs corrélations. A cette fin, on propose d'exploiter les bases de données liées au domaine d'étude afin d'extraire les relations possibles entre les entités pertinentes du domaine.

La richesse du contenu informatif des bases de données des patients (atteints ou non atteints du cancer de sein) permet de découvrir de nouvelles connaissances aux experts du domaine. Bien que, beaucoup de travaux [Mahmoodi, et al., 2016] [Gim, et al., 2015] [Dou, et al., 2015] [Nebot, et al., 2012] se basent sur l'utilisation des ontologies pour extraire les règles ou les motifs les plus pertinentes, il s'avère à l'inverse que peu de travaux visant à mettre à jour l'ontologie exploitent les techniques de fouilles de données.

Dans cette section, nous présentons, une contribution portant sur l'enrichissement relationnel de l'ontologie. Notre approche est fondée sur une idée directrice : transposer le modèle des règles d'associations utilisé en fouille de données, afin de découvrir des relations entre les concepts.

Ces règles d'association, sont dérivées, à partir des informations décrivant les expériences passées, reliant les items ou les valeurs d'attributs fréquents du domaine permettant, ainsi, de découvrir, sans connaissances préalables, des tendances implicatives fréquents sous la forme de motifs séquentiels.

Néanmoins, cet avantage d'une découverte non supervisée par les RAs est fortement dépendant de deux limitations majeures : La quantité prohibitive ainsi que la crédibilité des règles découvertes ce qui rend la tâche d'analyse et de vérification beaucoup plus compliquée. En effet, les différents tests montrent que les règles deviennent presque impossibles à utiliser

dès que leur nombre dépasse 100 [Marinica, 2010]. De plus, les connaissances apportées par ces règles d'association sont souvent connues par l'expert et/ou modélisées dans les bases de connaissances.

Vu que l'objectif de l'utilisation des RAs dans ce travail est bien précis (l'enrichissement relationnel de l'ontologie), l'évaluation des RAs (2-items) générées est étroitement liée aux connaissances prédéfinies dans l'ontologie de domaine afin d'en extraire de nouvelles. Cela exige la mise en correspondance des deux types de connaissances. D'autre part, l'acceptation de nouvelles connaissances est étroitement conditionnée par l'approbation de l'expert de domaine. A cette fin, une mesure de qualité des RAs également basée sur l'ontologie, est proposée afin de les classer selon l'ordre de leurs importances. Ce classement facilite l'analyse et la vérification par les décideurs en leur permettant de se concentrer sur les règles les plus intéressantes.

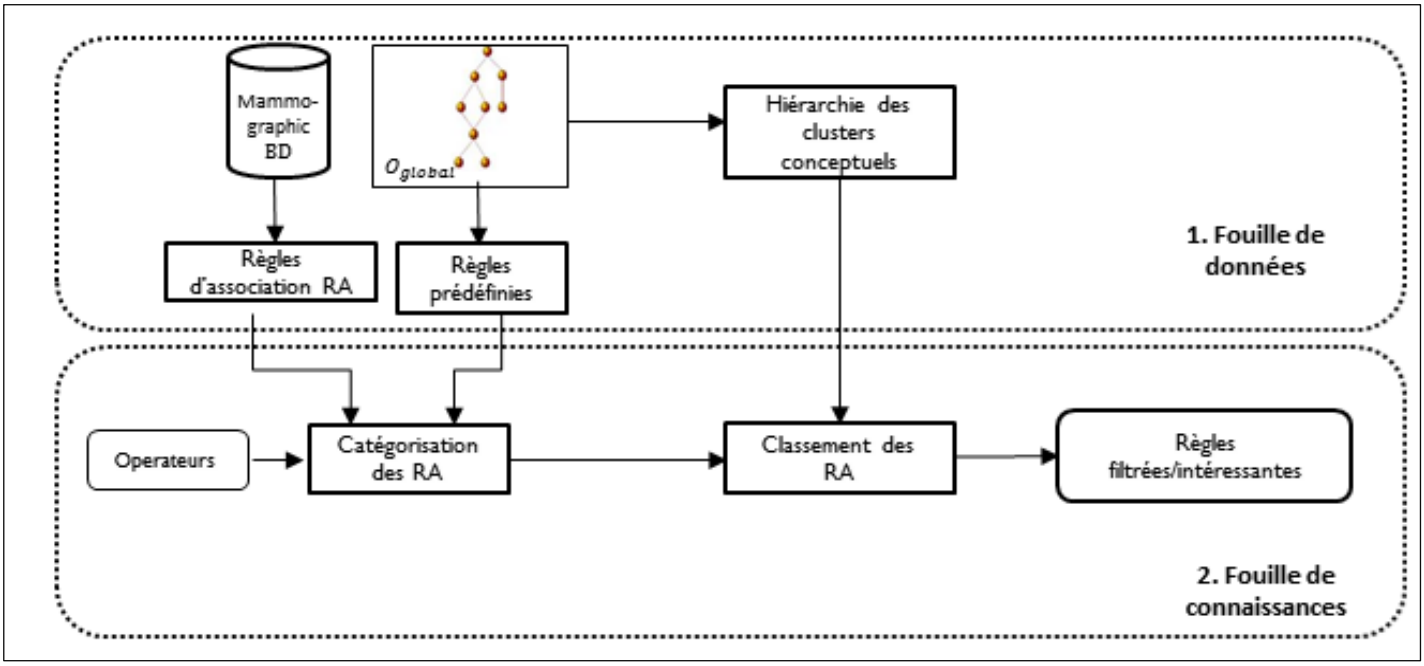

Figure III.20: Approche proposée pour l'extraction des règles nouvelles et intéressantes

Les principales contributions de ce travail (basé sur la fouille des connaissances des RAs) sont comme suit :

-La classification des RAs en des 'RAs connues' et 'nouvelles RAs' en se référant à l'ontologie qui représente un modèle sémantique de haut niveau. Cela dit que les connaissances de cette ontologie seront définie selon un formalisme proche de celui des RAs. Enfin, un ensemble d'opérateurs de traitement sont appliqués sur les deux sources de connaissances pour générer la classification.

-Une nouvelle mesure de qualité est également introduite. En effet, l'expert évalue l'importance d'une règle en fonction de la distance qui sépare ses termes. Plus les concepts introduits dans la RA sont sémantiquement distants, plus la règle est importante pour

l'expert. A cette fin, nous proposons une nouvelle distance sémantique basée sur la nouvelle réorganisation de l'ontologie proposée précédemment. Ceci dit, plus les concepts sont éloignés dans la hiérarchie, plus la règle les comportant est importante.

La Figure III.20 présente notre approche pour la fouille des connaissances des RA. L'approche se base sur deux modules. A partir de la base de données de domaine, le premier module consiste à extraire les RAs qui constituent l'entrée du deuxième module. Ce dernier permet de fouiller les implications générées, premièrement, en les catégorisant selon leur originalité, ensuite en les classant selon une nouvelle mesure de qualité. Les résultats de ce module seront introduits aux experts afin de valider les règles qui serviront pour l'enrichissement de l'ontologie. La fouille des RAs est basée principalement sur l'utilisation sémantique d'une ontologie du domaine.

5.1 Fouille de données : Extraction des règles d'association

Ces règles se présentent sous la forme $R: a \rightarrow b$; où a et b désignent des ensembles d'items disjoints; la règle R traduit que b est vrai quand a est vrai. .

Ainsi, le problème revient à extraire les règles les plus fiables en utilisant les deux critères cités ci-haut. Ceci est réalisé par l'algorithme d'extraction classique Apriori qui permet de filtrer les règles dont le support et la confiance sont respectivement supérieures aux seuils Min_Sup et Min_Conf prédéfinis par l'utilisateur, ces seuils sont connus par les contraintes d'extraction de règles. Le choix de l'algorithme Apriori est justifié par sa popularité et ses performances.

5.2 Fouille des connaissances des RAs

Une étude de l'état de l'art a été introduite dans le chapitre (II), sur les mesures de qualité proposées dans la littérature. On a d'une part, les mesures de qualité objective qui permettent l'évaluation des règles générées de point de vue des données (statistique et/ou descriptive). L'avantage de ces mesures, c'est leur autonomie (c.à.d. elles ne nécessitent pas l'intervention de l'utilisateur). Néanmoins, l'importance des règles varie d'un utilisateur à un autre. D'autre part, les mesures subjectives se basent sur la sélection des règles correspondantes aux attentes de l'utilisateur.

Dans le contexte de ce travail, l'objectif principal c'est de fouiller les règles portant de nouvelles connaissances par rapport aux connaissances de base (qui sont déjà modélisées au sein de l'ontologie). A cette fin, nous proposons, en premier lieu, de catégoriser les connaissances découvertes dans les RAs en deux catégories : des connaissances connues et des nouvelles connaissances. Cette catégorisation permet de filtrer les règles qui ne sont pas d'une importance dans le processus d'enrichissement. En deuxième lieu, les RAs portant de nouvelles connaissances sont classées selon leur importance afin de simplifier la tâche de vérification par l'expert.

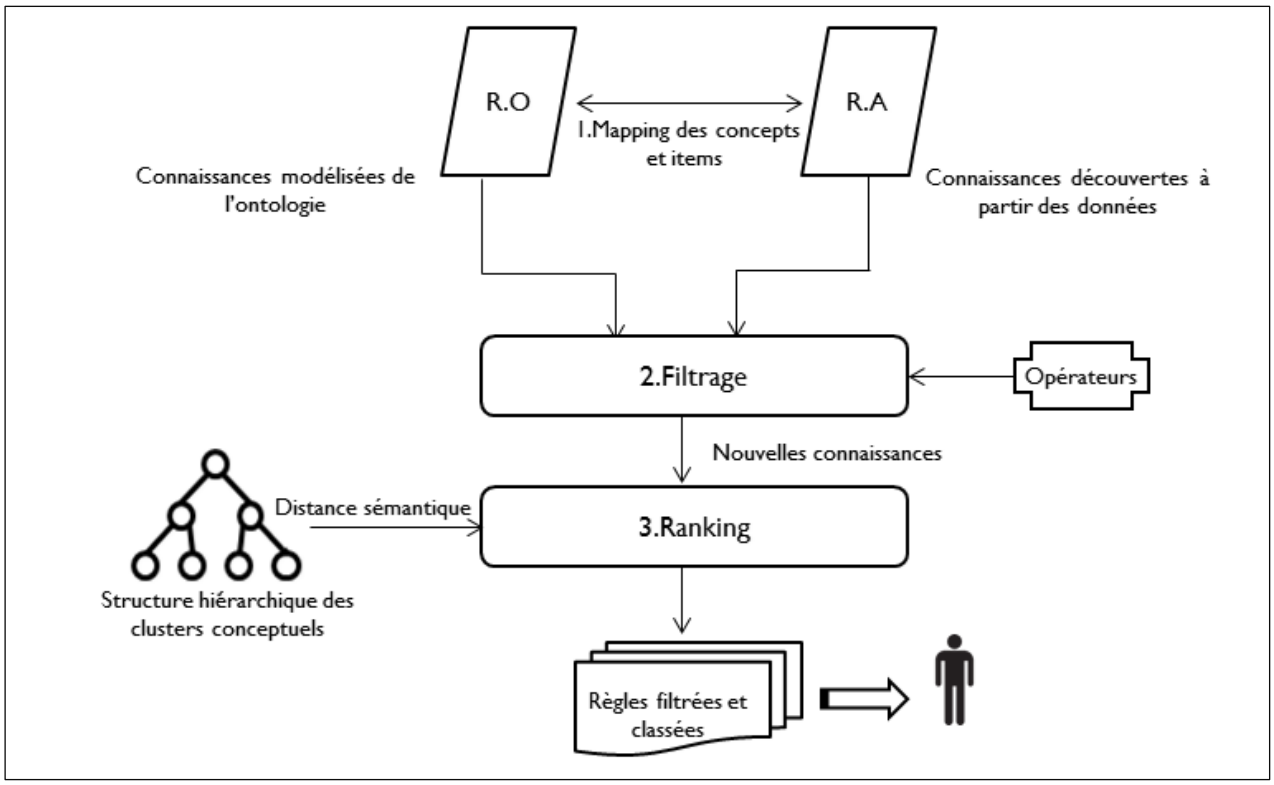

Figure III.21: Fouille de connaissances des RAs en se basant sur l'ontologie de domaine

La Figure III.21 présente l'approche proposée pour le post-traitement des R.A générées.

5.2.1 'Mapping' des concepts de l'ontologie et les items des RAs

Afin de pouvoir comparer les deux sources de connaissances en jeux (ontologie et RA), les items des RAs et les concepts ontologiques doivent être sémantiquement connectés [Idoudi, et al., 2016]. Rappelons que les concepts définis dans une ontologie peuvent être :

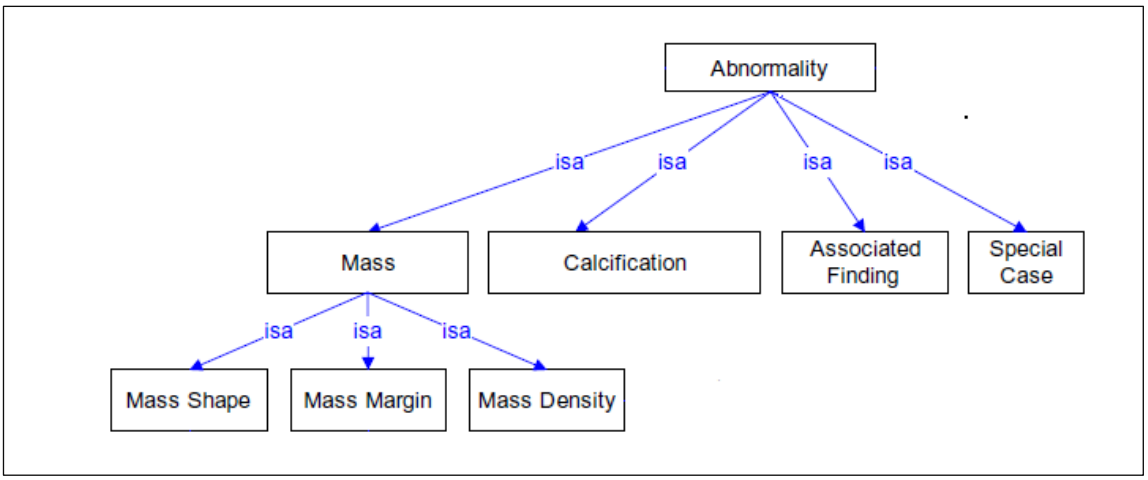

Figure III.22: Extrait d'une ontologie mammographique

- *Concepts généralisés :* Concepts disposants des relations 'is-a', par exemple, $Abnormality, Mass$ (voir Figure III.22)
- *Concepts feuilles* : Concepts sans descendants dans la taxonomie, par exemple, $Mass\ Shape, Mass\ Margin, Mass\ Density$ (voir Figure III.22)

De ce fait, le mapping est assuré comme suit : chaque concept C peut être connecté à un ou plusieurs items I de la base de données avec deux types de connexions:

- Une connexion directe $\Omega_{\text{direct}}: I \rightarrow C$

$$\forall\ i \in I,\ \Omega_{\text{direct}}(i) = c_i \mid i \equiv c_i$$

- Une connexion indirecte $\Omega_{\text{indirect}} I \rightarrow C$

$$\forall\ i \in I, \Omega_{\text{indirect}}(i) = \cup_i \{\ c_i\ (subsumant\ (c_0))\ \mid i \equiv c_0\}$$

La connexion directe assure la mise en correspondance entre les entités syntaxiquement équivalentes (les concepts feuilles et généralisés). Par exemple, soit $Mass\ Shape$, un item dans la base de données :

$$\Omega_{\text{direct}}(Mass\_shape) = \{Mass\_shape\}$$

La connexion indirecte permet d'établir une correspondance entre un item et les concepts généralisés du concept qui lui sont identiques. Ce type de connexion permet d'identifier les règles implicitement équivalentes.

$$\Omega_{\text{indirect}}(Mass\_shape) = \{Mass, Abnormality\}$$

Par conséquent, le mapping des concepts de l'ontologie avec les items de la base de données est défini comme suit [Idoudi, et al., 2016]:

$$\Omega = \Omega_{\text{direct}} \cup \Omega_{\text{indirect}}$$

5.2.2 Filtrage des règles d'association

Afin de pouvoir comparer les deux sources de connaissances en jeux (ontologie et RAs), les prédicats définis au sein de l'ontologie sont transformés en une représentation similaire à celles des RAs. Ainsi, les implications modélisées dans l'ontologie sont converties en des Règles Ontologiques RO comme suit :

$$RO: (A \rightarrow B)$$

Avec A, B représentent deux concepts de l'ontologie, '$\rightarrow$' représente une relation de type object-property.

Afin de pourvoir catégoriser les RAs, nous proposons l'utilisation des opérateurs définis par [Bing, et al., 1999], où l'objectif était de comparer les RAs avec les attentes de l'utilisateur. Un opérateur est définit comme étant l' action appliquée sur l'ensemble des RAs et ROs pour déterminer si une RA est conforme ou non à une RO. A cette fin, nous proposons Dans ce travail, on se limite à l'utilisation de deux opérateurs qui sont définis comme suit :

- *Conformité* CF: R_i et R_j sont conformes si leurs antécédents ainsi que leurs conséquences sont respectivement équivalents.

- *Non-conformité* NCF: Une R_i est sélectionnée si elle n'est pas conforme aux R_j au niveau de la partie antécédent et conséquence.

L'utilisation de ces opérateurs, dans ce contexte, est définie comme suit :

Appliqué sur les ensembles des RAs, l'opérateur de conformité CF détermine les RAs dont l'antécédent et la conclusion sont équivalents à ceux d'une RO. Soit :

$$RA_i : X \rightarrow Y \ , RO_j : A \rightarrow B$$

On a RA_i est conforme à RO_j; $CF(RA_i) = RO_j$, si : $\{A, B\} \subset \{\Omega(X) \cup \Omega(Y)\}$.

On a $NCF(RA_i) = RO_j$, si $\{A, B\} \not\subset \{\Omega(X) \cup \Omega(Y)\}$

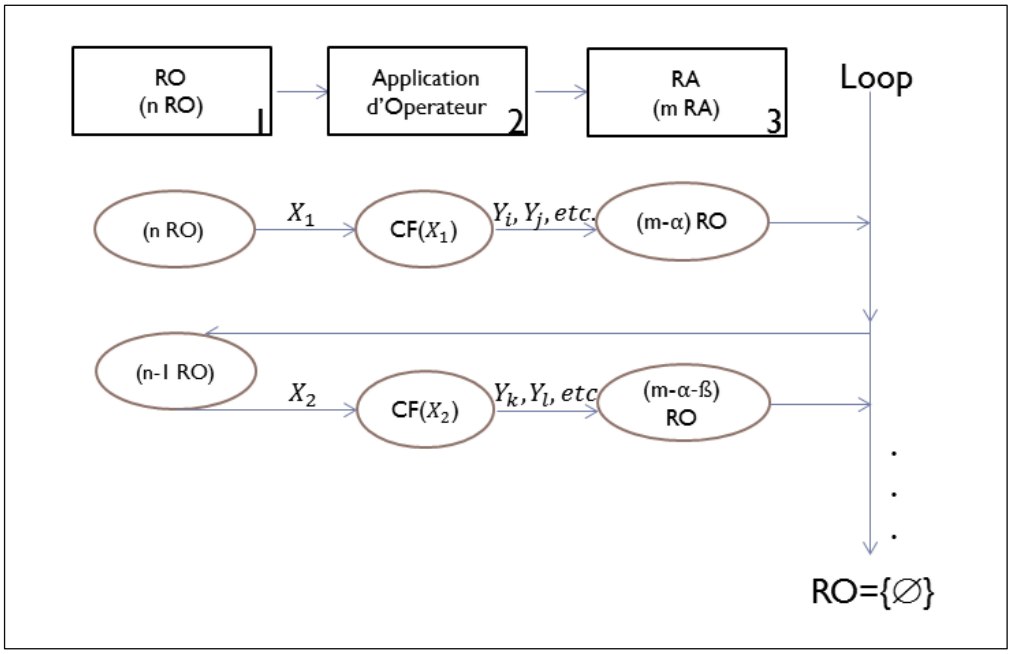

Figure III.23: Processus de filtrages des RAs

Par conséquence, les opérateurs appliqués permettent de diviser l'ensemble des RAs en deux catégories : les règles connues déterminées suite à l'application de l'opérateur CF et les règles portant de nouvelles connaissances déterminées suite à l'application de l'opérateur NFC.

Etant d'un grand intérêt, la deuxième catégorie des RAs sera introduite à l'expert afin de pouvoir valider ces connaissances pour l'enrichissement de l'ontologie. La Figure III.23 illustre le processus itératif proposé pour le filtrage des RAs. A chaque itération, une RO est sélectionnée, et l'opérateur de conformité est appliqué pour sélectionner les RAs qui lui sont similaires. Ces dernières sont ensuite supprimées de la liste des RAs. La liste finale des RAs représente les règles portant de nouvelles connaissances. Par ailleurs, nous proposons de classer ces règles en termes d'importance afin d'aider l'expert à se concentrer sur les règles les plus importantes.

5.2.3 Une nouvelle mesure d'intérêt pour les RAs

Le classement/ ranking des RAs permet de guider l'expert vers les connaissances potentiellement intéressantes dans les grands volumes de règles produites par les algorithmes de fouille de règles. A cette fin, les mesures de qualité des règles sont introduites. Ces indicateurs numériques permettent d'ordonner les règles des plus importantes au moins importantes.

Dans la littérature, plusieurs mesures d'intérêt ont été proposées. La plupart se base sur la structure des données en jeux pour évaluer la qualité statique de la règle sans tenir compte des connaissances de domaine. Afin de palier à ce problème, nous proposons dans cette section une nouvelle mesure des règles basée sur un support sémantique ' l'ontologie'. En effet, la mesure que nous introduisons, implique la similarité des concepts en utilisant la structure de l'ontologie. Selon l'expert de domaine mammographique, une règle est d'autant plus importante lorsque les items sont sémantiquement distants ou éloignés telles que les règles définissant des associations entre données cliniques/ radiologiques et la classification correspondante de la mammographie.

