République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université ABBAS LAGHROUR Khenchela

Faculté des Sciences et Technologie

Département d'Informatique

Projet de fin d'étude Master

Domaine : **Informatique**

Filière : **Informatique**

Spécialité : ***sécurité et technologie web***

Thème

Fouille de données sémantique : contribution dans le cadre de la méthode des règles d'association par l'utilisation d'une ontologie

---

Réalisé par :

Ghanem Aissam & Belguidoum Nadjib

Encadré par :

Mr Hemam Mounir

Année Universitaire : 2019/2020


بسم الله الرحمن الرحيم

Remerciement

Nos remerciements vont à l'adresse de tous ceux qui, de près ou de loin, ont aidé à la concrétisation de ce travail.

Nous remercions Dieu pour nous avoir permis d'aller jusqu'au bout de ce mémoire, nos parents pour leurs honnêtes et infaillibles sacrifices.

Nous vaudrions remercier particulièrement notre encadreur, Hemam Mounir, pour le soutien prodigué et lui exprimer notre reconnaissance et toute notre gratitude pour avoir encadré ce travail et pour la confiance qu'il nous a accordée, ses conseils et ses encouragements.

Nous remercions également le président du jury et ses membres pour avoir accepté de présider et prendre part à l'évaluation de ce travail.

Que tous ceux que nous n'avons pas nommément cités trouvent ici l'expression de notre profonde gratitude et notre salut éternel.

Dédicaces

Tous les lettres ne sauraient trouver les mots qu'il faut…

Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la reconnaissance….

Aussi c'est tout simplement que je dédie ce modeste travail :

Je dédie ce modeste travail a la mémoire de Mon père , ce grand homme , qui me manque tant…miséricorde et âme le repos de Dieu dans la paix éternelle

A la prunelle de mes yeux, qui m'a comblé d'amour et de tendresse, qui m'a encouragé et soutenue toute au long de mes études

A toi ma chère mère

A mes chères frères : foudhil , Mouhamed , Rabie , fouad , Sofian ,Lazher

A mes yeux mes sœurs : Chafia et Soumia et Saliha

À mon chère ami : Nacro

A tous mes camarades de classe avec lesquelles j'ai passé Cinque ans de bonheur

A toute ma famille

Dédicaces

On dédie ce mémoire :

À mes chères parents qu'ils méritent tout le bonheur du monde et je leurs dis « Vous avez tout sacrifié pour vous enfants n'épargnant ni santé ni efforts. Vous m'avez donné un magnifique modèle de labeur et de persévérance merci ».

À mes chères sœurs « Asma», « Sara », « Malak » et je leurs dis Je vous aime très fort.

À ma belle famille : ma belle mère que j'aime beaucoup, mon beau père.

À tout membre de ma grande famille : Grand-mère, tantes, oncles, cousins et cousines.

À mes enseignants de la filière Informatique qui m'ont instruit durant le cursus de l'ingéniorat.

À tous ceux qui ont contribué de près ou de loin à l'édition de ce mémoire.

AISSAM Ghanem

ملخص

تقع هذه الاطروحة في التقاء مجالين نشطين للبحث: هندسة المعرفة مع اهتمام خاص في علم الوجود واستخراج البيانات وبصورة ادق اسلوب قواعد الجمعية.

في تقنية التنقيب عن البيانات نحتاج الى توحيد مصطلحات المفردات لمجتمع التنقيب في البيانات. بالإضافة الى ذلك يرتبط التطبيق الفعال لعملية التنقيب في البيانات بالعديد من القرارات الفنية والصعبة بشأن اختيار الخوارزميات والمعلمات والتقييم...لذلك نحن تقدم المساعدة باستخدام الانطولوجيا لتحديد التحديات المذكورة اعلاه.

تم استخدام علم الوجود في مختلف مجالات ابحاث علوم الكمبيوتر بما في ذلك التنقيب عن البيانات. توضح هذه الاطروحة تطوير انطولوجيا المجال لاستخراج البيانات والتي ستضمن تعريفات للكيانات الاساسية لتقنيات التنقيب في البيانات مثل المهام والخوارزميات المستخدمة...لتوحيد شروط تقنيات التنقيب قي البيانات ومشاركة الفهم هذه التقنية شائعة وشرح ما يعتبر ضمنيا.