Dans [Razan, et al., 2014], les auteurs ont proposé une mesure d'intérêt basée sur une ontologie pour encoder l'utilité d'une règle afin de pouvoir sélectionner et classer les règles en fonction de leur importance. La mesure proposée s'appuie sur le calcul sémantique de la similarité entre deux items i_1, i_2 en se basant sur l'emplacement des concepts (qui leur correspondent) ainsi que leur ancêtre commun dans la hiérarchie ontologique. Formellement, la similarité est définie ainsi (Equation 3.14):

$$\boldsymbol{SemSim(i_1, i_2) = \frac{Dist(LCA(i_1,i_2),Root)}{Dist(i_1,i_2)+Dist(LCA(i_1,i_2),Root)}} \quad \textbf{(III.14)}$$

$LCA(i_1, i_2)$ désigne le petit ancêtre commun entre (i_1, i_2) ; $Dist(LCA(i_1, i_2), Root)$ désigne le nombre d'arcs séparant $LCA(i_1, i_2)$ et la racine ; $Dist(i_1, i_2)$ est la distance séparant i_1, i_2

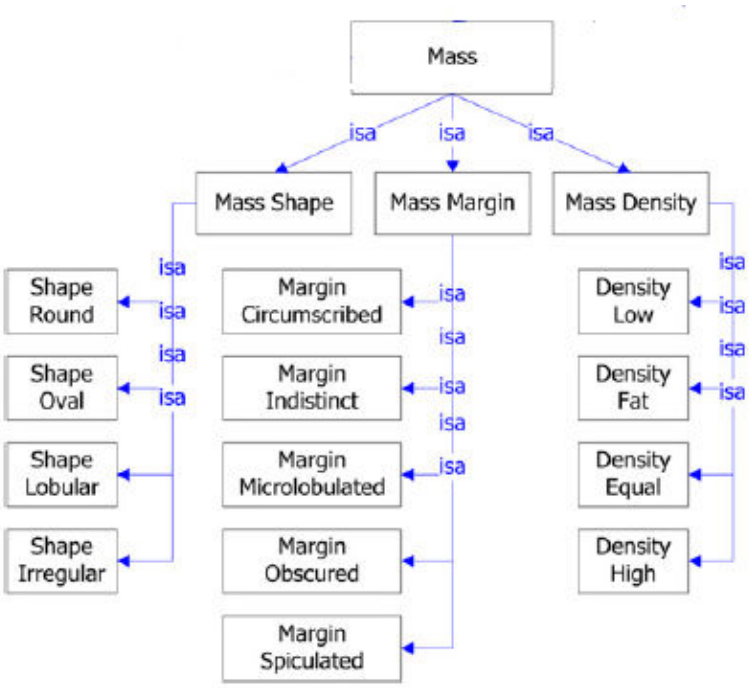

Figure III.24: Extrait d'une ontologie mammographique [Bulu, et al., 2012]

Par exemple, dans la Figure III.24, l'application de la similarité sémantique proposée dans (Equation 3.14) conclut que la distance entre les concepts 'Shape Round' et 'Shape Oval' est différente de celle entre 'Shape Round' et 'Margin spiculated'. Néanmoins, l'expert interprète ces deux paires de concepts comme étant sémantiquement proches puisque leur fonction est unique c'est de pouvoir décrire une forme conceptuelle.

Afin de palier à ce problème, nous proposons d'utiliser la structure hiérarchique des clusters conceptuels proposée dans la section 3.1. Comme on a déjà expliqué, la réorganisation proposée organise les clusters regroupant les concepts sémantiquement proches, dans une structure hiérarchique. Dans cette section, on n'exploite pas la notion floue des clusters afin d'éviter toute forme d'ambiguïté, de ce fait, les concepts sont assignés aux clusters auxquels les degrés d'appartenance sont les plus élevés.

Il convient de signaler que la distance proposée n'est pas seulement applicables aux types de règles sur lesquelles on travaille (disposant de deux items). Elle peut être également utilisée pour les règles multi-items.

La distance sémantique proposée est inspirée de celle de Wu&Palmer [Wu, et al., 1994] qui se base sur la notion des arcs entre les entités à comparer ainsi que leur plus bas ancêtre. Ainsi, on définit la mesure d'intérêt d'une règle $R_2: i_1 \rightarrow i_2$ par la distance entre ses items. Cette dernière basée sur la structure hiérarchique des clusters conceptuels est définit comme suit:

$$\mathbf{Interest}(R_2) = \boldsymbol{SemDist}(\boldsymbol{i_1}, \boldsymbol{i_2})$$

Avec $SemDist(i_1, i_2)$ est la distance sémantique entre les items i_1, i_2 définit par l'équation (III.18).

$$\boldsymbol{SemDist(i_1, i_2) = 1 - \frac{2*D(LCA(C(i_1),C(i_2)),Root)}{D(C(i_1),Root)+D(C(i_2),Root)}} \quad \textbf{(III.15)}$$

Avec :

- $LCC(C(i_1), C(i_2))$ désigne le plus petit cluster commun des clusters correspondants aux i_1, i_2 ;
- $D(LCC(C(i_1), C(i_2)), Root)$ désigne la distance séparant
- $LCC(C(i_1), C(i_2))$ et la racine ; $Dist(C(i_1), Root)$ est la distance séparant le cluster de i_1 et la racine ;
- $Dist(C(i_2), Root)$ est la distance séparant le cluster de i_2 et la racine.
- Le facteur 2 est utilisé pour la normalisation de la fonction.
- Si i_1, i_2 appartiennent au même cluster, leur similarité sémantique vaut 1.

Soit une règle multi-items R_m: $\{a_1, a_2..a_n\}$, la mesure d'intérêt de cette règle est définie comme suit :

$$\text{Interest } (\boldsymbol{R_m}) = \frac{\sum_{1\leq i,j\leq n;\, i\neq j} \mathbf{SemDist}(a_i,a_j)}{\sum_{k=1}^{n-1} k} \quad \textbf{(III.16)}$$

5.2.4 *Ranking* de règles d'association

Une fois la mesure d'intérêt est calculée pour chaque règle dans l'ensemble de RAs, les règles dont la mesure d'intérêt est bien supérieure à une valeur donné (fixé par l'expert) peuvent être transmis à la décision de domaine. Une autre façon consiste à générer les RAs top-k qui sont appréciées par la mesure proposée.

Ces règles classées, doivent être introduites vers l'expert qui, lui seul, peut juger de la pertinence de ces nouvelles connaissances. En effet, les modèles découverts peuvent révéler de nombreuses fausses règles, trompeuses, inintéressantes et insignifiantes. Ce problème se pose lorsque certaines RAs se produisent suite à une pure coïncidence résultant d'un certain caractère aléatoire dans le jeu de données en cours d'analyse.

5.3 Enrichissement relationnel

Les RAs : $X \rightarrow Y$ déterminées tentent de rapprocher, d'une manière certaine, les concepts candidats X, Y dans l'ontologie à travers l'association $\rightarrow$ sans nommer les relations qui (ne sont pas étiquetées). Notre contribution est fondée sur le déploiement de ces modèles de connaissances afin d'enrichir une ontologie de domaine existante et aboutir à un réseau conceptuel.

A cette fin, on introduit formellement des rôles/ labels aux relations d'association. Ceci dépend des concepts (ou items) auxquelles les concepts formant l'antécédent et la conséquence appartiennent. Par exemple, si une règle définit une implication entre deux concepts désignant des formes conceptuels tels que *mass*, *calcification*, *opacity*, etc. alors, nous

affectons le label '*occur_with* 'à la relation qui leur associe, la relation *associated_with()* associe un concept de type forme conceptuel et un concept de type classification mammographique. Ces nouvelles relations sont ainsi rajoutées afin de préserver la cohérence des relations préétablies.

6. Conclusion

Dans ce chapitre, nous avons proposé de suivre un processus de développement d'ontologie basé sur l'enrichissement d'une ontologie noyau par des connaissances issues des ressources ontologiques existantes ainsi que les bases de données liées au domaine d'étude.

Deux principales contributions ont été présentées dans ce chapitre. Au cours de la première contribution, nous avons proposé une nouvelle approche de fouille de connaissances ontologiques qui résulte en une structure hiérarchique des clusters conceptuels. Nous avons montré le rôle de l'approche pour la réduction de la complexité de la tache de recherche de correspondances entre les deux ontologies en jeux. Nous avons, aussi, montré une nouvelle méthode d'alignement. L'originalité de notre approche est qu'elle exploite les clusters des concepts sémantiquement proches pour mener la comparaison sur des clusters spécifiques de la hiérarchie source à travers la propagation de la similarité d'un niveau à un autre. Les nouvelles connaissances filtrées et validées sont ensuite utilisées pour l'enrichissement de la base de connaissances globale. Une deuxième originalité de notre approche est qu'elle permet de découvrir les nouvelles connaissances par groupement de concepts.

Dans la deuxième partie de ce chapitre, on s'est intéressé à l'exploitation des connaissances extraites à partir des bases de données grâce aux algorithmes d'ECD. Au cours de la deuxième contribution, nous avons montré le rôle de la fouille de connaissances dans la phase de post-traitement des RAs. Cette dernière a été explorée dans le but d'extraire de nouvelles connaissances filtrées et classées pour l'enrichissement de la base de connaissances globale. A cette fin, nous avons adapté l'approche proposée aux connaissances du domaine modélisées au sein de l'ontologie. L'étape de post-traitement est basée principalement sur le filtrage et le ranking des RAs. Nous avons également montré le rôle de la structure hiérarchique des clusters conceptuels proposée dans la première partie de ce chapitre dans l'étape de ranking des RAs.

Le chapitre suivant sera consacré à l'application et à l'évaluation de l'approche proposée sur un jeu de données réel lié au domaine mammographique afin de montrer la faisabilité et l'intérêt de notre approche.

IV. Chapitre 4 : Validation et Expérimentation_ Cas d'étude : le domaine mammographique

Sommaire :

1. Introduction

Ce chapitre présente les expériences que nous avons menées pour mettre en œuvre et valider l'approche proposée dans le chapitre précèdent, notamment : les phases de réorganisation d'ontologies, l'alignement, l'enrichissement conceptuel, l'extraction /le post traitement des RAs et l'enrichissement relationnel. Ce chapitre étale les différents jeux de données que nous avons utilisés pour l'enrichissement d'une ontologie mammographique cible, ainsi que les résultats obtenus dans chaque étape (préparation d'ontologies, alignement, enrichissement conceptuel, traitement des données, extraction des règles d'association, filtrage, ranking des règles d'association, enrichissement relationnel).

2. Description du scénario d'application

Le but du scénario illustré est de réaliser une ontologie mammographique suite à l'application du processus d'enrichissement d'une modélisation ontologique proposée dans la littérature. Le processus implique deux différentes sources de connaissances intervenant chacune à un niveau particulier du processus.

Nous avons choisi de réutiliser l'ontologie MAMMO et de la considérer comme étant l'ontologie noyau (cible) vue sa richesse en termes de concept et son niveau de structuration élevé. Pour être plus intensément exploitée, cette ontologie mérite d'être enrichie par de nouveaux concepts et relations.

Cette ontologie sera enrichie aux niveaux conceptuel et relationnel :

- **L'ontologie MAMMO** [Taylor, et al., 2012] : Cette ontologie a été proposée en 2010 et développée avec le langage OWL-DL. Elle a été conçue dans le but d'être intégrée dans un système d'apprentissage dédié aux radiologues stagiaires. Le processus de construction a été manuel suite aux nombreux interviews avec les experts du domaine. Cette ontologie est caractérisée par une hiérarchie bien structurée et assez compacte [Taylor, et al., 2012]. Elle contient 692 classes reliées par des liens de subsomption et 135 propriétés. La racine comportait, principalement, quatre fils directs, eux même pères directs d'autres concepts.

Dans ce qui suit, nous présentons les scénarios d'enrichissement conceptuel et relationnel associés aux différents jeux de données liés au domaine mammographique. La Figure IV.1 illustre les modules des processus d'enrichissement proposés ainsi que leurs sous étapes respectives.

Nous avons choisi d'utiliser le langage Java[4] et l'environnement de développement Eclipse[5]. Le choix du langage de programmation est justifié par le fait qu'il permet de manipuler plusieurs APIs tels que XML parser pour le traitement des fichiers XML, Jena[6] pour le chargement des ontologies OWL, et l'API WordNet. L'API Jena fournit toutes les

[4] http://java.sun.com/
[5] http://www.eclipse.org/
[6] http ://jena.sourceforge.net

fonctionnalités nécessaires à la manipulation des ontologies, la hiérarchie des concepts ainsi que leurs propriétés.

3. Enrichissement conceptuel de l'ontologie MAMMO

Nous présentons, dans cette partie, les différents résultats relatifs au module d'enrichissement de l'ontologie noyau au niveau conceptuel, à savoir les résultats du clustering hiérarchique des ressources ontologiques employées ainsi que les différentes correspondances trouvées dans les étapes d'ancrage et de dérivation.

3.1 Jeux de données pour l'enrichissement conceptuel

Concernant le choix de la ressource qui sert à enrichir l'ontologie noyau, nous avons retenu l'ontologie suivante:

-**L'ontologie BCGO** [Adina, 2010] : Cette ontologie a été proposée dans le contexte du projet 'European Virtual Physiological Human' en 2012 et développée avec le langage OWL-DL. Elle a été conçue dans le but de l'intégrer dans un système d'extraction d'information à base de cas. La hiérarchie est compacte contenant 531 concepts et 100 propriétés. La racine comportait quatre fils directs, eux même pères directs d'autres concepts. On considère cette ontologie comme étant l'ontologie source qui sert à enrichir l'ontologie noyau au niveau conceptuel.

3.2 Résultats de pré-Alignement : fouille de connaissances des ontologies cible et source

Nous présentons pour chacune des ontologies noyau (cible et source), les résultats de leurs réorganisations en des structures des clusters conceptuels flous.

La première étape de hiérarchisation consiste à diviser les concepts en des axes conceptuels de thématique générale caractérisant le domaine mammographique. Le nombre de clusters dans ce premier niveau est unifié pour les ontologies d'applications repérant quatre grands axes conceptuels typiques : les données cliniques, les données radiologiques, les évaluations et les classifications (Figure IV.2).

Ces axes conceptuels permettent de collectionner les termes du domaine dans des groupements de concepts et de suivre ainsi une méthode de construction descendante pour l'élaboration de la structure hiérarchique.

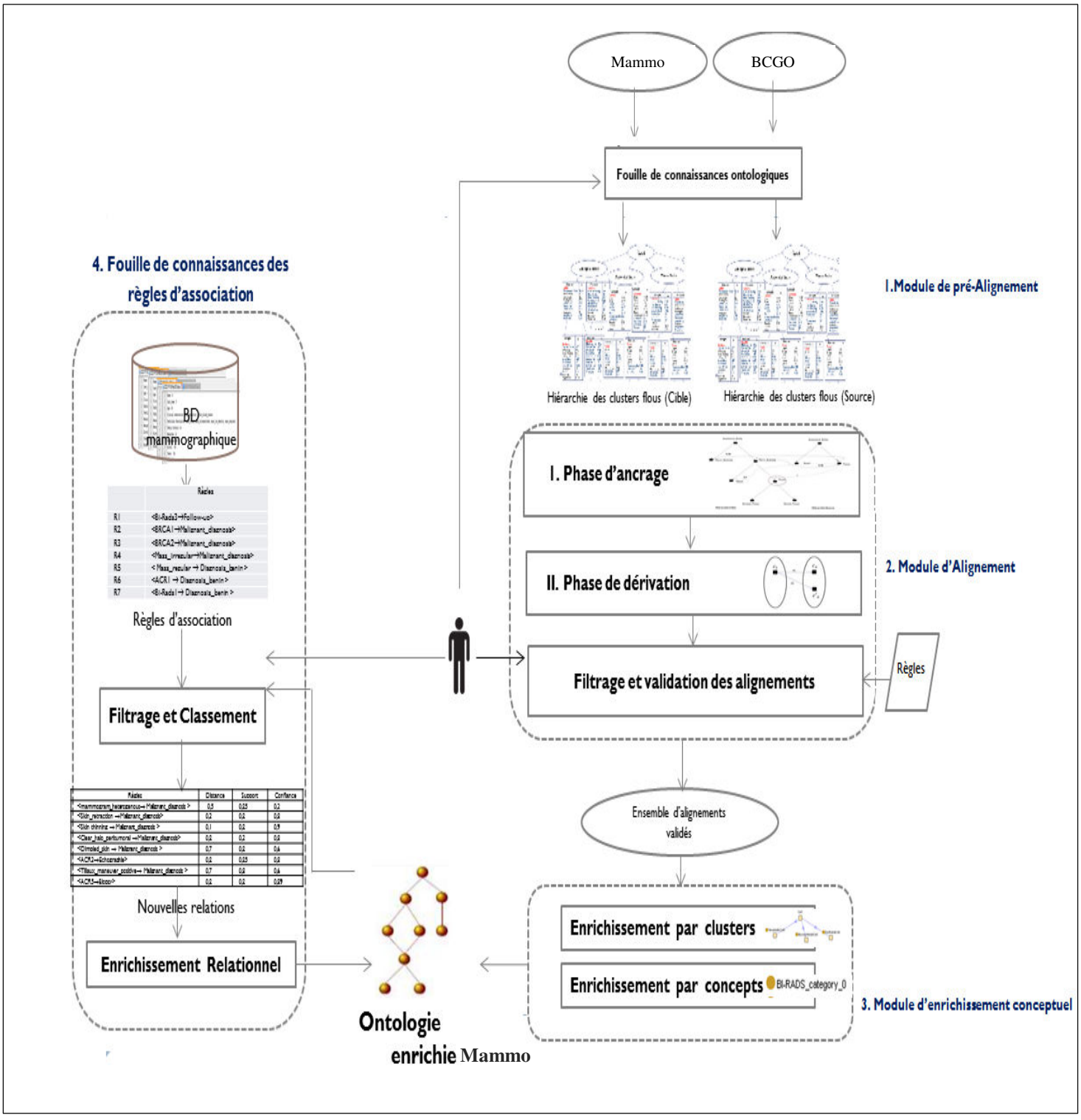

Figure IV.1: Architecture de l'approche proposée

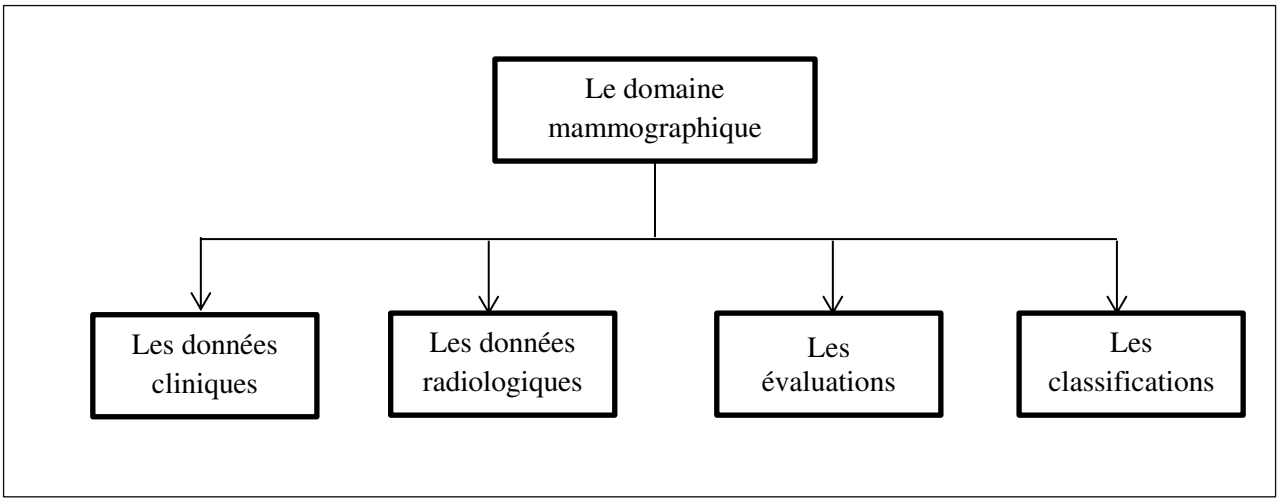

Figure IV.2: Le domaine mammographique

Dans chaque groupe conceptuel, nous allons étudier les clusters conceptuels qui ont en été générés. Les interprétations de ces classes sont présentées et expliquées dans le Tableau IV.1 :

Tableau IV.1: Les classes Top-Level des hiérarchies

Classes	**Signification**
Les observations cliniques : '**Clinical Observations**'	Cette classe représente les termes anatomiques liés au sein et ses caractéristiques : sa composition, sa géométrie, sa région, etc. Exemple : Tissue, nipple, etc.
Les observations radiologiques : '**Radiological Observations**'	Cette classe comprend les constations comprenant les formes radiologiques extraites d'une mammgraphie. Exemple : mass, micro_calcification, macro_calcification, lesion, etc.
Les évaluations : '**Assessment**'	Cette classe représente interprétations par rapport à l'existence ou non du cancer. Elles sont déduites suite aux résultats radiologiques et cliniques trouvés. Les diagnostics peuvent être bénins ou malignes. Malignant_diagnosis, diagnosis_benin.
Les classifications : '**Classification**'	Cette classe représente la classification des mammographies par rapport à la nature de tumeurs trouvées. Exemple : Bi-Rads1 Bi-Rads2, Bi-Rads3.

L'étape suivante consiste à élaborer les sous-hiérarchies conceptuelles correspondantes aux axes conceptuels présentés dans le Tableau IV.1.

Une boucle itérative est déclenchée, à chaque itération un cluster est sélectionné et re-clusterisé de nouveau si la mesure intra-cluster est supérieure ou égale à 0.7. A ce stade, seulement les concepts dont les degrés d'appartenance sont supérieurs à 0.5 constituent les objets du clustering. En effet, nous supposons que si un concept n'appartient pas vraiment à un cluster, alors il ne le sera pas pour les nouveaux clusters enfants.

```
<FuzzyCluster>
    <Id>...</Id>
    <Medoid>...</Medoid>
    <Concepts>
        <Concept>
        <Name>...</Name>
        <Membershipe_Degree>...</Membershipe_Degree>
        </Concept>
        <Concept>..</Concept>
        ...
    </Concepts>
</FuzzyCluster>
```

Figure IV.3: Représentation XML du cluster flou

A la sortie du système, un ensemble de clusters flous est généré avec estimation des medoides et des concepts associés avec des degrés d'appartenance aux clusters respectifs. Les résultats fournis en sortie sont enregistrés dans des fichiers de format XML qui servent à conserver formellement les résultats du clustering.La Figure IV.3 présente une représentation XML d'un cluster flou.

3.3.1 Résultats de fouille de connaissances de l'ontologie MAMMO

A l'issue de l'application de la méthode de clustering hiérarchique floue proposée, la structure hiérarchique finale correspondante à l'ontologie Mammo comportait 31 clusters flous. Ces derniers sont répartis sur 6 niveaux de hiérarchisation : 4 clusters au premier niveau et 17 clusters au niveau le plus bas. La Figure IV.5 présente un extrait de la représentation XML d'un cluster flou.

Ces clusters sont de différentes tailles et sont de plus en plus spécifiques et de tailles réduites en descendant dans la hiérarchie. La Figure IV.6 montre une visualisation hiérarchique des clusters obtenus à partir de l'ontologie MAMMO.

La Figure IV.4 montre un extrait de la hiérarchie globale désignant la répartition des sous clusters respectifs aux 'Anatomical_entity' sur les différents niveaux hiérarchiques. Les clusters sont représentés à travers leurs medoides.

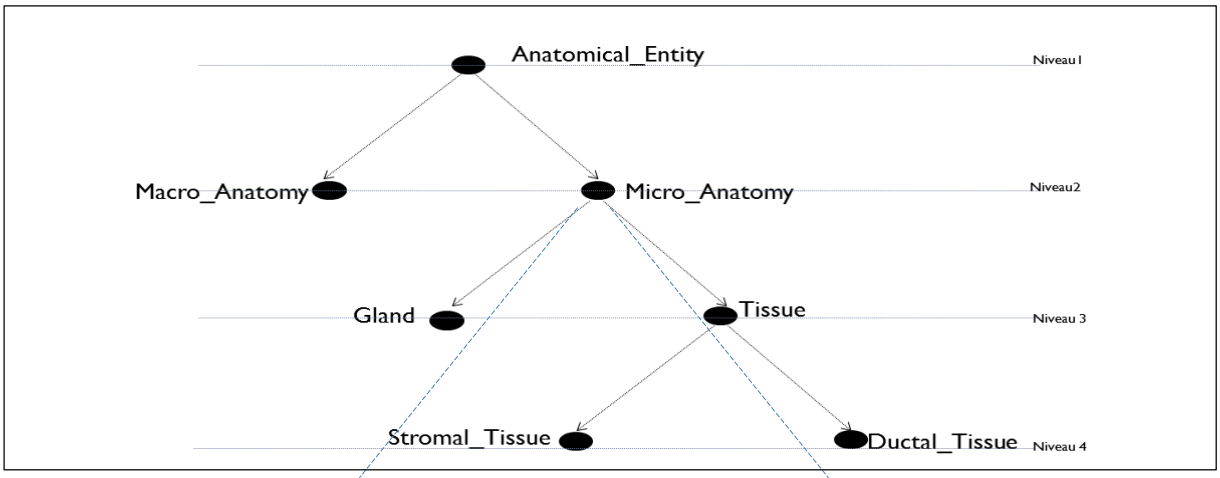

Figure IV.4: Sous-hiérarchie correspondante au cluster relatif aux entités anatomiques (Ontologie Mammo)

```
<FuzzyCluster>
    <Id>12</Id>
    <Medoid>Microanatomy</Medoid>
    <Concepts>
        <Concept>
        <Name>Gland</Name>
        <Membershipe_Degree>0.8</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Duct</Name>
        <Membershipe_Degree>0.69</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Lobule</Name>
        <Membershipe_Degree>0.69</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Tissue</Name>
        <Membershipe_Degree>0.59</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>MammaryGlandTissue</Name>
        <Membershipe_Degree>0.59</Membershipe_Degree>
        </Concept>
```

Figure IV.5: Extrait d'une représentation XML d'un cluster flou de la hiérarchie correspondante à l'ontologie MAMMO

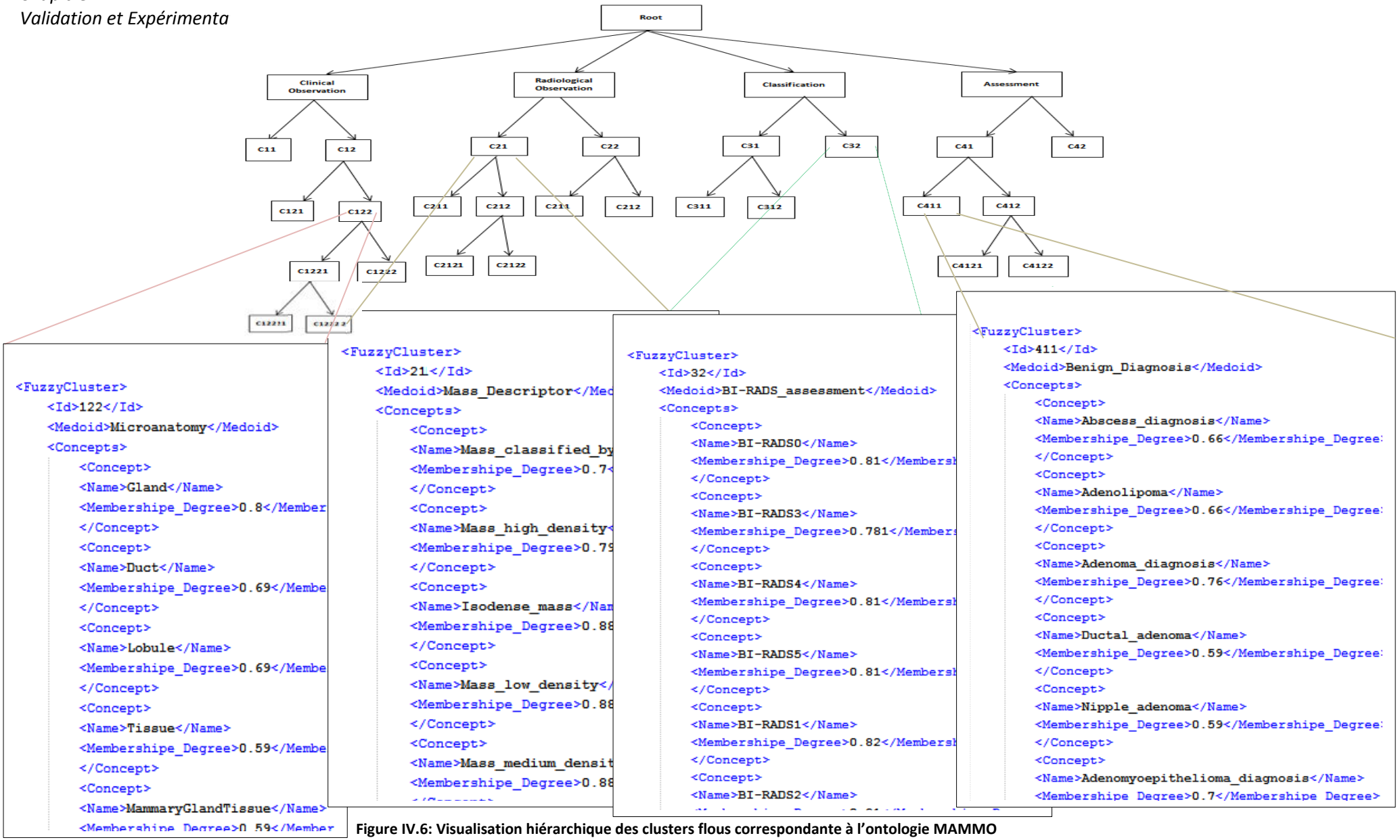

Figure IV.6: Visualisation hiérarchique des clusters flous correspondante à l'ontologie MAMMO

A ce titre, nous rappelons que les résultats que nous présentons, sont obtenus suite à l'application d'un seuil de mesure de qualité intra-cluster de 0.7. La sélection de cette mesure est réalisée suite à plusieurs expérimentations manuelles. Notons que le nombre d'itérations de l'algorithme FCMdd est sensible à l'étape d'initialisation des degrés d'appartenance et varie d'un cluster à un autre. Ce nombre diminue considérablement en descendant dans la hiérarchie, également pour le temps alloué à l'exécution. Les résultats de cette étape sont présentés dans le Tableau IV.2.

Tableau IV.2: Nombre de clusters par niveau dans l'ontologie MAMMO

Niveau	Nombre de clusters
Niveau 0	1
Niveau 1	4
Niveau 2	8
Niveau 3	13
Niveau 4	16
Niveau 5	17

3.3.2 Résultats de pré-alignement de l'ontologie BCGO

A l'issue de l'application de la méthode de clustering hiérarchique floue proposée, la structure hiérarchique finale correspondante à l'ontologie BCGO 27 clusters (Figure IV.9). Ces derniers sont répartis sur 5 niveaux. La Figure IV.7 présente un extrait de la représentation XML d'un cluster flou. Les résultats de cette étape sont présentés dans le Tableau IV.3.

Tableau IV.3: Nombre de clusters par niveau dans l'ontologie BCGO

Niveau	Nombre de clusters
Niveau 0	1
Niveau 1	4
Niveau 2	8
Niveau 3	13
Niveau 4	15

```
<FuzzyCluster>
    <Id>312</Id>
    <Medoid>Histopathological_Grading</Medoid>
    <Concepts>
        <Concept>
        <Name>Grading_Test</Name>
        <Membershipe_Degree>0.76</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Nottingham_Grading</Name>
        <Membershipe_Degree>0.76</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Grade_One</Name>
        <Membershipe_Degree>0.72</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Grade_Three</Name>
        <Membershipe_Degree>0.88</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Grade_Two</Name>
        <Membershipe_Degree>0.7</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Histopathological_Scoring</Name>
        <Membershipe_Degree>0.88</Membershipe_Degree>
        </Concept>
        <Concept>
        <Name>Criteria_Scoring</Name>
        <Membershipe_Degree>0.88</Membershipe_Degree>
        </Concept>
        <Concept>
```

Figure IV.7: Extrait d'une représentation XML d'un cluster de l'ontologie BCGO

La Figure IV.8 montre un extrait de la hiérarchie globale désignant la répartition des sous clusters respectifs au cluster de 'Anatomical_entity' sur les différents niveaux hiérarchiques. Les clusters sont représentés à travers leurs medoides.

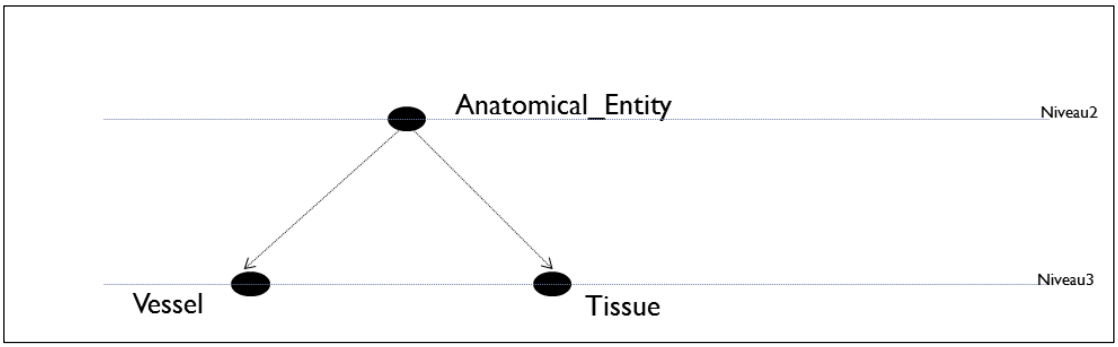

Figure IV.8: Sous-hiérarchie correspondante au cluster aux entités anatomiques (Ontologie Source)

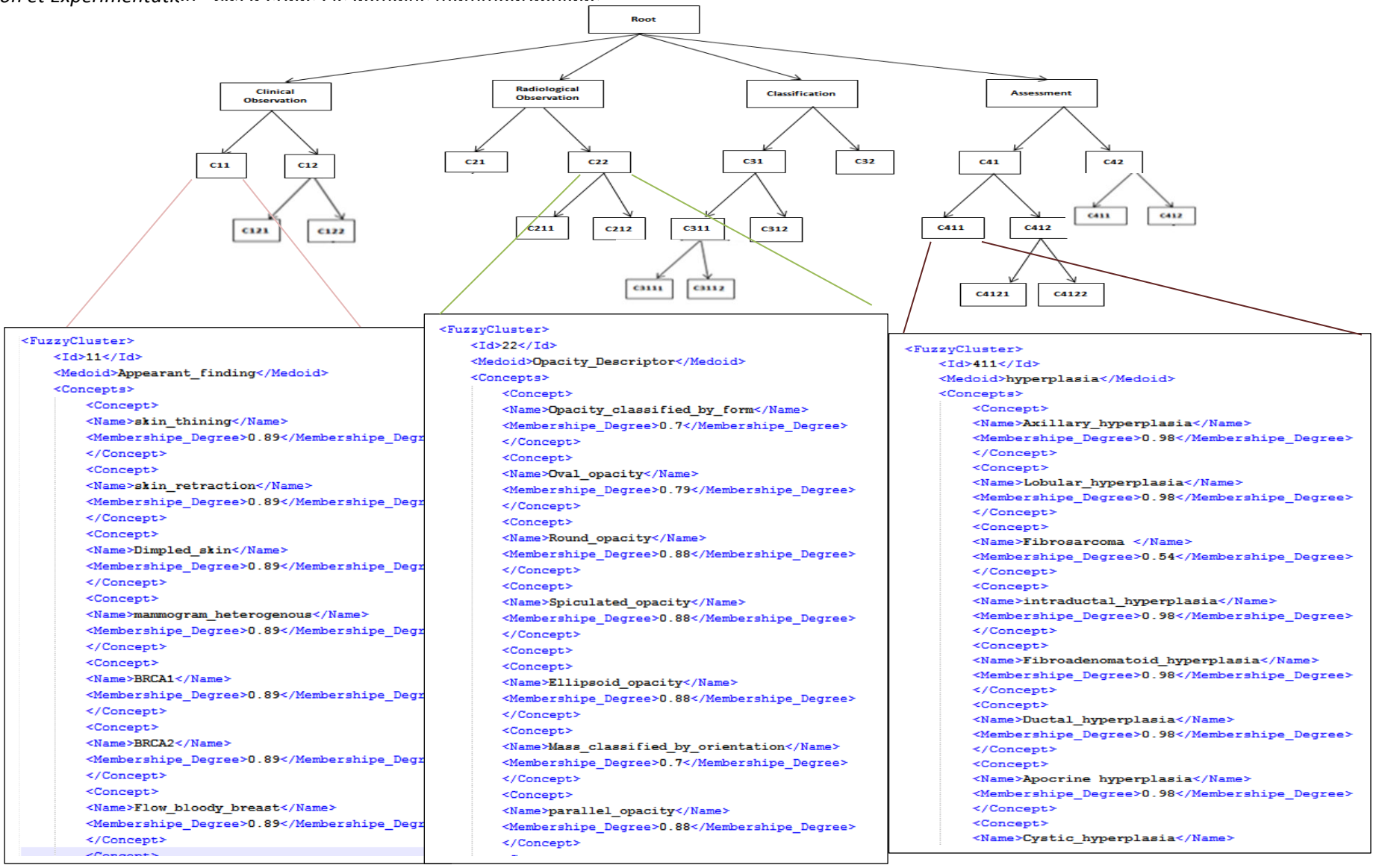

Figure IV.9: Visualisation hiérarchique des clusters flous correspondante à l'ontologie BCGO

3.4 Résultats d'alignement d'ontologies cible et source

Nous présentons ci-après un résumé sur les résultats d'alignement entre l'ontologie source BCGO et l'ontologie cible MAMMO en se basant sur leurs structures hiérarchiques des clusters générés dans la l'étape de pré-alignement.

3.4.1 Résultats de la phase d'ancrage

La phase d'ancrage est basée essentiellement sur la mesure de proximité sémantique entre les clusters présentée dans le paragraphe c du chapitre précèdent. Pour chaque cluster de la hiérarchie source, la structure hiérarchique cible est parcourue niveau par niveau, jusqu'à retenir le cluster le plus similaire. A l'issue de cette étape, on retient les couples des clusters c_i, c_j les plus proches.

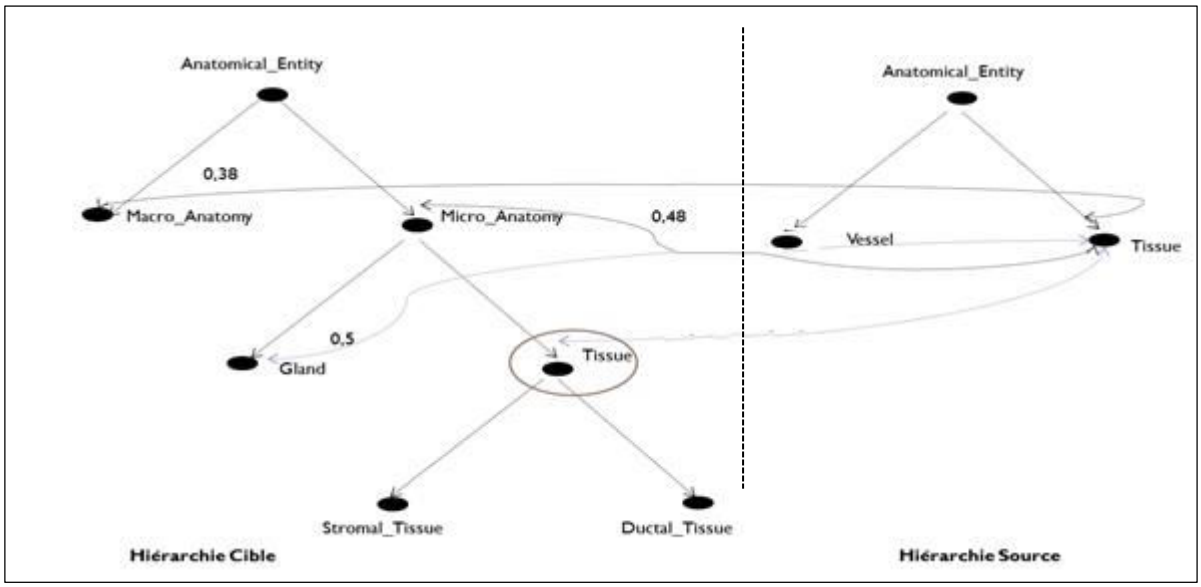

Figure IV.10: Alignement des clusters source et cible correspondants aux entités anatomiques

La Figure IV.10 illustre un exemple d'alignement du cluster source 'Tissue'. Ce type d'alignement correspond au scénario 'Multi-aligned, où ce cluster sans descendants est mis en correspondance avec le 'Micro-anatomy' du niveau 2, ensuite avec le cluster 'Tissue'. La paire de cluster à retenir dans cette étape est <Tissue, Tissue> qui sera l'entrée de l'étape de dérivation.

Nous présentons ci-après un résumé sur l'ensemble d'alignements d'équivalence sémantique trouvés entre les clusters de l'ontologie cible et celles dans l'ontologie source. Les calculs ont abouti à un nombre total de 67% des clusters de l'ontologie source qui ont été mis en correspondances avec les clusters de l'ontologie cible. Tableau IV.4 illustre les résultats des clusters mis en correspondances dans chaque niveau de l'ontologie source.

Tableau IV.4: Scénrios d'alignement des clusters de l'ontologie BCGO

	Aligné	Multi-aligned	Full aligned	Non aligned	Half aligned
Nombre de clusters	16	1	3	10	7

D'après les résultats du Tableau IV.4, 16 couples de clusters des ontologies BCGO et Mammo ont été mis en correspondance. Ces résultats prouvent la proximité élevée des ontologies du domaine mammographique. Néanmoins, 10 clusters n'ont pas été mis en correspondance. Ces clusters modélisent principalement les tumeurs, les évaluations des mammographies, et les diagnostics. Ceci s'explique par la différence de niveaux de connaissances entre les deux bases ce qui renforce le choix établi de l'ontologie source qui aide à complémenter les connaissances de la ressource noyau. Parmi les résultats d'alignement, on trouve 7 alignements de type half-aligned. Cet important nombre prouve la différence des niveaux de granularité dans les deux bases de connaissances, ce qui aide à rajouter de nouveaux concepts dans le noyau et établir un maximum de relations de subsomption.

3.4.2 Résultats de la phase de dérivation

Dans cette étape, les concepts des clusters source et cible sémantiquement les plus proches sont alignés à l'aide des techniques syntaxique, sémantique et structurelle. En d'autres termes, il s'agit de chercher la relation sémantique et intrinsèque entre les concepts ontologiques des clusters mis en correspondance en se basant sur leurs définitions au sein des ontologies. Les alignements découverts se présentent sous la forme $<e, e', Relation>$ (Figure IV.11) :

```
<Macro-biopsy, Biopsy, Subsomption>
<Intra_ lobular_cancer, intra_canal_cancer, Fraternité>
<Malignant_neoplasm, Breast_carcinoma, Fraternité>
<Adipose_Tissue, Tissue, Subsomption>
<Fibrous_Tissue, Fibrous_Tissue, Equivalence>
<Fibrous_Tissue, Tissue, Subsomption>
```

Figure IV.11: Exemple d'alignements générés

Notons qu'on dispose à ce stade de 8 couples de clusters à aligner. Les résultats finaux ont abouti à 180 relations d'équivalence, 30 relations de subsomption et 10 relations de fraternité. Ces alignements ont été validés par l'expert, dont la plupart sont jugés significatifs et cohérents, néanmoins, un certain nombre d'alignements générés sont invalides.

Tableau IV.5: Extrait des résultats d'alignement validés par l'expert

Numéro	Concept source, concept cible	Relation déduite
1	Calcic, Lipome	Equivalence
2	Micro-Calcification, Calcification	Subsomption
3	Macro-Calcification, Calcification	Subsomption
4	Lobe, Lobule	Equivalence
5	Nucleus, Mitosis	Equivalence

6	Tubule, Tubule	Equivalence
7	Lumina, Lumina	Equivalence
8	Neoplastic_cell, Cell	Subsomption
9	Micro_lobule, Lobule	Subsomption
10	Maco-lobule, Lobule	Subsomption
11	Grease, Fat	Equivalence
12	Microcyst, Cyst	Subsomption
13	Lesion, Mass	Fraternité
14	Opacity_Oval , Oval_mass	Equivalence
15	Opacity_round , Mass_Curved	Equivalence
16	Opacity_spicultaed , Mass_spiculated	Equivalence
17	Adenofibrome, Cyst	Equivalence
18	Opacity, Mass	Equivalence
19	Myoepithelial_cell, Cell	Subsomption
20	Epethelial_cell, Cell	Subsomption
21	Macro_biopsy, biopsy	Subsomption
22	Carcinoma, neoplasm	Fraternité
23	Radial_scar, surgical_scar	Fraternité
24	Dense _breast, fatty_breast	Fraternité
25	Carcinosarcoma, Chondrosarcoma	Fraternité
26	Fibrocystic, Fibroadenoma	Equivalence
27	Adenolipoma, Fibroadenoma	Fraternité
28	Adenosis, Chondroma	Fraternité
29	Adenoma, Ductal_adenoma	Subsomption
30	Adenoma, Nipple_adenoma	Subsomption

Le Tableau IV.5 présente un extrait des résultats d'alignement retenus portant sur les observations radiologiques.

On remarque, d'après les résultats obtenus, que les couples de concepts dotés d'une relation d'équivalence, sont souvent déterminés par la technique syntaxique. Ceci peut être expliqué par la non existence des termes synonymes dans le jargon lié au domaine mammographique ce qui permet de fiabiliser les techniques terminologiques pour la détermination des concepts similaires dans les domaines médicaux. Ce nombre important de concepts communs prouvent la proximité des deux ontologies en jeux. Egalement, certaines équivalences ont été déterminées par la technique structurelle, ce qui démontre la ressemblance de voisinage de ces couples de concepts. Par ailleurs, les plupart de relations de subsomption ainsi que les relations de fraternité ont été déterminées par la similarité sémantique. Ceci, montre que l'utilisation des techniques sémantiques est indispensable pour la recherche des mappings, soit pour approuver des résultats déterminés par d'autres techniques soit pour en découvrir. Les résultats de subsomption serviront pour augmenter la profondeur de l'ontologie noyau puisqu'ils amènent à l'ajout de nouveaux fils à des concepts initiaux.

3.4.3 Evaluations

Les résultats de clustering et d'alignement d'ontologies Mammo et BCGO sont évaluées en utilisant des métriques d'évaluation standards de la littérature.

3.4.3.1 Evaluation de la méthode de clustering

Dans cette section, nous nous sommes intéressés à l'évaluation de la qualité des clusters produits ainsi que la similarité sémantique proposée. A cette fin, nous avons utilisé les mesures de qualité suivantes:

- **Coefficient de Partition : PC** (indice à maximiser ↗) (Bezdek, 1981) : Cet indice mesure le 'fuzziness' du cluster. Il calcule la moyenne de la participation des entités parmi les paires de clusters flous [Zhang, et al., 2014]. Un score élevé du PC désigne une meilleure qualité de clustering :

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^{2}$$

  Les valeurs du PC varient dans l'intervalle $[\frac{1}{c}, 1]$, où c est le nombre de clusters.

- **Entropie de Partition : PE** (indice à minimiser ↘) [Bezdek, 1974]: Cette métrique mesure la distribution des entités dans les blocs et reflète la qualité de clustering. Les valeurs minimales de PE impliquent une bonne partition [Zhang, et al., 2014] :

$$PE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}[\mu_{ij}log_2\mu_{ij}]$$

  Les valeurs de PE varient dans l'intervalle $[0, log_2 c]$.