لإظهار فعالية نهجنا درسنا مجالا فرعيا رئيسيا للتنقيب ي البيانات : قواعد الجمعية من خلال انشاء علم الوجود لمساعدة عامل منجم البيانات خلال المراحل الرئيسية لعملية التعدين. تحتوي الانطولوجيا الخاصة بنا على المعرفة باستخدام البيانات و توفر المصطلحات التي يمكن مشاركتها و معالجتها بواسطة الباحثين في مجال التنقيب عن البيانات.

الكلمات المفتاحية: التنقيب في البياناتِ التنقيب عن البيانات قواعد الاتحادِ الانطولوجيا.

Resumé

Ce mémoire s'inscrit à la confluence de deux domaines actifs de recherche : l'Ingénierie des Connaissances, en s'intéressant particulièrement aux Ontologies et le Data Mining, plus précisément la technique des Règles d'Association.

Dans la technique de Data Mining, nous devons unifier les termes du vocabulaire pour la communauté de Data Mining. En plus, l'application efficace d'un processus de Data Mining est liée à de nombreuses décisions techniques et difficiles sur le choix des algorithmes, des paramètres, de l'évaluation…Par conséquent, nous proposons une assistance en utilisant des ontologies pour relever les défis mentionnés ci-dessus.

Les ontologies ont été utilisées dans divers domaines de recherche de l'informatique, y compris le Data Mining. Ce mémoire décrit le développement des ontologies de domaine pour le Data Mining qui incluront des définitions des entités de base des techniques de Data Mining, telles que les taches, les algorithmes utilisés…pour unifier les termes des techniques de Data Mining et pour partager une compréhension commune de ces technique et d'expliquer ce qui est considéré comme implicite.

Pour montrer l'efficacité de notre approche, nous avons étudié un sous-domaine majeur de Data Mining : les Règles d'Association en créant une ontologie nommé OntoAR pour aider le mineur de données tout au long des phases clés du processus de fouille. Notre ontologie contient les connaissances de Data Mining et fournit des terminologies qui peuvent être partagés et traités par les chercheurs de Data Mining.

Mots clés :Data Mining, fouille de données, Règles d'association, ontologie

Abstract

This memory is concerned with the merging of two active research domains: Knowledge Engineering (KE) with a main interest in Ontology and Data Mining, specifically the technique of Association Rules.

In Data Mining techniques, we need to unify the terminology for the Data Mining process is beset with many difficult and technical decisions about the choice of algorithms, parameters, evaluation…Therefore, we propose assistance by using ontologies for addressing the above-mentioned challenges.

Ontologies have been used in various research areas from computer science, including Data Mining. This memory describes the development of domain ontologies for Data Mining that will include definition of the basic entities of Data Mining techniques, such as tasks,algorithms used…to unify the terms of Data Mining techniques and to share a common understanding of these techniques and explain what is considered implicit.

To demonstrate the effectiveness of our approach, we have studied a major sub-domain of Data Mining: Association Rules by creating ontology named OntoAR to help the data miner throughout the key phases of the mining process. Our ontology contain the knowledge of Data Mining and provide common terminologies that can ne shared and processed by Data Mining researchers.

Key words: Data Mining, domain knowledge, association rules, ontology.

Table des matières

Liste des figures

Les tableaux

Liste des abréviations

APRIORI : Algorithme de recherche des items fréquents et règles d'association

ITEM : Article, attribut, littéral appartenant à un ensemble fini d'éléments distincts

ITEMSET : ensemble d'items.

PRONTODAM : Production **Onto**logie **Da**ta **M**ining

Ras : Règles d**'A**ssociations

OntoAR : **Onto**logy of **A**ssociation **R**ules

OWL: **O**ntology **W**eb **L**anguage

Introduction Générale

Plan

1. **Contexte de l'étude**
2. **Problématique**
3. **Objectif**
4. **Organisation du mémoire**

1. Contexte de l'étude

Cette mémoire s'inscrit à la confluence de deux domaines actifs de recherche : l'Extraction de connaissances à partir des Données (ECD), plus précisément les techniques de Data Mining et l'Ingénierie des connaissances en s'intéressant particulièrement aux ontologies.