Le calcul de ces mesures est réalisé pour chaque niveau des hiérarchies des ontologies, vu que la structure produite définit bien une hiérarchie entière dans chaque niveau. On donne la moyenne des valeurs pour chaque niveau.

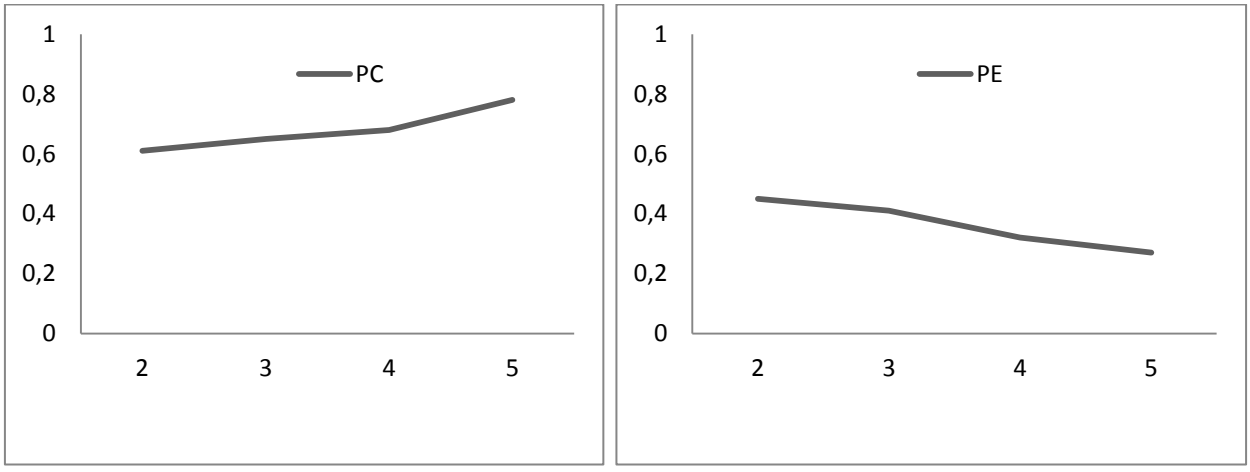

Figure IV.12: Evaluation des résultats du clustering relatifs à l'ontologie Mammo

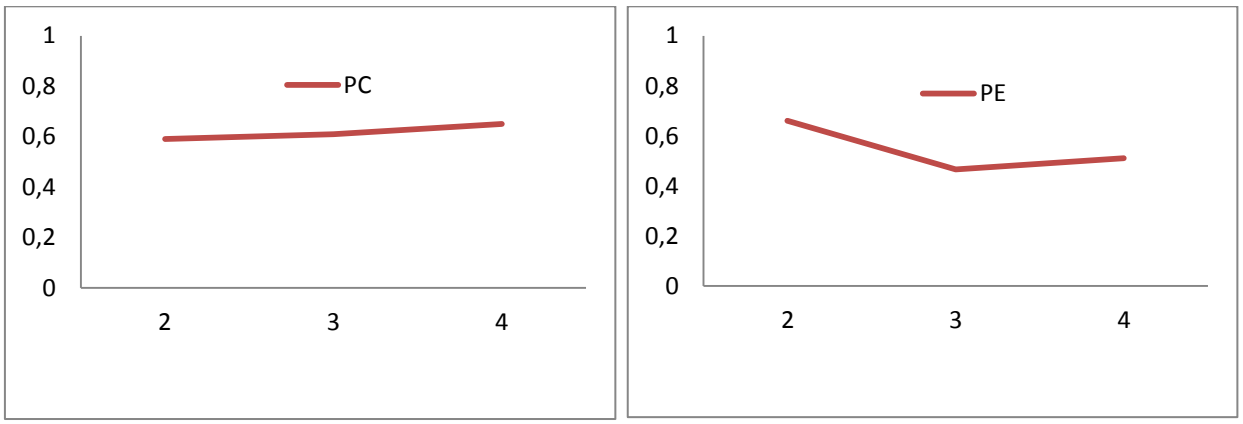

Figure IV.13: Evaluation des résultats du clustering relatifs à l'ontologie BCGO

Les résultats donnés dans les Figure IV.12 et Figure IV.13 présentent les valeurs des PC et PE dans chaque niveau hiérarchique des ontologies Mammo et BCGO. Nous remarquons qu'en descendant dans la hiérarchie, les valeurs du PC augmentent tandis que les valeurs de PE diminuent, ceci prouve l'amélioration de la cohésion des clusters hiérarchiques. Cette amélioration peut être, également expliquée par le fait que seulement les concepts disposant d'une forte appartenance aux clusters sont re-clusterisés.

Les résultats obtenus nous amènent à conclure que les deux indicateurs PC et PE peuvent être également utilisés comme un critère d'arrêt dans le clustering hiérarchique, où, le nombre de clusters produits sera choisi de manière à n'avoir aucun impact négatif sur la qualité du clustering.

Nous pouvons remarquer également que les valeurs du PE et du PC pour l'ontologie MAMMO sont légèrement meilleures que celles de l'ontologie BCGO. Ceci peut être expliqué par le niveau de structuration meilleure de l'ontologie Mammo par rapport à l'ontologie BCGO. Ceci est dû au fait que, l'ontologie MAMMO comportait un nombre important de relations de subsomption. Partant de ce constat, on peut conclure que la tâche du clustering de l'ontologie dépend étroitement de l'architecture/structure et de la qualité de l'ontologie en question.

3.4.3.2 Evaluation de la distance sémantique

Nous nous sommes intéressés également à comparer l'efficacité de la distance sémantique proposée par rapport à la distance de Wu&Palmer [Wu, et al., 1994]. Cette dernière a été utilisée dans de nombreux travaux de clustering orienté alignement, tels que [Kachroudi, et al., 2013] et [Wei, et al., 2008]. A cette fin, nous avons substitué la distance proposée par la distance basée sur la similarité de Wu&Palmer.

$$Distance_{WuPalmer}(c_1, c_2) = 1 - \frac{depth(LCA(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

Afin de pouvoir mener cette comparaison, nous avons appliqué l'algorithme FCMdd (non hiérarchique) en utilisant les deux distances en question (la distance proposé et celle de Wu&Palmer). Dans cette expérimentation, nous avons utilisé l'ontologie Mammo. Le nombre de clusters a été fixé empiriquement à c=10. La Figure IV.14 résume l'évaluation des clusters obtenus en utilisant les métriques introduites ci hauts : PC et PE.

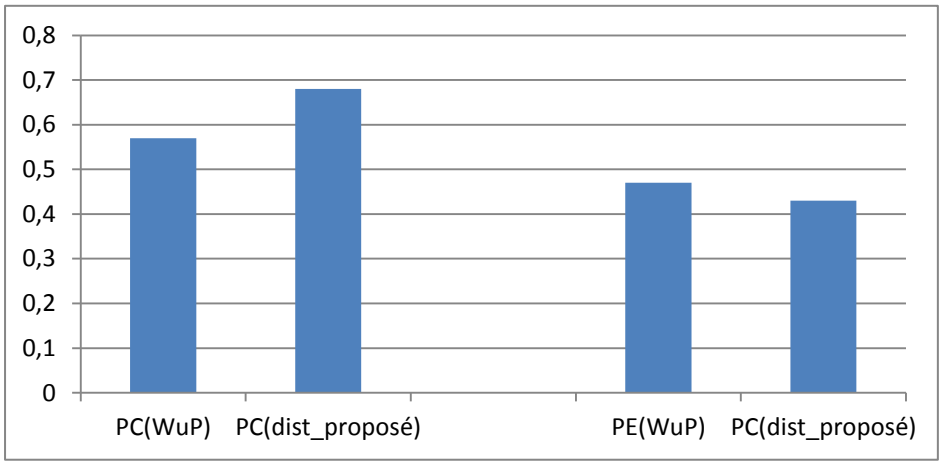

Figure IV.14: Comparaison des distances sémantiques

La Figure IV.14 montre que la méthode proposée a montré constamment de meilleures performances. Par ailleurs, il est intéressant de remarquer qu'avec l'utilisation de la distance structurelle (Wu&Palmer), les concepts de faible profondeur dans la hiérarchie ont tendance à avoir une faible appartenance aux différents clusters. Nous avons également constaté que les medoids sont associés, dans la plupart des cas, aux concepts les plus profonds de la taxonomie. Néanmoins, cela peut conduire à des medoids insignifiants et non représentatifs des clusters. En outre, en ne considérant que l'information portant sur le plus petit ancêtre pour la mesure de la dissemblance entre les concepts, ceci peut éliminer une grande quantité de connaissances explicites.

3.4.3.3 Evaluation de la méthode d'alignement : Phase de dérivation

Pour évaluer la qualité de l'alignement, nous procédons à la comparaison de notre méthode avec des méthodes existantes en utilisant la mesure de précision. Cette mesure représente la proportion de vrais positifs (parmi les éléments mis en correspondance) trouvés par la méthode. Elle permet d'évaluer la pertinence de la méthode proposée. Les valeurs proches de 1 indiquent une meilleure qualité d'alignement. Cette mesure est définie comme suit:

$$\text{Precision} = \frac{|M \cap R|}{|M|}$$

Où :

- M : désigne l'ensemble des correspondances entre les entités ontologiques découvertes par la méthode.

- R : est l'ensemble des correspondances de référence.

Nous n'avons pas calculé le rappel (la proportion de vrais positifs parmi tous les éléments correspondant à l'alignement de référence) car nous ne disposons pas des alignements de référence. Nous avons comparé les performances de notre méthode d'alignement d'ontologies en termes de précision avec deux systèmes existants qui sont open source : FALCON-AO et S-match:

Tableau IV.6: Comparaison des méthodes d'alignement

	Nombre d'alignements trouvés	**Précision**
Falcon-AO	209	0.80
S-Match	138	0.55
Méthode proposée	230	0.84

Le Tableau IV.6 présente les résultats d'évaluation de notre algorithme par rapport aux systèmes Falcon-AO (cette méthode adopte la technique de clustering) et S-Match (cette méthode n'utilise pas le clustering) en utilisant l'indice de précision sur le couple d'ontologie BCGO et MAMMO.

Ces valeurs indiquent que les performances de notre méthode et celles de Falcon-AO sont proches ce qui prouve l'avantage de l'utilisation de la technique de clustering dans l'alignement. Néanmoins, nous observons que pour le système de Falcon, certains alignements ont été perdus à cause de clustering 'rigide'. Tandis que l'utilisation des clusters flous provoque une amélioration de la valeur de la précision. Les adaptations portant sur la hiérarchisation, la mesure de similarité contextuelle et la proximité sémantique améliorent les résultats générées.

Notons aussi que l'algorithme adopté ne parcourt pas tous les clusters de deux ontologies pour la recherche des clusters similaires, la technique de hiérarchisation permet de propager la comparaison sur des clusters spécifiques contrairement au Falcon-AO.

Pour le système S-match, on remarque qu'il est le moins performant par rapport aux autres systèmes ceci s'explique par la complexité de l'espace de recherche qui permet de découvrir moins d'alignements valides. Par exemple, 40% des alignements par rapport à notre méthode n'ont pas été trouvés.

Le problème des alignements non trouvés par notre application peut être dû en particulier à la différence de structuration entre les ontologies, ce qui influence les résultats du clustering des concepts et par la suite sur la qualité d'alignement produits. En outre, le problème de mauvais alignements peut être également expliqué par la stratégie de bi-clustering appliquée dans la version actuelle de notre application. Cette limitation peut être palliée par une approche de détection automatique du nombre de clusters à chaque niveau.

3.3 Résultats d'enrichissement conceptuel de l'ontologie cible

L'enrichissement de l'ontologie source est réalisé en deux étapes: Enrichissement par clusters et Enrichissement par concepts. Le premier type d'enrichissement concerne les clusters cibles qui n'ont pas été alignés avec les clusters sources dans la phase d'ancrage, ceci signifie que ces derniers comportent des connaissances différentes de celles contenues dans les clusters cibles. Ainsi, ces clusters sont rajoutés dans la hiérarchie de l'ontologie source selon le niveau correspondant. Les résultats déduits de cette étape sont affichés dans le Tableau IV.7.

Tableau IV.7: Métriques de la hiérarchie cible avant et après enrichissement

	Avant Enrichissement	**Après Enrichissement**
Niveau 0	1	1
Niveau 1	4	4
Niveau 2	8	10
Niveau 3	13	17
Niveau 4	16	22
Niveau 5	17	23

D'après ce tableau, on remarque que la structure hiérarchique source a été enrichie en plusieurs niveaux augmentant ainsi la taille de la structure hiérarchique cible. Ceci permet de mieux conduire le niveau de détail assuré par l'ontologie de base.

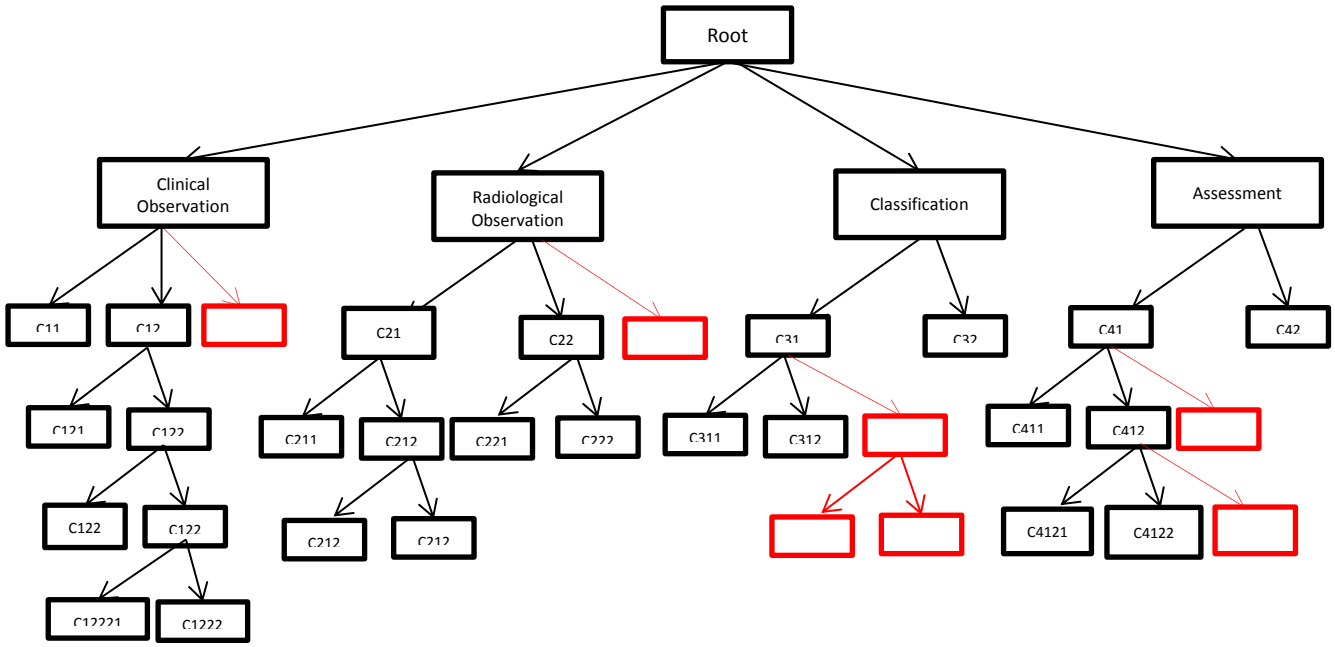

Figure IV.15: Enrichissement de la hiérarchie cible par de nouveaux clusters

Parmi ces nouveaux clusters, la Figure IV.16, montre un exemple de cluster dont les concepts ont été rajoutés dans l'ontologie cible.

Nous citons d'autres clusters qui ont servies pour enrichissement tels que les classifications histologiques, les entités macro anatomiques, les entités micro anatomiques, etc. Le cluster portant sur classifications histologiques permet de regrouper les cellules cancéreuses du sein. Ce cluster est placé dans la hiérarchie source comme descendant du cluster 'Assessement'. Par conséquence, il est ajouté comme un cluster fils de la classe 'Assessement'. Le cluster 'Histopathological_classification' disposait lui-même de deux clusters descendants dans l'ontologie source à savoir, 'Histopathological_Grading' et 'Histopathological_Scoring' (voir Figure IV.15). Parmi les clusters qui ont été le plus enrichis, nous trouvons ceux liés aux diagnostics et évaluations dans l'ontologie cible. En effet, le développement de ces deux familles est indispensable pour assurer l'exhaustivité d'une base de connaissances représentant le domaine mammographique.

Suite au calcul de la similarité globale des éléments alignés, la phase de dérivation nous a apporté 60 nouveaux concepts repartis sur différents clusters ce qui permet d'augmenter la profondeur de l'ontologie par l'ajout de nouveaux fils aux concepts d'origine. Parmi ces concepts, nous citons les concepts *micro-calcification* et *macro-calcification* comme étant des concepts fils du concept *Calcification*, les concepts *Myoepithelial_cell, Neoplastic_cell* et *Epethelial_cell* fils du concept *Cell (voir* Figure IV.17*)*. Ces concepts ajoutés aux nombreux niveaux hiérarchiques de l'ontologie, augmentent, ainsi, la profondeur de l'ontologie résultante.

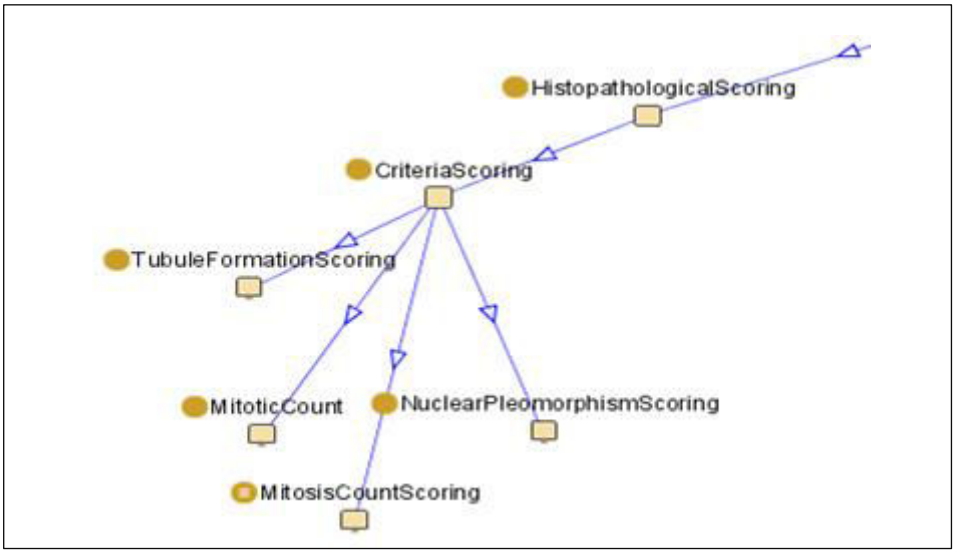

Figure IV.16: Exemple des clusters conceptuel source 'Histopathological_Scoring' rajouté dans l'ontologie cible

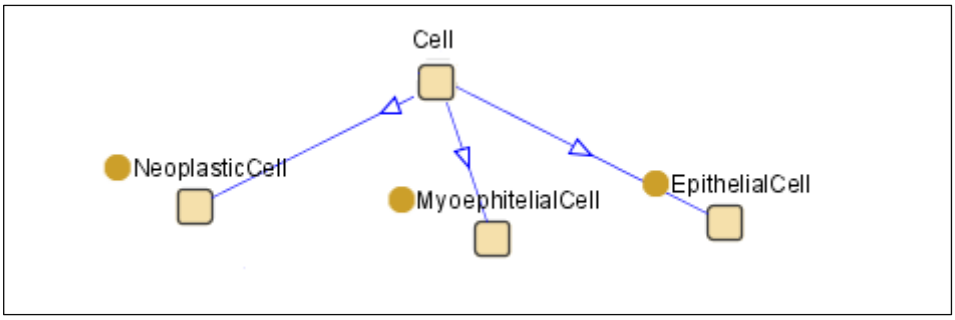

Figure IV.17: Rajout des concepts 'Neoplastic_cell', 'Myoepithelial_cell' et 'Epethelial_cell' fils du concept 'Cell'

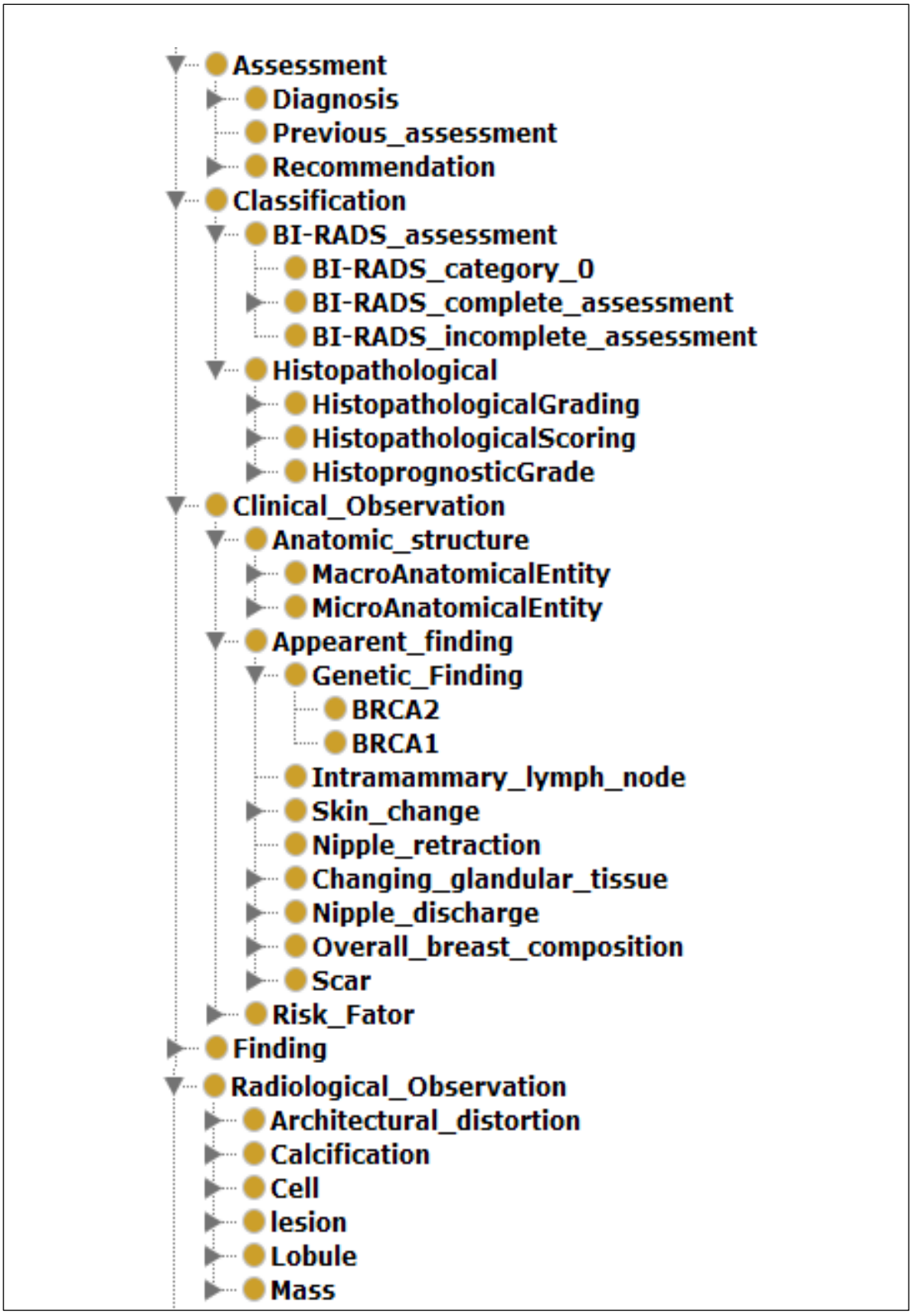

Figure IV.18: Extrait de l'ontologie cible après enrichissement conceptuel

Suite à l'étape de placement de groupements de concepts (enrichissement par clusters) ainsi que des concepts individuels (enrichissement par concepts), l'enrichissement conceptuel de l'ontologie cible a généré une augmentation du nombre initial de concepts de 110 concepts

sur tous les niveaux hiérarchiques de l'ontologie. La Figure IV.18 montre un extrait de la nouvelle ontologie enrichie.

4. Enrichissement Relationnel de l'ontologie Mammo

Cette partie porte notamment sur l'enrichissement relationnel de l'ontologie mammographique avec des RAs portant sur:

-La coexistence des observations radiologiques/ cliniques (facteurs de risque) et la classification ou le diagnostic approprié.

-La coexistence des observations radiologiques.
-La coexistence des observations cliniques.
-La coexistence des observations radiologiques et cliniques.

Ceci en suivant étape par étape la méthodologie présentée dans le Chapitre 3 (section 3.2). Nous soulignons ici l'importance de la participation de l'expert du domaine tout au long du processus afin de mieux saisir l'intérêt des règles extraites, permettant ainsi de tirer les meilleures relations des concepts dans le domaine.

Les résultats de notre étude, présentés dans cette section, pourront être améliorés en cas d'augmentation de la base de données considérée, ou en cas de certaines précisions apportées par les experts du domaine. La Figure IV.19 illustre les étapes du processus adopté.

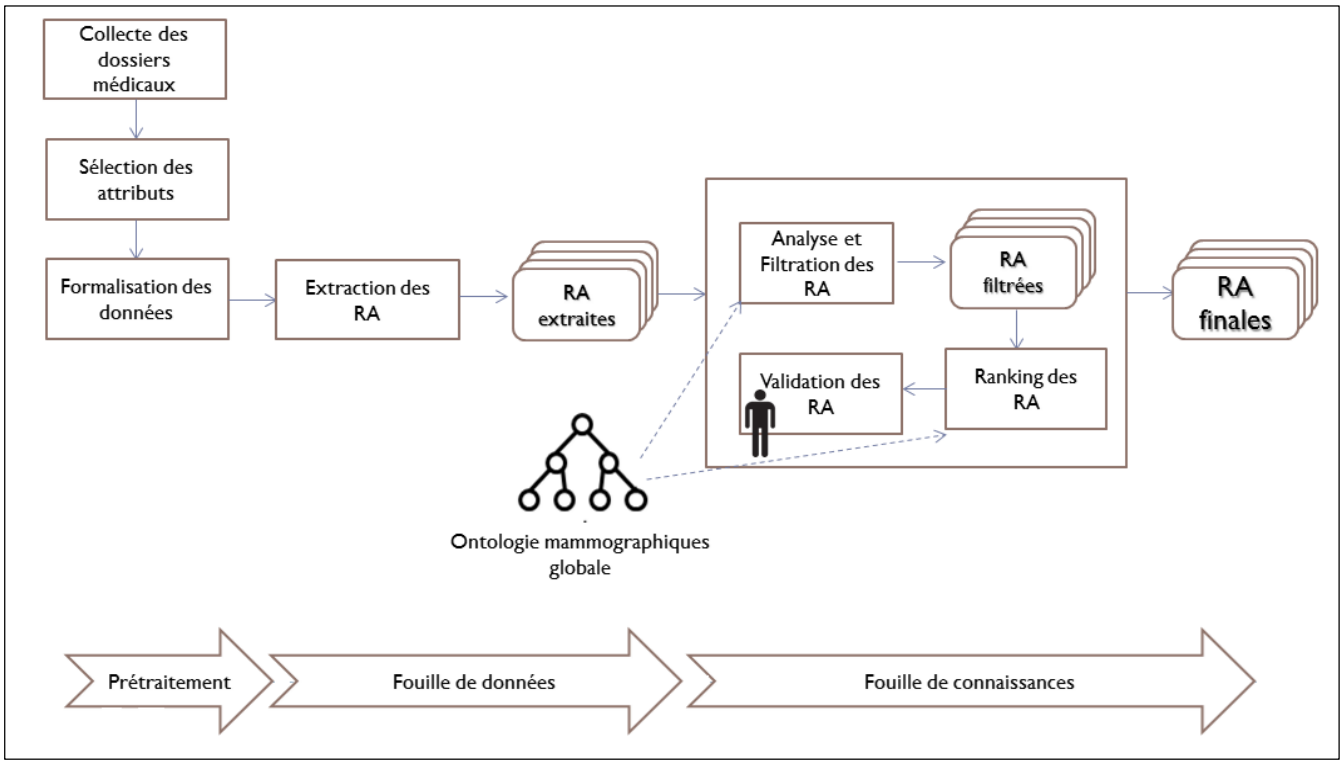

Figure IV.19: Processus d'extraction des règles pour l'enrichissement relationnel de l'ontologie

4.1 Jeux de donnés pour l'enrichissement relationnel

Le processus proposé commence, dans un premier temps par l'étape de prétraitement des données en jeux. Cette étape dépend fortement des objectifs attendus du système afin de pouvoir sélectionner l'information (transactions, attributs ou variables) utile.

Les données utilisées dans notre cas d'étude sont les dossiers médicaux provenant des hôpitaux Ben Arous et Taher Sfar en Tunisie. Cette base comporte les données relatives aux patients atteints du cancer de sein et des patients sains. Ces dossiers ont été collectés et numérisés dans un travail de mastère réalisé au sein de laboratoire LR/SITI [Baccour, 2013]. La base collectée comportait 312 dossiers médicaux des patients atteints du cancer du sein, ou sains. Chaque dossier comporte des données textuelles décrivant les données radiologiques de la mammographie ainsi que les données cliniques relatives au patient. La Figure IV.20 montre un extrait d'un dossier médical de la base de données.

```
Name: S
Last_name: Y
Age: 60
Clinical observation: Skin_retraction , Flow_bloody_breast
Radiologic Description: mass_round, mass_circumscribed, mass_low_density, mass_regional
Family history: no
Menarche: 15
Menopause: 40
Alcool : No
Tabac : No
Hormonal treatment: No
```

Figure IV.20: Exemple de dossier médical

4.2 Prétraitement des données

Rappelons que notre objectif est de découvrir de nouvelles corrélations entre les concepts de l'ontologie mammographique globale. Nous nous sommes intéressés particulièrement, dans cette partie, aux associations de coexistence entre les observations radiologiques/ cliniques et les diagnostics/ classification.

Le choix des attributs est une tâche très importante puisqu'elle permet de décider concernant la nature des connaissances à extraire. Dans cette étape, on a sélectionné 16 attributs comportant différentes valeurs numériques et catégoriques qui sont en relation avec ces classes *: les observations radiologiques, les observations cliniques, les diagnostics et les classifications* (Le Tableau IV.8 illustre les attributs choisis). Ensuite, nous avons procédé à une étape de transformation de l'ensemble de données semi-structurée dans un formalisme approprié (fichier de type .ARFF) pour l'application de l'algorithme, où chaque dossier médical représente une transaction ou un enregistrement modélisé par un ensemble de valeurs d'attributs ou item. Dans cette étape, nous n'avons gardé que les attributs d'intérêt et nous avons corrigé les inexactitudes et/ou erreurs dans les données. Nous avons également analysé certaines valeurs d'attributs numériques et étudié la possibilité de les regrouper en des intervalles, comme par l'exemple, les attributs âge, âge de ménopause, âge de ménarche, (en

effet, ça sera plus approprié de les généraliser que les analyser valeur par valeur). Il est à noter que nous nous sommes basés sur les labels des concepts dans l'ontologie pour la définition des attributs (par exemple l'attribut masse ovale est réécrit '*Oval_mass*' comme dans l'ontologie).

Nous présentons dans le Tableau IV.9 une illustration de l'étape de transformation des données brutes (dossiers médicaux) en des données formelles où les lignes représentent les transactions (T1→T7), les colonnes (A→E) représentent les attributs qui peuvent être manquants ou présents dans chaque transaction. Dans la transformation, les attributs manquants sont présentés par « ? ».

Bien que cette partie du processus requière un temps de traitement important, son déroulement est déterminant afin de pouvoir extraire les interactions entre les valeurs des attributs choisis (qui représentent les concepts du domaine dans l'ontologie).

Tableau IV.8: Attributs sélectionnés

Id	Attribut	Description	Exemple de valeurs possibles
1	AGE	Age de la patiente	[50-59][60-69], etc.
2	MNPS	Age de la ménopause de la patiente	[50-59][60-69], etc.
3	BREAST	Nature de sein	'Adipose', 'dense', etc.
4	ANTCDT	Antécédent familial de la maladie	'Yes', 'No'
5	CLINICAL_EX	Examen clinique	'Breast_pain', 'Dimpled_skin ', 'halo_peritumoral', ' Skin_thinning', etc.
6	TUMOR	Classe de la tumeur	'Mass' , 'Calcificaion',etc.
7	SIZE	Taille de la tumeur	'Large_mass', 'Small_mass',etc.
8	SHAPE	Forme de la tumeur	'Round_mass', 'Oval_mass',etc.
9	CONTOUR	Contour de la tumeur	'Circumscribed_mass', 'Microlobulated_mass',etc.
10	DNSTE	Densité de la tumeur	'mass_low_density', 'Isodense_mass',etc.
11	DISTRIBTN	Distribution de la tumeur	'Linear_mass',etc.
12	MRPHLGIE	Morphologie de la tumeur	'polymorph_mass', 'amorph_mass',etc.
13	GPMENT	Type de groupement	'uni_mass', 'pluri_mass',etc.
14	ClASS	Classification de la mammographique	'BI-RADS0', 'BI-RADS1',etc.
15	Typ_Cancer	Type de cancer	'Diagnosis_benin', ' Malignant_diagnosis'.
16	Exmn	Examen proposé	'Follow-up', 'Biopsy', 'Echography', etc.

Tableau IV.9: Formalisation des données

Transaction	Liste des items
T1	A, C, E
T2	A, D
T3	B, C, D, E
T4	A, B, C
T5	A, C
T6	C, D, E
T7	A, B, C, D, E

Transformation

	A	B	C	D	E
T1	x	?	x	?	x
T2	x	?	?	x	?
T3	?	x	x	x	x
T4	x	x	?	?	x
T5	x	?	x	?	?
T6	?	?	x	x	x
T7	x	x	x	x	x

La phase de préparation concerne également l'extraction des prédicats déclarés dans l'ontologie. Ces prédicats serviront de base pour la sélection de nouvelles RAs qui feront l'objet d'enrichissement. A cette fin, ces prédicats seront transformés en des règles ontologiques. En effet, chaque prédicat se présente sous la forme :

Prédicat : $<Domain, object - property, Range>$

Exemple : <Lesion, has_diagnosis, diagnosis>

Où $domain$ et $range$ représentent des concepts dans l'ontologie, $object - property$ est la relation associant les deux concepts. Ce prédicat est ainsi transformé en une règle ontologique qui se présente sous la forme :

RO : $Domain \rightarrow Range$

Exemple : Lesion $\rightarrow$ diagnosis

Où $\rightarrow$ désigne l'implication associée ente les deux concepts, la partie antécédent est ainsi présentée par $Domain$, $Range$ est assignée à la partie conclusion.

A l'issue de cette étape, nous avons obtenu 30 ROs, notons que nous n'avons gardé que les ROs dont l'antécédent et la conclusion appartiennentt à la liste des attributs sélectionnés pour le jeu de données. Le Tableau IV.10 illustre quelques exemples des règles extraites à partir de l'ontologie.

Tableau IV.10: Extrait des règles ontologiques RO

	Règles
R1	<BI-RADS3→Follow-up>
R2	<BRCA1→Malignant_diagnosis>
R3	<BRCA2→Malignant_diagnosis>
R4	<Mass_irregular→Malignant_diagnosis>
R5	<Mass_regular → Diagnosis_benin >
R6	<BI-RADS1 → Diagnosis_benin >
R7	<Parallel_mass→Diagnosis_benin>
R8	<mammogram_heterogenous→ Malignant_diagnosis >

4.3 Mapping des concepts de l'ontologie et les items

Cette étape permet de connecter les concepts dans l'ontologie avec les items définis dans les règles. La connexion a été réalisée manuellement à cause du nombre restreint des attributs ainsi que de leurs valeurs. Néanmoins, nous convenons que pour de grandes bases de données, une connexion manuelle peut être fastidieuse.

Comme nous l'avons déjà mentionné, les valeurs des attributs (items) sont directement connectées avec les concepts *X* qui lui sont identiques ($ƒ_{direct}$ (Oval_mass)=Oval_mass). Ensuite, ces items sont connectés aux concepts pères généralisant de X, $ƒ_{indirect}$ (Oval_mass)=mass. Ce type de connexion permet d'identifier les règles implicitement identiques entre les deux bases de connaissances.

Prenons cet exemple ; le concept Sclerosing_Adenosis est déclaré un concept fils de Adenosis. Ce dernier est en relation avec le concept Diagnosis-benin ; cette connaissance sera modélisée comme suit :

- RO : 'Adenosis →Diagnosis-benin'.

Soit une RA indiquant :

- RA: 'Sclerosing_Adenosis →Diagnosis-benin'.

La connexion établie de l'item Sclerosing_Adenosis est: ƒ(Sclerosing_Adenosis)={ Sclerosing_Adenosis, Adenosis }. On peut déduire ainsi que la RO et RA sont sémantiquement similaires.

Les concepts ontologiques connectés aux items de la base de données feront l'objet d'enrichissement relationnel puisqu'ils représentent les jeux de données dans le processus d'extraction des RAs.

4.4 Extraction des règles d'associations

Dans ce paragraphe, nous procédons à l'extraction des règles d'association en appliquant l'algorithme classique Apriori avec l'éditeur Weka 3.6.1[7]. Afin de contrôler les règles extraites, deux valeurs seuils sont fixés au préalable à savoir: MinSup et MinConf. Le Tableau IV.11 présente, les résultats du nombre de RAs générées suite à plusieurs variations des valeurs seuils.

Tableau IV.11: Nombre de RAs générées en fonction des MinSup et MinConf

Itération *i*	1	2	3	4	5	6	7	8	9	10
MinSup	0.7	0.3	0.1	0.7	0.3	0.1	0.3	0.09	0.07	0. 05
MinConf	1	1	1	0.9	0.9	0.9	0.8	0.7	0.7	0.7
RAs	0	0	74	2	23	536	73	2167	3395	6552

Les résultats, illustrés dans le Tableau IV.11 montrent l'influence de ces mesures sur la qualité et le nombre des règles déduites. D'après ces résultats, on peut constater, d'une part, que plus on diminue les valeurs de MinConf et MinSup, plus important est le nombre de

[7] http://sourceforge.net/projects/weka/files/weka-3-6/3.6.1/

règles obtenues. Même en maintenant, la valeur de MinConf constante et en diminuant le MinSup, le nombre de règles reste considérable (comme pour les itérations 8,9 et 10). Néanmoins, ces règles peuvent ne pas être fiables et robustes à cause de leur faible fréquence d'occurrence dans la base de données. Par ailleurs, en ayant un MinSup et MinConf importants (itérations 1, 2 et 4), on peut noter que le nombre de règles obtenues est beaucoup moins important (par exemple MinSup=0.7 aucune règle n'a été générée). Nous pouvons remarquer que ces règles sont robustes puisque l'antécédent et la conclusion sont presque toujours associés.

Dans ce qui suit, nous avons empiriquement fixé les valeurs MinSup et MinConf respectivement aux 0.1 et 0.7, ce qui conduit à l'extraction de 1756 RAs dont 415 RAs sont des règles de 2 items. Rappelons que nous nous intéressons, dans cette partie, aux règles de deux items pour l'enrichissement relationnel (appelées aussi règles de base [Ruiz, 2014]), vu que notre objectif c'est d'extraire des implications qui seront modélisées comme des object-property dans l'ontologie cible. Un extrait de ces règles est présenté dans Tableau IV.12.

Tableau IV.12: Extrait des règles d'association obtenues (avec MinSup= 0.1, MinConf=0.7)

Règles	**Support**	**Confiance (%)**
<Oval_mass→ benign_diagnosis>	0.6	0.85
<Age [50-59]→ Malignant_diagnosis>	0.41	0.83
<Age [60-69]→ Breast_pain>	0.42	0.75
<Oval_mass→mass_density-low>	0.56	0.76
<Mass_spiculated→BI-RADS5>	0.61	0.75
<Oval_mass→→BI-RADS4>	0.4	0.74
< Age [60-69]→Menopause[40-49]>	0.43	0.96
<micro_calcificaion_irreguler→Biopsy >	0.22	0.75

Les connaissances extraites sont des RAs entre les concepts de la forme : $concept_i \rightarrow concept_j$. Chaque RA est associé à ses mesures de support et de confiance.

R1: Oval_mass→ benign_diagnosis sup=0.6, conf=0.85

Interprétation: On peut constater que les deux items *'Oval_mass'*, '*benign_diagnosis'* sont souvent présents dans la base de données (sup=60%). Cette règle permet de savoir que les masses de forme ovale conduisent souvent à un diagnostic bénin. Une explication possible de

cette relation réside dans le fait que les formes lisses des lésions sont souvent révélatrices de bénignité.

R2: Age [50-59]→ Malignant_diagnosis sup=0.41, conf=0.83

Interprétation: Cette règle indique que les patientes âgées de 50 à 59 ans sont atteintes du cancer du sein. Ces deux items sont souvent liés dans la base de données (sup=0.41). En effet, 41% des femmes dont l'âge varie entre 50 et 59 correspondent à cette situation. Une explication possible de cette relation réside dans le fait que l'incidence du cancer du sein augmente avec l'âge. En effet, selon un expert, la ménopause survient entre l'âge de 45 et 55 ans en moyenne, après l'âge de 55 ans, on parle de ménopause tardive, ceci constitue un facteur de risque, contrairement au ménopause précoce c un facteur protecteur

R3: Age [60-69]→ Breast_pain sup=0.42, conf=0.75

Interprétation: Cette règle indique qu'un pourcentage important des femmes âgées de 60 à 69 souffre de douleur de sein. Cette règle met en association deux items appartenant à la classe de facteurs de risque (observations cliniques). Bien que cette règle dispose d'une valeur de support importante, elle n'est pas validée par l'expert, puisqu'elle ne peut pas être généralisée et par la suite rajoutée dans l'ontologie.

R4: Micro_calcificaion_irreguler→Biopsy sup=0.22, conf=0.75

Interprétation: Selon les experts, cette règle peut être interprétée comme suit : les micro calcifications sont de petites dépôts de calcium dans le sein. Elles peuvent laisser croire que l'activité est accrue dans certaines cellules du sein. Lorsque ces cellules sont plus actives, elles absorbent plus de calcium du corps. Si les micro-calcifications semblent suspectes, le radiologiste peut suggérer une biopsie.

Nous remarquons, d'après ces résultats, que certaines connaissances apportées par les règles sont considérées significatives pour l'expert. Néanmoins, quelques relations sont surprenantes et/ou non significatives et/ou limitées à la base de données et ne peuvent pas être généralisées. Egalement, certaines règles extraites ont un faible support ce qui montre qu'il y a moins de probabilité pour que ces règles soient triviales. En même temps, la valeur de confiance élevée montre que les règles sont importantes et qu'elles doivent être considérées. Par conséquent, l'intervention de l'expert pour l'analyse et la validation de ces règles est cruciale pour pouvoir enrichir l'ontologie.

On peut noter également que chacune de ces règles présentées, exprime une relation particulière modélisée par l'implication '→'. L'interprétation de cette implication dépend étroitement des items trouvés dans l'antécédent et la conclusion de la règle. Egalement, on doit signaler que l'expert peut effectuer des permutations des items entre la partie 'antécédent' et la partie 'conclusion' des règles (dans ce cas la valeur du support de la règle reste intacte, mais la valeur de confiance peut changer).

4.5 Post-traitement de 2-items RAs

A ce stade l'expert doit analyser l'utilité des RAs extraites ce qui est impossible manuellement. A cette fin, on présente, dans cette section, les étapes de filtrage et de ranking appliquées sur les RAs afin de simplifier leur vérification et garder celles les plus utiles à l'enrichissement de l'ontologie.

4.5.1 Filtrage des règles d'association

L'expert s'intéresse aux RAs différentes de celles prédéfinies dans l'ontologie. A cette fin, les RAs portant des connaissances connues sont éliminées. La comparaison entre les deux sources de connaissances se fait à travers la comparaison de leurs prémisses et conclusions. Deux règles ayant les antécédents et les conclusions respectivement équivalents ou inversement équivalents sont conformes.

Chaque RO dans l'ensemble des règles ontologiques est confrontée à l'ensemble des RAs pour générer la ou les règles qui lui sont semblables CF(RO)= $\{RA_i, RA_j, etc.\}$. Les règles générées par l'opérateur de conformité sont éliminées de l'ensemble des RA. La confrontation des deux ensemble 30 ROs et 415 RAs a induit à la suppression 14 RAs soit 3% de l'ensemble totale des RAs. Le Tableau IV.13 illustre un extrait quelques RAs qui ont été filtrées par l'opérateur de conformité.

Tableau IV.13 : Extrait des RAs déduites par l'opérateur de conformité

	Règles
R1	<BI-RADS3→Follow-up>
R2	<Oval_mass→Diagnosis_benin>
R3	<BI-RADS1 → Diagnosis_benin >
R4	<Mass_irregular→Malignant_diagnosis>
R5	<Breast_pain → mastalgia>
R6	<fibrocystic_typic → normal_breast>
R7	<BI-RADS4→Biopsy>
R8	<Flow_bloody_breast→Malignant_diagnosis>

Le Tableau IV.14 illustre un extrait des règles d'association qui n'ont pas été déterminées par l'opérateur de conformité.

On remarque également que la majorité des ROs n'ont pas été mises en correspondance avec des RAs. Ceci est expliqué par le non exhaustivité de la base de données utilisée qui représente un échantillon restreint des cas possibles. Nous convenons que la taille de la base de données peut affecter le nombre des ROs portant de nouvelles connaissances. De ce fait, plus la base de données est volumineuse et diversifiée, plus des règles pertinentes peuvent être découvertes.

Tableau IV.14 : Extrait des RAs portant de nouvelles connaissances

	Règles
R1	<mammogram_heterogenous→ Malignant_diagnosis >
R2	< Skin_retraction →Malignant_diagnosis>
R3	< Skin_thinning → Malignant_diagnosis >
R4	< halo_peritumoral →Malignant_diagnosis>
R5	<Cyst→ Diagnosis_benin >
R6	< Dimpled_skin → Malignant_diagnosis >
R7	<BI-RADS3→Echography>
R8	<Tillaux_maneuver_positive→ Malignant_diagnosis >
R9	<BI-RADS5→Biopsy>
R10	<Micro_calcificaion_irreguler→Biopsy >
R11	<Cyst→ BI-RADS2>

4.5.2 Ranking des 2_items RAs

Cette étape permet d'aider l'expert de se focaliser sur les règles les plus importantes parmi les 391 RAs filtrées. On peut noter que les RAs générées peuvent mettre en association des concepts (items) appartenant à la même catégorie (classe) ou à différentes catégories. Le but de cette étape est d'aider l'expert à porter une attention particulière sur les règles de grande importance. A cette fin, nous appliquons une mesure de qualité qui présente le comportement des RAs en termes d'importance. Selon l'expert, plus les items sont distants (n'appartenant pas au même cluster), plus la règle est importante. Cette distance dépend, ainsi, de la distance des clusters correspondants dans la hiérarchie.

Tableau IV.15: Extrait des RAs classées selon la distance des items

Règles	**Mesure d'intérêt**	**Support**	**Confiance**
< Cyst-→BI-RADS2>	1	0.11	0.83
<Skin_retraction →Malignant_diagnosis>	1	0.22	0.75
<Skin_thinning → Malignant_diagnosis>	1	0.36	0.76
< Halo_peritumoral →Malignant_diagnosis>	1	0.21	0.75
<Dimpled_skin → Malignant_diagnosis >	1	0.40	0.74
<Blood_invasion→ Malignant_diagnosis >	1	0.12	0.81
<Axillary_node→ Malignant_diagnosis >	1	0.11	0.77

<BI-RADS3→Echography>	0.66	0.13	0.96
<Malignant_diagnosis → Biopsy >	0.83	0.31	0.75
<BI-RADS5→Biopsy>	0.66	0.24	0.74

Bien que la mise en relief de l'utilité et de la validité des règles demande une excellente connaissance du domaine auquel nous n'avons pas accès, nous convenons que l'utilisation d'un support sémantique telle que l'ontologie pour l'évaluation et/ou le post traitement semble très prometteuse. Cette dernière permet de quantifier l'apport de la RAs par rapport au domaine d'étude.

4.5.3 Sélection des « meilleures » règles

Pour toute valeur σ fixée, on retourne un sous-ensemble de règles tel que la distance des items pour chacune des règles soit supérieure ou égale à σ. Ainsi, en observant l'évolution de la taille de l'ensemble des règles les plus intéressantes en fonction de la valeur σ, ceci nous permet d'évaluer la qualité des RAs générées.

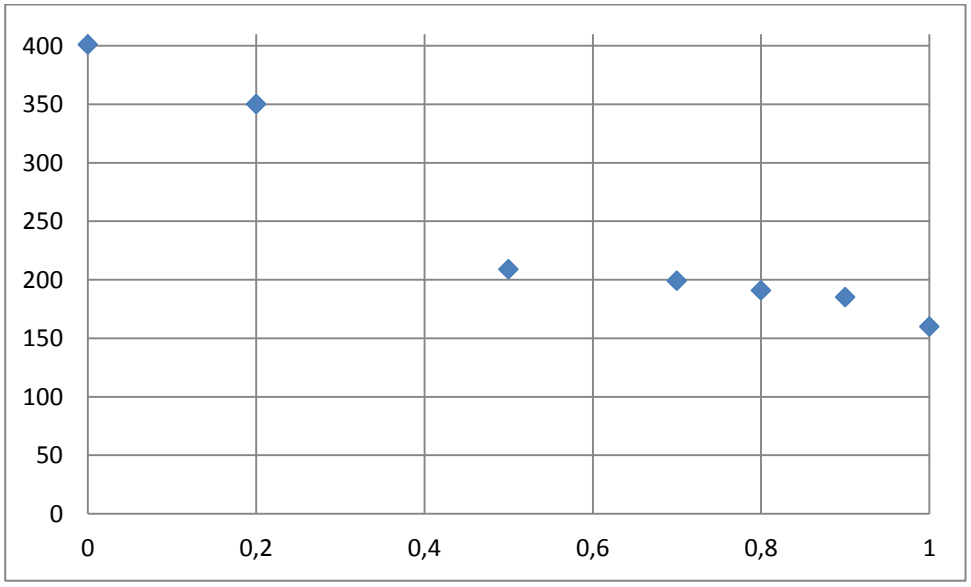

Figure IV.21 : Evolution des règles intéressantes en fonction du seuil σ

La Figure IV.21 présente l'évolution de la taille de cet ensemble. On distingue un changement distinctif de comportement lorsque σ dépasse 0.4. Ceci signifie que 50% des RAs disposent d'une distance sémantique entres les items supérieure à 0.5.

4.6 Enrichissement Relationnel de l'ontologie cible

Les RAs classées ont été introduites à l'expert afin de les analyser, interpréter et garder celles les plus pertinentes. A l'issue de cette étape, l'expert a pu valider 19 RAs parmi les 391 RAs en entrée, soit 5% de l'ensemble original. Les corrélations validées ont été introduites

dans l'ontologie en fonction de la relation établie entre les deux concepts (domain-range), où un label a été assigné.

Le Tableau IV.16 présente un extrait des relations introduites au sein de l'ontologie et décrivant les différentes corrélations ainsi que les noms de prédicats associés.

Tableau IV.16: Labels des relations enrichies dans l'ontologie

Label de la relation	<Domaine →range>
Co_occur	<Breast_pain→ mastalgia>
O_has_Assessment	<skin_retraction →Malignant_diagnosis>
R_ has_Assessment	<Oval_mass→Diagnosis_benin>
D_ has_Assessment	< Malignant_diagnosis → Biopsy>
Require	<BI-RADS3→Echography>
O_has_Assessment	<skin_lesion→ Biopsy>
O_Associated_with	<Skin_thinning → Malignant_diagnosis>
R_ has_Assessment	<Cyst-→BI-RADS2 >

4.7 Etude des multi-items RAs

Nous avons également procédé à l'analyse des RAs disposant de plusieurs items. Ces règles ont été classées selon leur importance en utilisant l'équation III.16. Le Tableau IV.17 présente un extrait des RAs multi-items avec leurs mesures d'intérêt respectives. Ces modèles de connaissances sont de grand intérêt et peuvent, ultérieurement, enrichir la base de connaissances de référence par de nouvelles règles et/ou axiomes. A cette fin, le post traitement et l'étude de la pertinence de ces connaissances par les experts du domaine sont envisageables afin de dégager les connaissances utiles à l'enrichissement.

Tableau IV.17: Extrait des RAs multi-items classées selon la mesure d'interet

Règles	Interprétation sémantique	Mesure d'intérêt
Age [60-69], mass_circumscribed→ Bi-rads4	L'association de la masse circonscrite et âge [60-69] entraînent une classification de Bi-rads 4.	1
Age[60-69], Oval_mass, mass_circumscribed→ Diagnosis_benin	Les masses ovales et bien circonscrites sont hautement prédictives d'une lésion bénigne.	0.83
mass_low_density, mass_circumscribed→ Bi-rads4	Les masses circonscrites et de faible densité entraînent une classification Bi-rads 4.	0.66
Round_mass, mass_circumscribed→ Bi-rads4	Les masses rondes et circonscrites entraînent la classification Bi-rads 4.	0.66

5. Synthèse

L'application de l'approche proposée dans ce travail au domaine de la mammographie a permis d'enrichir une ontologie existante par de nouveaux éléments pertinents et utiles à la compréhension du domaine afin d'améliorer la couverture des connaissances. Outre les concepts, de nombreuses relations ont été extraites pour mieux structurer l'ontologie.

En premier lieu, les experts ont vérifié si les concepts sont pertinents pour le domaine. Ils ont ensuite validé les liens taxinomiques les rattachant aux autres concepts du noyau ontologique. En deuxième lieu, les relations non taxonomiques issues de l'étape de la fouille de données les experts ont été aussi analysés, ce qui a permis de dégager des connaissances additionnelles approuvées telles que :

-La relation de coexistence entre certains symptômes/caractéristiques génétiques (BRCA1, BRCA2, etc.) et nature de diagnostic.

-Déduire que certaines observations cliniques/symptômes peuvent être présentes de manière concurrente.

-Déduire que quelques observations cliniques (telle que la rétraction de la peau ou amincissement de la peau) peuvent révéler une malignité. (Ce qui permet un diagnostic précoce du cancer).

-Déduire que la forme, la texture et l'opacité de la tumeur peuvent révéler sa bénignité ou malignité.

Ontologie	Métriques	But de conception	Caractéristiques
MammOnto	-non disponible	- fournir un vocabulaire communément admis et des définitions formelles afin de décrire les mammographies.	- Définition des résultats anormaux des examens et des évaluations médicales. -Utilisation du lexique Bi-Rads.
BCGO	-531 concepts -100 propriétés	- Intégration dans un système de classification des mammographies.	-Elle est associée à des règles écrites en langage SWRL pour le raisonnement inférentiel. -Modélisation exhaustive des concepts liés à l'évaluation et la catégorisation des mammographies (ACR).
GIMI Mammography Ontology'	-Core Mammographic Ontology -Mammography Learning Ontology	- Intégration dans un système d'apprentissage pour comparer les annotations des stagiaires avec celles des experts	-Elle liste différents concepts liés aux facteurs de risque. -Bien structurée et contient un grand nombre de concepts.
Core Mammographic Ontology	-692 concepts -135 propriétés,	-Description des concepts pertinents au domaine mammographique.	- Les concepts s'articulent principalement autour de ces grandes familles: les entités anatomiques, les entités conceptuelles et les diagnostics.
Mammography e-Learning Ontology	-740concepts -142 propriétés	-Comparaison des annotations des stagiaires par rapport à celles de l'expert.	-Cette ontologie ne peut être fonctionnelle. Elle est limitée au contexte de l'application dont laquelle elle est intégrée.
Mammographic Ontology'	-48 concepts	- La recherche et l'extraction d'information	- Combler le fossé sémantique entre les fonctions de bas niveau extraites par les méthodes de traitement d'image et les concepts de haut niveau (la sémantique).
MAO	-non disponible	- Intégration dans un système d'annotation des anomalies observées dans une mammographie.	-Les concepts relatifs aux anomalies évoquées dans l'ontologie sont très restreints à savoir les masses et les calcifications.
Breast-Cancer Ontology	-non disponible	-Intégration dans une application destinée aux usagers non professionnels	-Cette ontologie met l'accent sur les concepts de nature informative aidant les usagers de santé à formuler leurs problèmes.
L'ontologie proposée	**-802 concepts** **-154 propriétés**	**-Modélisation formelle et exhaustive des connaissances mammographiques.** **-Cette ontologie peut être exploitée dans plusieurs applications (Indexation et recherche d'information, systèmes d'aide à la décision)**	**-Elle est basée sur la fusion des connaissances provenant des sources ontologiques existantes (du même domaine).** **-Elle est riche en relations taxonomiques et non taxonomiques. Ces derniers sont issus des données réelles du domaine.** **-Elle couvre les différents aspects du domaine mammographique.** **-Les connaissances sont validées par un expert de domaine.**

Tableau IV.18: Positionnement par rapports aux ontologies mammographiques ontologique

6. Conclusion

Dans ce chapitre, nous avons appliqué le processus d'enrichissement proposé dans le chapitre précèdent en utilisant des bases de connaissances et des bases de données réelles liées au domaine de la mammographie. Nous nous sommes intéressés dans cette étude aux ontologies mammographiques existantes pour l'enrichissement conceptuel et aux dossiers médicaux provenant des hôpitaux (Ben Arous de Tunis et Taher sfar de Mahdia) pour l'enrichissement relationnel.

Egalement, on a essayé d'évaluer notre méthode, à savoir les modules de clustering et d'alignement, en utilisant des métriques standards issues de la littérature, ce qui nous a permis de positionner nos travaux de recherche. Ensuite, nous avons montré l'intérêt de nos contributions de fouille de connaissances ontologiques ainsi que le post traitement des règles d'association afin de sélectionner celles les plus importantes.

Nous considérons que ces premiers résultats de l'application de notre approche dans le domaine mammographiques sont très prometteurs. En effet, le processus de fouille de connaissances proposé nous a permis dans un premier temps de présenter les connaissances de l'ontologie d'une manière plus simplifiée et dynamique pour l'expert de domaine. Dans un second temps, ce processus nous a aidé à interpréter et quantifier l'intérêt des RAs conformément aux attentes des experts.

Il est également envisagé d'augmenter la liste des sources de connaissances en incluant d'une part de nouvelles ontologies du domaine mammographique et d'autre part, d'autres dossiers médicaux.

La méthode implémentée est une méthode semi-automatique qui dépend de l'intervention de l'utilisateur. Ainsi, nous admettons que l'automatisation du processus suggéré et la minimisation de l'intervention de l'expert sont fortement recommandées afin de simplifier l'utilisation de l'application et pouvoir se positionner par rapport aux méthodes d'alignements existants en termes de temps d'exécution.

V. Conclusions & Perspectives

Conclusions

Avec l'avènement des appareils de radiologie de sein, les radiologues rencontrent des difficultés dans la prise de décision. Ceci est dû principalement à l'ambiguïté des informations médicales ainsi que la complexité du domaine d'étude. De ce fait, les experts de domaine se retrouvent face à une nécessité grandissante en systèmes d'analyse d'images mammographiques et d'aide à la décision.Néanmoins, un système d'aide au diagnostic, doit se baser sur une modélisation de connaissances fiable et exhaustive du domaine. Parmi les modélisations des connaissances possibles, nous citons les ontologies.

La richesse du contenu informatif de plusieurs domaines tel que le domaine mammographique a induit ; d'une part à plusieurs modélisations ontologiques conçues différemment, et modélisant différents parties de domaines et d'autre part, une grande base de données des patients atteints et non atteints du cancer.

A la croisée de l'ingénierie des connaissances et du web sémantique, les travaux de cette thèse s'intéressent à la gestion des ontologies où l'on cherche avant tout à réutiliser des ressources de connaissances hétérogènes existantes où l'hétérogénéité sémantique de ces ressources constitue un frein à leur utilisation.

Dans nos travaux de thèse, nous exploitons la richesse du contenu informatif de ces sources hétérogènes pour une meilleure compréhension du domaine. L'originalité de notre approche s'introduit à travers la définition d'un nouveau concept : *la fouille de connaissances*. Cette notion, a été employée dans le but d'extraire de nouvelles connaissances à partir des connaissances existantes. Partant de cette notion, l'approche que nous proposons se base sur l'exploitation de deux principaux axes portant sur l'enrichissement relationnel et conceptuel d'une ontologie existante.

La nouvelle approche que nous avons proposée consiste en un processus de couplage entre le processus d'enrichissement conceptuel et le processus d'enrichissement relationnel à partir

d'une base de données du domaine. Les contributions apportées dans le cadre de ce travail sont :

Dans le volet « enrichissement relationnel »:

- Le pré-alignement ou la préparation d'ontologies basé sur le clustering hiérarchique flou des concepts de l'ontologie et ayant pour objectif la réorganisation de la structure ontologique de manière à grouper les concepts similaires dans des clusters répartis sur différents niveaux de granularité.
- L'alignement des ontologies en se basant sur les structures hiérarchiques des clusters qui leurs correspondent.
- L'enrichissement de l'ontologie de référence par de nouveau concepts.

Dans le volet « enrichissement relationnel »:

- L'extraction des connaissances à partir des données sous la forme de règles,
- La confrontation des deux sources de connaissances, provenant de l'ontologie de référence et des règles d'association extraites, afin d'élaguer les règles portant des connaissances déjà connues
- L'évaluation et le *ranking* des règles d'association en utilisant des mesures d'intérêts. Le calcul de la mesure d'intérêt est basé sur une nouvelle distance sémantique exploitant la structure hiérarchique des clusters proposée dans le premier volet.
- L'enrichissement de l'ontologie par de nouvelles relations.

Le premier axe se base sur l'exploitation des ressources ontologiques de domaine pour l'enrichissement conceptuel d'une ressource cible. Dans un premier temps, nous avons proposé une approche d'alignement qui vise à pallier aux problèmes d'hétérogénéité entre les modélisations ontologiques. Nous avons considéré les problèmes de voluminosité et de complexité de domaine d'étude en introduisant une étape de pré-alignement d'ontologies basée sur la fouille de connaissances. Au cours de cette étape, les structures ontologiques à aligner ont été transformées en des structures hiérarchiques des clusters conceptuels flous. Ceci vise à répartir les concepts sémantiquement similaires en des blocs afin de faciliter la tâche de recherche de correspondance. Par rapport à la littérature, notre approche proposée se caractérise par deux spécifications : l'utilisation des clusters flous ainsi que l'utilisation de la hiérarchisation des clusters.

Le concept flou des clusters permet d'assigner une entité à plusieurs clusters simultanément ce qui évite la perte d'information lors de la tache de clustering. Ceci est du également au fait qu'un concept défini au sein de l'ontologie possède plusieurs attributs et propriétés permettant de l'affecter à plusieurs classes.

Parallèlement, la hiérarchisation des clusters permet de mener la comparaison du cluster cible sur des clusters spécifiques de la hiérarchie source à travers la propagation de la similarité d'un niveau à un autre.

Nous avons également proposé une nouvelle mesure de similarité sémantique permettant de quantifier la similitude entre deux concepts appartenant à la même ontologie. Cette mesure est étroitement liée au contexte relationnel définissant une entité ontologique. Selon les expérimentations, les performances de notre mesure sont meilleures que celle proposée par Wu & Palmer.

Une fois cette étape de pré alignement est achevée, nous procédons à l'étape d'alignement qui consiste en deux phases. La phase d'ancrage qui permet de rapprocher les clusters sémantiquement comparables de deux ontologies (c.à.d. contenant des éléments majoritairement liés par des liens d'équivalence) et la phase de dérivation qui permet d'identifier les différents alignements entre les éléments des clusters similaires à travers l'utilisation de différents techniques de similarité.

La phase d'alignement a pour objectif de déterminer les nouvelles connaissances et les rajouter dans l'ontologie choisie comme ontologie source afin d'obtenir une ontologie exhaustive liée au domaine mammographique.

Ces contributions ont donné naissance à un outil d'alignement d'ontologies basé sur trois modules exécutant chacun une fonction bien particulière du processus : Module de chargement d'ontologie, module de clustering d'ontologie et module d'alignement.

Enfin, nous avons validé les différentes propositions de ce travail en comparant les résultats obtenus par rapport aux systèmes existants. A cet effet, nous avons testé l'algorithme de clustering hiérarchique flou proposé, ainsi que l'algorithme d'alignement d'ontologie.

Nous avons montré, également, l'apport de l'utilisation de fouille de connaissances pour l'amélioration des résultats d'alignement.

En deuxième partie, nous avons proposé d'exploiter les données des patients antérieurement étudiés et diagnostiqués par des experts de domaine dans le but d'extraire de nouvelles associations entre les termes pertinents de domaine d'application. Pour se faire, nous avons proposé d'utiliser les règles d'association en tant qu'un outil de fouille de données. Ce type de règles permet non seulement d'identifier les cooccurrences des items mais aussi les relations d'implications entre les observations dans une mammographie et la classification correspondante.

Les principaux avantages de la méthode d'extraction de règles d'association sont dus au fait que le modèle des motifs extraits est relativement simple et compréhensible par un utilisateur non spécialiste en data mining (puisque des relations sous la forme de règles sont proches du raisonnement humain). Cependant, plusieurs limites ont été rencontrées lors de sa mise en œuvre, tels que la tendance à découvrir un grand nombre de règles d'association ou le fait que toutes les règles ne sont pas pertinentes ou intéressantes pour les utilisateurs finaux; les règles extraites pouvant s'avérer redondantes, incohérentes ou encore contradictoires ; et la difficulté pour interpréter les règles en tenant compte des mesures d'intérêt ou des indicateurs statistiques.

Afin de palier à ces inconvénients, nous avons proposé un nouveau processus d'évaluation et de post traitement des connaissances générées afin de garder celles qui sont différentes des connaissances existantes. Notre contribution dans ce sens réside dans la proposition d'une démarche structurée qui prend en compte trois manières d'évaluer les règles:

- Une évaluation objective : cette étape est basée sur les mesures d'intérêt associées aux règles (nous avons utilisé dans nos travaux les deux mesures classiques prises en compte dans les algorithmes d'extraction de règles : le support et la confiance),
- Une évaluation sémantique : cette étape considère les connaissances existantes du domaine,
- Une évaluation subjective : dans cette étape nous avons proposé une nouvelle mesure d'intérêt basée sur la distance entre les items constituant la règle. Selon les experts de domaine, cette mesure est d'autant plus intéressante, lorsque les valeurs d'attributs sont différentes. A cet effet, cette distance est mesurée en se basant sur la hiérarchie des clusters conceptuels générés dans la première partie de nos travaux. Plus les clusters correspondants aux concepts sont éloignés dans la hiérarchie, plus la règle d'intérêt est importante.

Les règles filtrées et classées, sont finalement présentées à l'expert, qui lui seul peut juger la pertinence de ces nouvelles connaissances. Les connaissances générées sont finalement, enrichies dans la ressource ontologie cible, en rajoutant de nouvelles relations entre les concepts mis en jeux.

Perspectives

Comme perspectives, plusieurs axes de recherche peuvent être envisageables à partir de ce travail de recherche actuel. Ces perspectives se rapportent d'une manière générale à l'approche mise en œuvre dans ce travail de thèse, et d'une manière spécifique à l'approche de clustering conceptuel d'ontologie.

L'approche mise en œuvre est une approche extensible, c.à.d. elle pourrait être dotée pour permettre l'expression d'autres types d'enrichissement d'ontologie, tel que l'enrichissement axiomatique. Une vision de notre travail serait donc de chercher d'autres perspectives d'enrichissement de l'ontologie mammographique étudiée et appliquer, également, cette approche à d'autres domaines d'application.

Une utilisation intensive de cette approche (en sollicitant d'autres ontologies de domaine) conduira à multiplier le nombre de nouvelles connaissances (concepts et relations). Ceci nous amène donc à une seconde perspective de ce travail. En présence d'un nombre important des entités candidats d'enrichissement, il devient nécessaire de concevoir une aide automatisée pour la sélection et l'ajout de ces éléments.

Concernant nos travaux portant sur le clustering hiérarchique des concepts d'ontologies, une perspective intéressante et envisageable serait d'automatiser la détermination du nombre de sous clusters générés dans chaque niveau. Pour l'instant, dans nos travaux, le nombre de sous clusters est fixé, par défaut, à deux. Ceci constitue un point faible qu'il faut résoudre en l'automatisant ou l'outillant. Le nombre de sous clusters peut être défini de différentes façons. Il peut s'agir de fixer, à chaque fois, un nombre de clusters et évaluer à chaque fois la qualité des clusters.

Toujours concernant nos travaux sur le clustering, une perspective envisageable constitue à rendre l'approche de clustering orientée alignement, c'est-à-dire de considérer l'objectif d'alignement lors de processus de clustering. Ceci peut se faire, par exemple, par le co-clustering des ontologies en jeux en même temps. Une étude approfondie de cette approche est envisageable. Une comparaison par rapport aux résultats obtenus de la version actuelle de notre approche sera également menée.

Nous projetons, également, d'implémenter et mettre en œuvre les étapes de mise à jour de la hiérarchie après l'enrichissement ainsi que l'enrichissement conceptuel par clusters. Cette dernière est de grande importance puisqu'elle permet d'enrichir la base de connaissances par de groupement de concepts à la fois, contrairement aux méthodes d'enrichissement cassiques de la littérature (enrichissement concept par concept). Néanmoins, cette étape est très compliquée puisque la MAJ affecte tous les clusters concernés par l'enrichissement (y inclut les medoides des clusters ainsi que les degrés d'appartenance des concepts) ce qui rend le processus de validation des mappings lourd et moins sûr. Partant de ce constat, il faudrait engager une méthodologie efficace, fiable et non couteuse pour la maintenance et la MAJ de la hiérarchie. Ceci représente un axe de recherche à explorer.

Par rapport au module d'enrichissement relationnel proposé, la prise en compte du caractère incertain des jeux de données utilisés pour extraire les règles d'association semblent aussi une problématique intéressante qui pourrait apporter une information additionnelle lors de l'analyse et l'interprétation des règles afin de tenir compte de la fiabilité des résultats obtenus pour la prise de décisions.

Finalement, la mise à jour ou la maintenance de l'ontologie de domaine est un dernier point décisif qui devrait être abordé et étudié afin d'estimer la cohérence et la validité des connaissances en jeux. La vérification de la cohésion et la manière avec laquelle les connaissances sont établies est indispensable, i.e comment les nouvelles connaissances extraites participent à la préservation, actualisation et amélioration des connaissances stratégiques de domaine.

Pour cela, la mise à jour de la base de connaissances doit se faire fréquemment. Ces analyses doivent conduire à une stratégie pour valider les nouvelles connaissances, et supprimer/ou modifier, les connaissances incohérentes.

VI. ANNEXES

1. Annexe 1 : Les mesures de similarité d'ontologie

La similarité sémantique dans une ontologie se réfère à la proximité de deux concepts indiquant la distance qui les sépare par rapport à la structure ontologique. Plusieurs similarités ont été proposées. Dans cette section, on énumère celles les plus connues dans la littérature.

a. La mesure de Rada :

Dans [Rada, et al., 1989], les auteurs proposent la méthode du plus court chemin pour calculer la similarité sémantique entre les concepts en raison de sa simplicité. La mesure est donnée par la fonction suivante :

$$\mathbf{sim}(\mathbf{c_1}, \mathbf{c_2}) = \mathbf{2Max} - \mathbf{L}$$

Où c_1, c_2 sont les deux concepts ; Max est le chemin le plus long séparant c_1 et c_2 ; L est le chemin le plus court séparant c_1 et c_2.

Dans [Sussna, 1993], l'auteur propose une extension de la mesure ci-dessus en introduisant la notion des liens pondérés. Cette pondération dépend de la profondeur de la hiérarchie ainsi de la densité de la taxonomie. La distance est obtenue en additionnant les poids des liaisons traversées. Ces dernières méthodes sont simples et faciles à implémenter, néanmoins, elles tiennent compte seulement des relations hiérarchiques de type 'is-a' sans considération d'autres types de liaison sémantique.

b. La mesure de Wu et Palmer :

Dans [Wu, et al., 1994], la similarité est basée sur les distances N_1 et N_2 qui séparent deux concepts c_1 et c_2 par rapport à la racine et la distance N qui sépare le concept subsumant les concepts c_1 et c_2 de la racine. La distance est définie par la fonction suivant :

$$\mathbf{sim}(\mathbf{c_1}, \mathbf{c_2}) = \frac{\mathbf{2N}}{N_1 + N_2}$$

Cette méthode repose sur le fait que plus les concepts sont profonds dans la hiérarchie ontologique, plus ils sont similaires. Cette distance a été utilisée dans [Wei, et al., 2008] pour le partitionnement 'rigide' d'ontologie. Elle a été utilisée également dans [Hamdi, 2012] pour effectuer le partitionnement d'ontologie autour des ancres qui ont été déterminés par une mesure lexicale. Cependant, en général, les termes de même profondeur n'ont pas nécessairement la même spécificité, et les liens de même niveau ne correspondent pas forcément à la même distance sémantique.

c. *La mesure basée le contenu informationnel :*

Des mesures de similarité sémantiques entre deux concepts c_1 et c_2 ont été proposées en se basant sur le contenu informationnel (IC) de leur concept commun le plus spécifique (c) définie par la formule suivante:

$$\mathbf{IC(c) = -log\ P(c)}$$

où P(c) représente la probabilité d'apparition de c dans un corpus. Le contenu informationnel peut être également calculé en divisant le nombre des instances de (c) par le nombre total des instances. Par conséquence, plus l'information est partagée par deux concepts, plus ces derniers sont similaires. Cependant, cette méthode est seulement appropriée pour la hiérarchie d'ontologie dont les relations sont de même type [Mingxin, et al., 2013]. Dans [Lin, 1998], l'auteur propose une extension de cette méthode en considérant, également, le contenu informationnel des concepts candidats donné par cette formule :

$$\mathbf{sim(c_1, c_2) = \frac{2LogP\ (c)}{Log\ P(c_1)+Log\ P(c_2)}}$$

Cette méthode a été appliquée dans [Havens, 2010] pour le partitionnement de l'ontologie 'Gene Ontology.' Cependant, la considération des concepts d'intérêts provoque une dépendance au niveau d'annotation. Par conséquent, l'exactitude des résultats ne peut être garantie que lorsque les concepts dans l'ontologie sont précisément décrits.

Une autre mesure a été appliquée dans la 'Gene Ontology' dans [Havens, 2010], la méthode considère chaque terme de l'ontologie comme un vecteur des gènes annotés et mesure la similarité entre deux termes en calculant le produit scalaire des deux vecteurs correspondants.

Dans [Tversky, 1977], les auteurs ont proposé une mesure de similarité basée sur les caractéristiques des concepts au sein de la hiérarchie ontologique en tenant compte des caractéristiques communes et disjointes des concepts. Cependant, cette méthode basée sur l'alignement des attributs ne tient pas compte de leurs ordonnancements.

2. Annexe 2 : Exemple de dossier médical (Hôpital Régional de Ben Arous)

Ministère de la santé publique

---***---

Hôpital Régional de Ben Arous
Service de Radiologie

001

DATE DE L'EXAMEN : 7/11/2012 N° DE FICHE : N° DE MATRICULE :
NOM PRENOM :
AGE : 43 ans
TEL : ADRESSE :
SERVICE : MATERNITÉ DISP : AUTRE : .....
MEDECIN :

ATCD :
G 0 P 0 / Poids : ETAT DE GROSSESSE : ALLAITEMENT
AGE de 1ère GROSSESSE :
ALLAITEMENT AU SEIN : NON
OUI DUREE
THS : OUI NON PRECISE : .....
FACTEURS DE RISQUE : Aucun 1er Degré 2ème Degré Mutation Non Précise
RESULTATS ANAPATH ANTERIEURS : NON
OUI Cytoponction Microbiopsie Chirurgie

CLINIQUE :
NORMAL X RESSAUT PLACARD ERYTHEMATEUX RETRACTION ADP AXILLAIRE
NODULE CICATRICE ECOULEMENT AUTRE : Les hétérogène Lésion bilatérale

MAMMOGRAPHIE :
TYPE DE SEIN : BIRADS 1 BIRADS 2 BIRADS 3 BIRADS 4
NORMALE
MASSE OPACITE SURCROIT D'OPACITE.
SIEGE : QII QIE QSI QSE UQI INF UQ INT UQE UQ SUP
Nombre :
Forme : Rond Ovale Irreguliere Lobulee
Contours : Nets Masqués Microlobulés Indistincts Spiculés
Densité : Forte Moyenne Faible Contenu Graisseux.
CALCIFICATIONS : Nom
Typiquement bénigne
Suspecte.
X Foyer de microcalcifications
Distorsion architecturale

CONCLUSION :
A DROITE: ACR 1 ACR2 ACR3 ACR4 ACR5

ECHOGRAPHIE :

Masse : Nombre : Taille :

Localisation : Quadrant Horaire Distance /Mamelon

Forme : Ronde Ovale Irrégulière

Echostructure : Anéchogène Hypoéchogène Isoéchogène Hyperéchogène Mixte

Orientation du grand axe : Horizontal Oblique Vertical

Contours : Nets Microlobulés Spiculés Macrolobules Estompés Indistincts

Limites : Nettes Halo Hyperéchogène

Particularités des Echos postérieurs : Atténuation Renforcement Aucune Aspect Combiné.

Tissu environnant : Dilatation Canalaire Distorsion Architecturale Œdème

Calcifications : Micro Macro Intra Lésionnelles Extra Lésionnelles

Vascularisation : Absente Périphérique Centrale Anarchique

Conclusion échographique : ACR 1 ACR2 ACR3 ACR4 faible ACR 4 fort ACR5

CONCORDANCE : Clinique-Mammo-Echo Clinique-Mammo Clinique-Echo Pas De Corrélation.

MICROBIOPSIE

Aiguille : 14G 15G 16G Mammotome

Contrôle Aiguille : Satisfaisant Non Satisfaisant Non Précise

Nombre de carottes : 1 2 3 4 Non Précise

Disparition de la cible : Complète Incomplète Pas De Modifications

Incidents : non

Cible non atteinte Hématome Saignement

RESULTAT ANAPATH :

P. BENIGNE ADENOFIBROME MFK LESION INFLAMMATOIRE FIBROSE FOCALE AUTRES

P. Maligne CIS CCI GRADE BAS HAUT INTERMEDIAIRE

CLI : AUTRE :

LESION PAPILLAIRE : AUTRE :

RECEPTEURS :

SUIVI :

CHIRURGIE TYPE

HISTOLOGIE DE LA PIECE OPERATOIRE

RADIOTHERAPIE CHIMIOTHERAPIE AUTRE

IRM MAMMAIRE + GADO:

MASSE : Taille :

Localisation : Quadrant Horaire Distance /Mamelon

Forme : Ronde Ovale Irrégulière

Contours : Nets Microlobulés Spiculés Macrolobules Estompés Indistincts

Rehaussement interne : en anneau plein homogène hétérogène septa septa non rehaussée

FOCUS : Taille Nombre Localisation

REHAUSSEMENT NON MASSE : Suspect Non Suspect

REHAUSSEMENT MATERIEL DE FOND : Masquant Non Masquée

COURBE : type 1 type 2 type 3

Conclusion : ACR 1 ACR2 ACR3 ACR4 faible ACR 4 fort ACR5

Ministère de la santé publique
****
Hôpital Régional de Ben Arous
Service de Radiologie

Date : 02/11/2012

Nom: _____ Prénom : _____ Age :

IRM MAMMAIRE

INDICATION

- Tumeur du sein droit.
- Foyers de micro calcifications suspectes du sein gauche.
- Recherche d'autre foyer tumoral.

TECHNIQUES

Acquisitions multi planaire explorant les deux seins sans puis après injection de gadolium.

RESULTATS

*** Sein droit :**

- Deux masses contigües du QSE droit de formes lobulés de contours irréguliers spiculés se rehaussant de façon intense et précoce après injection de gadolinium avec un Wash out tardif mesurant 3 x 1, 45 x 1,35 cm (sans spicules) et 4 x 2,6 x 2,3 cm (spicules comprises).
- Masse contigüe aux deux précédentes du QSE droit mesurant 8 x 4 mm ovalaire de contours réguliers se rehaussant tardivement après injection de produit de contraste avec un hile vraisemblablement graisseux cadrant le plux probablement avec un ganglion intra mammaire.
- Micro kystes prédominants au niveau des quadrants externes.
- Adénomégalies axillaires à centres graisseux d'allure réactionnelle dont la plus volumineuse mesure 8 x 5,8 mm.

***Sein gauche :**

- Absence de masse ou de prise de contraste suspecte.
- Microkystes du QSE.
- Adénomégalies axillaires à centres graisseux d'allure banale.

CONCLUSION

Ministère de la santé publique
****
Hôpital Régional de Ben Arous
Service de Radiologie

COMPTE RENDU D'EXAMEN RADIOLOGIQUE

Date : 03/10/2012

Nom : _____ Prénom : _____ Age : 43 ans

Dossier : Service :

EXAMEN RADIOLOGIQUE ECHOGRAPHIE MAMMAIRE+MAMMOGRAPHIE

RENSEIGNEMENTS CLINIQUES

- 43 ans
- ATCD s néo du sein =0
- Mastodynie

RESULTATS

Mammographie

- Seins graisseux type II de BIRADS.

-Sein droit :

Deux masses contigües du QSE droit denses à contours spiculés :

La première supérieure mesure 7 mm de grand axe avec un foyer de microcalcification en regard la deuxième inférieure mesure 1.4 x1 cm avec un foyer de microcalcifications en regard .

- Microcalcifications éparses .

-Foyer de microcalcifications polymorphes à distribution segmentaire superficielles au niveau du QSE.

-Sein gauche :

- Surcroit d'opacité prepectorale gauche peu dense sans microcalcifications en

3. Annexe3 : Exemple de dossier médical (Hôpital Taher Sfar Mahdia)

Centre Hospitalo-Universitaire Tahar Sfar MAHDIA
Service d'Imagerie Médicale

Date 29/01/2014

Chef de service : Pr Hamza H.

Pr Ag : Dr Jerbi Omezzine S.
M.S.S.P : Dr Bouslah M.
Dr Ben Rhouma K.
Dr Bourogaa S.
Dr Chouchene N.
COOP : Dr Ryglewicz A.

Nom et Prénom : [redacted]

Age : 60 ANS

Service ou consultation : *C GYNECO*

ECHO-MAMMOGRAPHIE
N°3634

Clinique : Examen de dépistage.

Résultats:

Mammographie :
Seins en involution graisseuse.
Absence de microcalcifications groupées en foyers.
Absence d'opacité d'allure suspecte.
Absence d'anomalie du revêtement cutané.
Ganglion axillaire droit.

Echographie :
Légère ectasie canalaire rétro-aréolaire droite anéchogène homogène (0,38 mm de diamètre).
Petite formation kystique à contenu échogène rétro-aréolaire gauche mesurant 5 mm de grand axe.
Absence de masse d'allure suspecte.

Conclusion :
Petite formation kystique à contenu échogène rétro-aréolaire gauche (ACR 3). A contrôler dans 4 mois.
Légère ectasie canalaire rétro-aréolaire droite (ACR 2).

DR JERBI

CENTRE HOSPITALO-UNIVERSITAIRE TAHAR SFAR MAHDIA
SERVICE D'IMAGERIE MEDICALE

Chef de service : Pr Hamza H.

A.H.U : Dr Jerbi Omezzine S.
M.S.S.P : Dr Bouslah M.
Dr Bourogaa S.
Dr Chouchene N.
COOP : Dr Ryglewicz A.

Date : 22/10/2011

Nom et Prénom : [redacted]

Age : 50 ANS

Service ou consultation : C GYNECO

ECHO MAMMOGRAPHIE
N° /11

Clinique : Contrôle de nodule du QII gauche il y a un ATCD : RAS.

Résultats

Mammographie :
Seins denses type 4 de Birads.
Absence de micro calcifications groupées en foyer.
Absence d'opacité d'allure suspecte.
Absence d'anomalie du revêtement cutané.
Creux axillaire libre.

Echographie :
Sein droit
Absence de masse d'allure suspecte.
Absence de désorganisation architecturale.

Sein gauche
Masse infero-interne hypoéchogène atténuante et spiculée, elle mesure 20 mm de grand axe.
Un ganglion axillaire.

Conclusion :
Sein droit : absence d'anomalie écho-mammographique : ACR 1.
Sein gauche : masse du QI interne suspect : ACR 4 : une ponction histologique étant indispensable.

DR ACHOUR/DR ZRIG

VII. Publications

Publications Internationales

- '*Fuzzy Clustering based Approach for Ontology Alignment'* **Rihab Idoudi**, Karim Saheb Ettabaa, Kamel Hamrouni and Basel Solaiman, 18th International conference on Entreprise Information systems, ICEIS2016, 24-28 April 2016, Rome, Italy- *Classe C*

-'*Ontology Knowledge mining based Association Rules Ranking*' **Rihab Idoudi**, Karim Saheb Ettabaa, Basel Solaiman and Kamel Hamrouni, 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom- *Classe B*

- '*Image Based Mammographic Ontology Learning'* Yosra Ben Salem, **Rihab Idoudi**, Kamel Hamrouni, Basel Solaiman, Sana Bousetta, 11th International Conference on intelligent systems: Theories and applications (SITA), 19-20 October 2016, Mohammedia - Morocco

Revues

- '*Ontology Knowledge Mining for Ontology Conceptual Enrichment'* **Rihab Idoudi**, Karim Saheb Ettabaa, Basel Solaiman and Kamel Hamrouni (**en cours de révision**) Knowledge Management Research & Practice, **Springer** ***(IF: 0.595)***

- '*Ontology Knowledge Mining for Ontology Alignment'* **Rihab Idoudi**, Karim Saheb Ettabaa, Basel Solaiman and Kamel Hamrouni, 'International Journal of Computational Intelligence Systems, Vol 9(5): pp 876-887 ***(IF: 0.547)***

-'*Association rules-based Ontology Enrichment'* **Rihab Idoudi**, Karim Saheb Ettabaa, Kamel Hamrouni and Basel Solaiman, International Journal Web Applications, Vol 8(1): pp16-25 (2016)

Publications Nationales

- '*Ontological approach to mammographic knowledge representation*', **Rihab Idoudi**, Karim Saheb Ettabaa, Kamel Hamrouni and Basel Solaiman , IEEE Advanced Technologies for Signal and Image Processing (ATSIP'14), 17-19 March, Sousse, Tunisie, Page(s): 31-34, 2014

-'*Application de l''algorithme de C-moyenne Flou Pour le partitionnement d'Ontologies*', **Rihab Idoudi**, Karim Saheb Ettabaa, Kamel Hamrouni and Basel Solaiman, Traitement et Analyse de l'Information Méthodes et Applications (Taima'15), 11-16 Mai, Hammamet

VIII. References

Adina Branici Représentation et raisonnement formels pour le pronostic basé sur l'imagerie médicale microscropique. Application à la graduation du cancer du sein, Doctoral dissertation, Université de Franche-Comté, 2010.

Agrawal Rakesh, Imielinski Tomasz and Swami Arun Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD'93. - Washington, D.C., USA : ACM, 26-28 mai 1993. - pp. 207–216.

Algergawy Alsayed, Moawed Seham and Sarhan Ameny Improving Clustering-Based Schema Matching Using Latent Semantic Indexing, Large-Scale Data- and Knowledge-Centered Systems. - 2014. - pp. 102-123.

Algergawy Alsayed, Nayak Richi and Gunter Saake Element similarity measures in XML schema matching, Information Sciences. - 2010. - Vol. 180. - pp. 4975-4998.

Ana Cristina, Bicharra Garcia and Inhauma Ferraz From data to knowledge mining, Artificial Intelligence for Engineering Design, Analysis and Manufacturing. - 2009. - 4 : Vol. 23. - pp. 427 - 441.

Baccour Mariem Plateforme de segmentation et de construction de bases d'images mammographiques, Ecole Nationale d'ingénieurs de Tunis. - 2013. - pp. 1-80.

Baesens Bart, Viaene Stijin and Vanthienen Jan Post-processing of association rules, Workshop on Post-Processing in Machine Learning and Data Mining: Interpretation, visualization, integration, and related topics with in Sixth ACM SIGKDD, Int. Conf. on Knowledge Discovery and Data Mining. - 2000. - pp. 20-23.

Baviskar Sagar, Lokhande Biyani and Kabra Bayesian Network Model for Epidemiological Data (Radiation exposure and circulatory disease risk: Hiroshima and Nagasaki atomic bomb survivor data), GJSFR-D: Agriculture and Veterinary. - 2013. - 2 : Vol. 13.

Bellandi Andrea, Furletti Barbara and Grossi Valereo Ontological Support for Association Rule Mining, 26th IASTED International Conference on Artificial Intelligence and Applications. - Innsbruck, Austria : 2008. - pp. 110–115.

Ben Abbes Sarra Construction d'une cartographie de domaine à partir de ressources sémantiques hétérogènes , Doctoral dissertation, Université Paris 13. - 2013. - tel-01146324.

Ben Abbès Sarra, Scheuermann Andreas and Meilender Thomas Characterizing Modular Ontologies, International Conference on Formal Ontologies in Information Systems : Springer-Verlag, July 2012. - pp. 13-25.

Bendaoud Rokia Construction et enrichissement d'une ontologie à partir d'un corpus de textes, Actes des Rencontres des Jeunes Chercheurs en Recherche d'Information (RJCRI'06). - Lyon, 2006. - pp. 353-358.

Bendaoud Rokia, Hacene Amine and Toussaint Yannick Construction d'une ontologie a partir d'un corpus de textes avec l'ACF , Ingénierie des connaissances. - 2007.

Bezdek James Cluster validity with fuzzy sets, Journal of Cybernets. - 1974. - Vol. 3. - pp. 58–73.

Bezdek James Fuzzy Mathematics in Pattern Classification [Report] : Doctoral dissertation. - Cornell University, Ithaca : Applied Math Center, 1973.

Bezdek James Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press. - New York, 1981.

Bing Liu, Hsu Wynne and Wang Ke Visually aided exploration of interesting association rules, Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). - 1999. - pp. 380–389.

Bo Hu, Dasmahapatra Srinandan and Shadbolt Nigel From Lexicon To Mammographic Ontology: Experiences and lessons, International Workshop of Description Logics-DL'03. - 2003.

Bordogna Gloria and Gabriella Pasi Hierarchical-Hyperspherical Divisive Fuzzy C-Means (H2D-FCM) Clustering for Information Retrieval, IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. - 2009. - Vol. 1. - pp. 614-621.

Borst W.N Construction of Engineering Ontologies Doctoral dissertation, University Of Twente. - Netherlands , 1997.

Bottcher Mirko, Ruß Georg and Nauck Detlef From change mining to relevance feedback : A unified view on assessing rule interestingness, Post-Mining of association rules : Techniques for effective knowledge extraction.IGI Global, 2009. - pp. 12–37.

Bouquet Paolo, Giunchiglia Fausto and Harmelen Frank Contextualizing Ontologies, Journal of Web Semantics. - 2004. - 1 : Vol. 1. - pp. 325-343.

Brin Sergey, Motwani Rajeev and Silverstein Craig Beyond market baskets: Generalizing association rules to correlations, Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD '97. - Tucson, Arizona, USA : ACM Press, 13-15 mai 1997. - pp. 265–276.

Bulu Hakan, Alpkocak Adil and Balci Pinar Ontology-based mammography annotation and Case-based Retrieval, Expert Systems with Applications. - 2012. - 4 : Vol. 39. - pp. 11194–11202.

Castano Silavana, Ferrara Alfio and GianPaolo Messa Results of the H-match ontology matchmaker , Workshop on Ontology Matching OM2006. - 2006.

Catarci Tiziiana and Lenzerini Maurizio Representing and using interschema knowledge in cooperative information systems, Journal of Intelligent and Cooperative Information Systems. - 1993. - 4 : Vol. 2. - pp. 375–398.

Chin-Yew Lin and Hovy Eduard The automated acquisition of topic signatures for text summarization, Proceedings of the 18th conference on Computational linguistics. - Stroudsburg,PA, USA : Association for Computational Linguistics, 2000. - Vol. 1. - pp. 495-501.

Choudhary Alok, Hardin Jenny and Tiwari Manoj Kumar Data mining in manufacturing: a review based on the kind of knowledge, Journal of Intelligent Manufacturing. - 2009. - 5 : Vol. 20. - pp. 501–521.

Di Jorio Lisa, Abrouk Lylia and Fiot Céline Enrichissement d'ontologie basé sur les motifs séquentiels, 2007. - lirmm-00176073.

Dou Dejing, Wang Hao and Liu Haishan Semantic data mining: A survey of ontology-based approaches,IEEE International Conference on Semantic Computing (ICSC). - February 2015. - pp. 244-251.

Euzenat Jérome and Shvaiko Pavel Ontology matching, Ontology Management. - 2008. - pp. 177-206.

Euzenat Jerome, Ferrara Alfio and Meilicke Christian Results of the Ontology Alignment Evaluation Initiative 2010, Proceeding of the 5th ISWC Workshop on Ontology Matching (OM-2010). - 2010. - pp. 1-35.

Fabrice Guillet Qualite, Fouille et Gestion des Connaissances [Report] : Habilitation a diriger des recherches / Université de Nantes. - 2010. - pp. 207-216. - tel-00481938.

Fahad Muhammed, Abdul-Qadir Muhammed and Noshairwan Muhammed DKP-OM: A semantic based ontology merger, Proc. 3rd International Conference I-Semantics. - 2007. - pp. 313-322.

Fanizzi Nicola, d'Amato Claudia and Esposito Floriana Fuzzy Clustering for Categorical Spaces, 18th International Symposium, ISMIS 2009 : Springer, September 14-17 2009. - Vol. 5722. - pp. 161-170.

Farah Imed Riadh, Messaoudi Wassim and Saheb Ettabâa Karim, Solaiman, Basel Satellite Image Retrieval Based on Ontology Merging, ICGST-GVIP Journal. - 2008. - 2 : Vol. 8. - pp. 45-53. - 1687-398.

Fareh Meassaouda, Boussaid Omar and Chalal Rachid Merging ontology by semantic enrichment and combining similarity measures, International Journal of Metadata, Semantics and Ontologies. - 2013. - 1 : Vol. 8. - pp. 65-74.

Fauré Clément, Delprat Sylvie and Mille Alain Utilisation des reseaux bayesiens dans le cadre de l'extraction de regles d'association, Proceedings of the French-speaking Conference on Knowledge Discovery and Management. - 2006. - pp. 569–580.

Fayyad Usama, Shapiro Gregory and Smith Padhraic From data Mining to Knowledge Discovey Databases, Advances in knowledge discovery and data mining: AAAI Press P. Smyth Menlo Park, 1996. - pp. 1-30.

Ferraz Inhauma Naves, Cristina Ana and Garcia Bicharra Ontology in association rules, SpringerPlus. - 2013. - 2 : Vol. 452.

Freitas Alex On Objective Measures of Rule Surprisingness, Principles of Data Mining and Knowledge Discovery. - 1998. - pp. 1-9.

Gim Jangwon, Jung Hanmin and Jeong Don-Heon XOnto-Apriori: An Effective Association Rule Mining Algorithm for Personalized Recommendation Systems, Springer Berlin Heidelberg. - Computer Science and its Applications , 2015. - pp. 1131-1138.

Giudici Paolo Applied data mining: Statistical methods for business and industry, John Wiley & Sons Ltd, 2003.

GiunchiGlia Fausto, shvaiko Pavel and Yatskevich Mikalai Discovering missing background knowledge in ontology matching, ECAI. - 2006. - pp. 382-386.

Gruber Thomas A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition. - 1993. - 2 : Vol. 5. - pp. 199-220.

Grüninger Michael and Fox Mark M.S. Methodology for the Design and Evaluation of Ontologies, Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95. - Montreal, 1995.

Guan-yu Li, Shu-peng Liu and Yan Zhao Formal concept analysis based ontology merging method, Computer Science and Information Technology (ICCSIT) 3rd IEEE International Conference. - 2010. - Vol. 8. - pp. 279-282.

Guarino Nicola Some organizing principles for a unified top-level ontology, AAAI Spring Symposium on Ontological Engineering. - 1997. - pp. 57-63.

Hajlaoui Kafil Dispositifs de recherche et de traitement de l'information en vue d'une aide à la constitution de l'information en vue d'une aide à la constitution de réseaux d'entreprises, doctoral dissertation, tel-00770878 / Ecole Nationale Supérieure des Mines de Saint-Etienne. - 2009.

Hamania Mohamed Said, Maamrib Ramdane and Kissoumc Yacine Unexpected rules using a conceptual distance based on fuzzy ontology, Journal of King Saud University - Computer and Information Sciences. - January 2014. - 1 : Vol. 26. - pp. 99-109.

Hamdi Fayçal Améliorer l'interopérabilité sémantique : Applicabilité et utilité de l'alignement d'ontologies, Doctoral dissertation, Université Paris-Sud. - France, 2012.

Hamdi Fayçal, Zargayouna Haifa and Safar Brigitte TaxoMap in the OAEI alignment contest alignment contest, Ontology Alignment Evaluation Initiative (OAEI). - Karlsruhe, Germany, 2008. - pp. 26-30.

Han Jiawei and Kamber Micheline Data mining: concepts and techniques, Morgan Kaufmann Publishers. - San Francisco, CA, 2006. - Vol. 2.

Havens Timothy Clustering in relational data and ontologies, Doctoral dissertation, Missouri, 2010.

Heydari Seyed Taghi, Ayatollahi Sayed Muhamed and Zare Najef Comparison of Artificial Neural Networks with Logistic Regression for Detection of Obesity, Journal of medical systems. - 2012. - 4 : Vol. 36. - pp. 2449-2454.

Hou Xiangdan, Gu Junhua and Shen Xueqin Application of Data Mining in Fault Diagnosis Based on Ontology, Third International Conference on Information Technology and Applications. - Washington, USA, 2005. - pp. 260–263.

Hu Wei, Zhao Yuanyuan and Qu Yuzhong Partition-based block matching of large class hierarchies, Proceedings of Asian Semantic Web Conference- ASWC. - 2006. - pp. 72-83.

Idoudi Rihab, Karim Saheb Ettabaa, Kamel Hamrouni, Basel Solaiman Association rules-based Ontology Enrichment [Journal] International Journal of Web Applications. - 2016. - 1 : Vol. 8. - pp. 16-25.

Idoudi Rihab, Karim Saheb Ettabaa, Kamel Hamrouni, Basel Solaiman L'algorithme de C-moyenne Flou Pour le partitionnement d'Ontologies, Traitement et Analyse de l'Information Méthodes et Applications. - Hammamet, Tunisie , 2015.

Idoudi Rihab, Hamrouni Kamel and Solaiman Basel Ontological approach to mammographic knowledge representation, IEEE Advanced Technologies for Signal and Image Processing (ATSIP'14). - Sousse, Tunisie, 17-19 March 2014. - pp. 31-34.

Kachroudi Marouan, Zghal Sami and Ben Yahia Sadok OntoPart: at the cross-roads of ontology partitioning and scalable ontology alignment systems, International Journal of Metadata, Semantics and Ontologies. - 2013. - 3 : Vol. 8. - pp. 215–225.

Kandpal Ankita, R.H Goudar and Chauhan Rashmi Effective ontology alignment: an approach for resolving the ontology heterogeneity problem for semantic information retrieval, Intelligent Computing, Networking, and Informatics : Springer, 2014. - pp. 1077-1087.

Kiu Ching-Chieh and Lee Chien-Sing OntoDNA: Ontology Alignment Results, Journal of Educational Technology & Society. - 2007. - 16p : Vol. 9. - pp. 27-42.

Kong Hyn, Hwang Myunggwon and Kim Pankoo A new Methodology for Merging the Heterogeneous Domain Ontologies based on the Wordnet, Proceedings of the International Conference on Next generation Web Services Practices, IEEEXplore. - 2005.

Koskinen Kaj U Problem absorption as an organizational learning mechanism in project based companies: Process thinking perspective, International Journal of Project Management. - 2012. - 3 : Vol. 30. - pp. 308–316.

Kotis Konstantinos and Vouros George The HCONE approach to ontology merging, The Semantic Web: Research and Applications Springer Berlin Heidelberg. - 2006. - pp. 137-151.

Levenshtein V.I Binary codes capable of correcting deletions, insertions, and reversals, Technical Report. - 1966. - 8.

Lin Dekang An information-theoretic definition of similarity, Proceedings of the 15th International Conference on Machine Learning. - 1998. - pp. 296–304.

Maedche Alexander and Staab Steffen Axioms are objects too: Ontology engineering beyond the modeling of concepts and relations, Research report 399, Institute AIFB. - Karlsruhe , 2001.

Maedche Alexander and Staab Steffen Measuring Similarity between Ontologies, Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management , ed. Springer-Verlag. - London, UK : Ontologies and the Semantic Web, 2002. - pp. 251-263.

Maedche Alexander and Staab Steffen Mining ontologies from text, 12th European Workshop on Knowledge Acquisition, Modeling and Management. Springer-Verlag. - 2000. - Vol. 1937.

Mahmoodi Sayed Abbas, Mirzaie Kamal and Mahmoud Seyed Mostafa A new algorithm to extract hidden rules of gastric cancer data based on ontology, SpringerPlus, 2016. - 1 : Vol. 5. - pp. 1-21.

Maiz Nora, Fahad Muhammed and Boussaid Omar Automatic Ontology Merging by Hierarchical Clustering and Inference Mechanisms, Proceedings of I-KNOW. - Graz, Austria , September 1-3 2010. - pp. 81-93.

Mansingh Gunjan, Osei-Bryson Kweku-Muata and Reichgelt Han Using Ontologies to Facilitate Post-processing of Association Rules by Domain Experts, Information Sciences 181. - 2011. - pp. 419–434.

Marinica Claudia Association Rule Interactive Post-processing using Rule Schemas and Ontologies - ARIPSO, Doctoral dissertation, Artificial Intelligence. - 2010. - pp. 1-209. - tel-00912580.

Markandey Versha Tracking Medicine Purchase Trends Using Data Mining, IUP Journal of Computer Science. - 2015. - 1 : Vol. 9.

Massmann Sabine, Raunich Salvatore and Aumüller David Evolution of the COMA match system, Ontology Matching. - 2011. - Vol. 49.

McGuinness Deborah, Fikes Richard and Rice James An environment for merging and testing large ontologies, Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference Morgan Kaufmann Publishers. - San Francisco, CA, 2000.

Messai Radja Ontologies et services aux patients : Application à la reformulation des requetes, Doctoral dissertation, Université Joseph Fourier – Grenoble I. - 2009. - pp. 1-153.

Messoudi Wassim and Faleh Imed Riadh Proposition d'une annotation sémantique floue guidée par ontologie pour l'interprétation des images de télédétection: Application à la gestion des risques naturels, Doctoral dissertation, 2013. - p. 157. - 2013telb0236.

Miller George WordNet: a lexical database for English. Communications of the ACM, 1995. - 11 : Vol. 38. - pp. 39-41.

Mingxin Gan, Dou Xue and Rui Jiang From Ontology to Semantic Similarity: Calculation of Ontology-Based Semantic Similarity, The Scientific World Journal. - January 2013. - p. 11. - 793091.

Mohd Izwan Mohd, Hadzic Fedja and Dillon Tharam Interestingness measures for association rules based on statistical validity, Knowledge-Based Systems. Elsevier, 2011. - Vol. 24. - pp. 386–392.

Narayana, Varma and Govardhan An improved technique for ranking semantic associations, International Journal of Web & Semantic Technology. - 2013. - Vol. 4. - pp. 93-106.

Natarajan Rajesh and Shekar A framework for evaluating knowledge-based, Fuzzy Optimization and Decision Making. - 2004. - 2 : Vol. 3. - pp. 157–185.

Nebot Victoria and Berlanga Rafael Finding association rules in semantic web data, Knowledge-Based Systems : Elsevier, 2012. - 1 : Vol. 25. - pp. 51-62.

Neeches Robert, Fikes Richard and Finin Tim Enabling technology for knowledge sharing, AI Magazine. - 1991. - 12 : Vol. 3.

Noy Natalya and Musen Mark PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment, Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00) SMI technical report, 2000.

Otero-Cerdeira Lorena, Rodríguez-Martínez Francisco and Gómez-Rodríguez Alma Ontology matching: A literature review, Expert Systems with Applications. - 2015. - 42 . - pp. 949–971.

Padmanabhan Balaji and Tuzhuilin Alexander A belief-driven method for discovering unexpected patterns, 4th International Conference on Knowledge Discovery and Data Mining. - 1998. - pp. 94 – 100.

Padmanabhan Balaji and Tuzhuilin Alexander Discovery the minimal set of unexpected patterns, Knowledge Discovery and Data Mining. - 2000. - pp. 54–63.

Pasha Maruf and AbdulSattar Building Domain Ontologies From Relational Database Using Mapping Rules, International Journal of Intelligent Engineering & Systems. - 2012. - 1 : Vol. 5.

Petasis Georgios, Karkaletsis Vangelis and Paliouras Georgios Ontology population and enrichment: State of the art, Knowledge-driven multimedia information extraction and ontology evolution. - Springer-Verlag, 2011. - pp. 134-166.

Rada Roy, Mili Hafedh and Bicknell Ellen Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man and Cybernetics. - 1989. - Vol. 19. - pp. 17-30.

Raikwal and Saxena Kanak Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set, International Journal of Computer Applications. - 2012. - Vol. 20. - pp. 35-39.

Raunich Salvatore and Rahm Erhard ATOM: Automated Target-driven Ontology Merging, 27th International Conference on Data IEEEXplorer. - 2011.

Razan Paul, Tudor Groza and Hunter Jane Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain, Journal of Biomedical Semantics. - 2014. - 8 : Vol. 5. - pp. 1-13.

Resnik Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research. - 1999. - Vol. 11. - pp. 95-130.

Ruiz Potes and Andrea Paula Generation de connaissances à l'aide de retour d'eperience: Application à la maintenance industrielle, Doctoral dissertation Université de Toulouse. - Toulouse : Institut National Polytechnique de Toulouse, 2014. - pp. 1-182.

Ruiz Potes, Foguem Bernard and Grabot Bernard Generating Knowledge in Maintenance from Experience Feedback, Knowledge-Based Systems. - 2014. - Vol. 68. - pp. 4-20.

Sampson Jennifer Comprehensive framework for ontology alignment quality, Doctoral dissertation, University of Science and Technology. - Trondheim, Norway 2007.

Sowa John Top-level ontological categories, International Journal of Human and Computer Studies. - 1995. - Vol. 34. - pp. 669-685.

Stoean Ruxandra and Stoean Catalin Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection, Expert Systems Applications. - 2013. - 7 : Vol. 40. - pp. 2677-2686.

Stumme Gerd, Hotho Andreas and Berendt Bettina Semantic web mining : State of the art and future directions, Web Semantics : Science, Services and Agents on the World Wide Web. - June 2006. - 2 : Vol. 4. - pp. 124-143.

Sussna Michael Word sense disambiguation for free-text indexing using a massive semantic network, Proceedings of the 2nd International Conference on Information and Knowledge Management. - Washington, DC, USA, November 1993. - pp. 67-74.

Swartout Bill, Patil Ramesh and Knight Kevin Towards Distributed Use of Large Scale Ontologies, Stanford University . - CA : Spring Symposium Series on Ontological Engineering, 1997.

Taylor Paul and Toujilov Igor Mammographic Knowledge Representation in Description Logic, Springer. - August 2012. - pp. 158-169.

Thangaraju and Mehala Novel Classification based approaches over Cancer Diseases, International Journal of Advanced Research in Computer and Communication Engineering. 2015. - 3 : Vol. 4.

Timon C.Du, Feng Li and King Irwin Managing knowledge on the Web – Extracting ontology from HTML Web, Decision Support Systems. - 2009. - 4 : Vol. 47. - pp. 319-331.

Touma Rizkallah, Romero Oscar and Jovanovic Petar Supporting data integration tasks with semi-automatic ontology construction, Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP : ACM, 2015. - pp. 89-98.

Tversky Amos Features of similarity, Psychological Review. - 1977. - 4 : Vol. 84. - pp. 327-352.

Uschold Mike and King Martin Towards a methodology for building ontologies, Workshop on Basic Ontological Issues in Knowledge Sharing in conjunction with IJCAI-95. - 1995.

Wadii Boulila Extraction de connaissances spatio-temporelles incertaines pour la prediction de changements en imagerie satellitale, Doctoral dissertation, Telecom Bretagne, Universite de Rennes 1. - France : 2012. - tel-00741990.

Wei Hu, Yuzhong Qu and Gong Cheng Matching large ontologies: A divide-and-conquer approach, Data & Knowledge Engineering. - January 2008. - Vol. 67. - pp. 140-160.

Wong Wilson, Liu Wei and Bennamoun Muhammed Ontology learning from text: A look back and into the future, ACM Computing Surveys (CSUR). - 2012. - 4 : Vol. 44.

WordNet WordNet a lexical database for the English language [Online]. - 2009. - http://wordnet.princeton.edu/.

Wu Zhibiao and Palmer Martha Verb semantics and lexical selection, 32nd Annual Meeting of The Association for Computational Linguistics. - Las Cruces 1994. - pp. 133-138.

Zeman Martin, Ralbovsky Martin and Svatek Vojtech Ontology-Driven Data Preparation for Association Mining - In Proceedings of the 8th Znalosti Conference 2009. - pp. 1-12.

Zhang Dawei, Ji Min and Yang Jun A novel cluster validity index for fuzzy clustering based on bipartite modularity, Fuzzy Sets and Systems. - 2014. - Vol. 253. - pp. 122-137.

Résumé

Face à la complexité significative du domaine mammographique ainsi que l'évolution massive de ses données, le besoin de contextualiser les connaissances au sein d'une modélisation formelle et exhaustive devient de plus en plus impératif pour les experts. C'est dans ce cadre que s'inscrivent nos travaux de recherche qui s'intéressent à unifier différentes sources de connaissances liées au domaine au sein d'une modélisation ontologique cible.

D'une part, plusieurs modélisations ontologiques mammographiques ont été proposées dans la littérature, où chaque ressource présente une perspective distincte du domaine d'intérêt. D'autre part, l'implémentation des systèmes d'acquisition des mammographies rend disponible un grand volume d'informations issues des faits passés, dont la réutilisation devient un enjeu majeur. Toutefois, ces fragments de connaissances, présentant de différentes évidences utiles à la compréhension de domaine, ne sont pas interopérables et nécessitent des méthodologies de gestion de connaissances afin de les unifier. C'est dans ce cadre que se situe notre travail de thèse qui s'intéresse à l'enrichissement d'une ontologie de domaine existante à travers l'extraction et la gestion de nouvelles connaissances (concepts et relations) provenant de deux courants scientifiques à savoir: des ressources ontologiques et des bases de données comportant des expériences passées.

Notre approche présente un processus de couplage entre l'enrichissement conceptuel et l'enrichissement relationnel d'une ontologie mammographique existante. Le premier volet comporte trois étapes. La première étape dite de pré-alignement d'ontologies consiste à construire pour chaque ontologie en entrée une hiérarchie des clusters conceptuels flous. Le but étant de réduire l'étape d'alignement de deux ontologies entières en un alignement de deux groupements de concepts de tailles réduits. La deuxième étape consiste à aligner les deux structures des clusters relatives aux ontologies cible et source. Les alignements validés permettent d'enrichir l'ontologie de référence par de nouveaux concepts permettant d'augmenter le niveau de granularité de la base de connaissances. Le deuxième processus s'intéresse à l'enrichissement relationnel de l'ontologie mammographique cible par des relations déduites de la base de données de domaine. Cette dernière comporte des données textuelles des mammographies recueillies dans les services de radiologies. Ce volet comporte ces étapes : i) Le prétraitement des données textuelles ii) l'application de techniques relatives à la fouille de données (ou extraction de connaissances) afin d'extraire des expériences de nouvelles associations sous la forme de règles, iii) Le post-traitement des règles générées. Cette dernière consiste à filtrer et classer les règles afin de faciliter leur interprétation et validation par l'expert vi) L'enrichissement de l'ontologie par de nouvelles associations entre les concepts. Cette approche a été mise en œuvre et validée sur des ontologies mammographiques réelles et des données des patients fournies par les hôpitaux Taher Sfar et Ben Arous.

Les travaux de recherche présentés dans ce manuscrit s'inscrivent dans le cadre de la valorisation et fusion des connaissances issues des sources hétérogènes afin d'améliorer les processus de gestion de connaissances.

Mots clés : Fouille de connaissances, Enrichissement conceptuel d'ontologie, Alignement d'ontologies, Clustering hiérarchique conceptuel flou, Enrichissement relationnel d'ontologie, Extraction de règles d'association, Post-traitement des règles d'association.

Abstract

Facing the significant complexity of the mammography area and the massive changes in its data, the need to contextualize knowledge in a formal and comprehensive modeling is becoming increasingly urgent for experts. It is within this framework that our thesis work focuses on unifying different sources of knowledge related to the domain within a target ontological modeling.

On the one hand, there is, nowadays, several mammographic ontological modeling, where each resource has a distinct perspective area of interest. On the other hand, the implementation of mammography acquisition systems makes available a large volume of information providing a decisive competitive knowledge. However, these fragments of knowledge are not interoperable and they require knowledge management methodologies for being comprehensive. In this context, we are interested on the enrichment of an existing domain ontology through the extraction and the management of new knowledge (concepts and relations) derived from two scientific currents: ontological resources and databases holding with past experiences.

Our approach integrates two knowledge mining levels: The first module is the conceptual target mammographic ontology enrichment with new concepts extracting from source ontologies. This step includes three main stages: First, the stage of pre-alignment. The latter consists on building for each input ontology a hierarchy of fuzzy conceptual clusters. The goal is to reduce the alignment task from two full ontologies to two reduced conceptual clusters. The second stage consists on aligning the two hierarchical structures of both source and target ontologies. Thirdly, the validated alignments are used to enrich the reference ontology with new concepts in order to increase the granularity of the knowledge base. The second level of management is interested in the target mammographic ontology relational enrichment by novel relations deducted from domain database. The latter includes medical records of mammograms collected from radiology services. This section includes four main steps: i) the preprocessing of textual data ii) the application of techniques for data mining (or knowledge extraction) to extract new associations from past experience in the form of rules, iii) the post-processing of the generated rules. The latter is to filter and classify the rules in order to facilitate their interpretation and validation by expert, vi) The enrichment of the ontology by new associations between concepts. This approach has been implemented and validated on real mammographic ontologies and patient data provided by Taher Sfar and Ben Arous hospitals.

The research work presented in this manuscript relates to knowledge using and merging from heterogeneous sources in order to improve the knowledge management process.

Keywords : Knowledge mining, Ontology conceptual enrichment, Ontology hierarchical fuzzy conceptual clustering, Relational ontology enrichment, Association rules extraction, Post-processing of association rules.

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

N° d'ordre : 2017IMTA0005

IMT Atlantique Bretagne-Pays de la Loire - www.imt-atlantique.fr

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 03
T +33 (0)2 29 00 11 11
F +33 (0)2 29 00 10 00

Campus de Nantes
4, rue Alfred Kastler - La Chantrerie
CS 20722
44307 Nantes Cedex 3
T +33 (0)2 51 85 81 00
F +33 (0)2 51 85 81 99

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
T +33 (0)2 99 12 70 00
F +33 (0)2 99 12 70 08