Aujourd'hui, nous vivons dans un monde essentiellement défini par les connexions entre des personnes, des appareils et des machines gérées généralement par le biais de l'Internet. Dans ce contexte, des quantités importantes de données sont générées dans différent domaines tels que les données de transaction de supermarchés, les dossiers de crédit d'utilisation des cartes, les détails d'appels téléphoniques. Comme le cerveau humain a des capacités limitées à traiter les données des outils performants ont vu le jour ; .Il s'agit des outils de la fouille de données.

La fouille de données est un domaine qui est apparu avec l'explosion des quantités d'informations stockées, avec le progrès important des vitesses de traitement et des supports de stockage. La fouille de données vise à découvrir, dans les grandes quantités de données, les informations précieuses qui peuvent aider à comprendre les données ou à prédire le comportement des données futures. La fouille de données utilise depuis son apparition plusieurs outils de statistiques et d'intelligence artificielle pour atteindre ses objectifs.

La fouille de données s'intègre dans le processus d'extraction des Connaissances à partir des Données ECD ou (KDD : Knowledge Discovery from Data en anglais).Ce domaine en pleine expansion est souvent appelé le Data Mining.

Nous répertorions plusieurs techniques de Data Mining telles que : la technique des arbres de décision, Réseaux de neurones, Clustering, etc.…

Nous intéressons dans cette mémoire sur la technique des Règles d'Association(RA).Cette technique a pour but de découvrir des tendances implicatives parmi les items d'une base de données transactionnelle.

2.Problématique

Nous avons pensé à utiliser les Ontologies qui sont utilisées pour formaliser un modèle de représentation des connaissances et elles définissent un vocabulaire commun pour les chercheurs qui ont besoin de partager l'information dans un domaine.

Dans ce mémoire, nous allons créer un nouveau modèle d'ontologie : OntoAR(Ontology of Association Rules) pour unifier le vocabulaire et pour représenter respectivement les connaissances de la technique des Règles d'Association afin de représenter les concepts de base et de définir la sémantique des relations entre les entités de base de cette technique.

Notre Ontologie (OntoAR) vise à être une Ontologie de référence pour cette technique de Data Mining.Nous essayons de développer une représentation des entités clés de cette technique aussi générique que possible. Par conséquent, nous espérons de contribuer à l'état de l'art de DM par le développement de cette Ontologie de domaine OntoAR afin que nous puissions représenter les concepts de base de cette technique (taches,algorithmes,étapes…) pour guider les utilisateurs de cette technique de DM.

En bref, les raisons de la construction de notre nouvelle ontologie(OntoAR) sont :

- Décrire la technique des Règles d'Association. Cette Ontologie incluse des définitions des entités de base de cette technique, telles que les taches, les algorithmes,etc… pour partager une compréhension commune de ces méthodes et expliquer ce qui est considéré comme implicite.

- Structurer et organiser le domaine Data Mininig qui se caractérise par un volume de données très important.
- Partager une structure de compréhension commune des informations entre les personnes ou les logiciels.
- Crée un support et un guide pour les chercheurs,les mineurs de données et surtout les débutant de cette technique de DM.

Donc, la principale question à laquelle nous essayons de répondre dans cette mémoire est alors : Comment les ontologies peuvent contribuer au domaine de Data Mining, en particulier la technique des Règles d'Association.

3 . Organisation du mémoire

Ce mémoire est organisé de la façon suivante :

Ce premier chapitre courant introduit la notion des ontologies, leurs objectifs, leurs outils de modélisation, ainsi que leur utilisation dans la fouille de données.
Le chapitre 2 introduit la définition du Data Mining son domaine d'application, ses taches et ses outils. Un plus détail est donné sur la technique d'extraction des règles d'association : définition, application, processus d'extraction des règles d'association ainsi que l'algorithme APRIORI et ses variantes.
Le chapitre 3 décrit la conception de notre ontologie avec la methondologie de Fernandez
Le chapitre 4 fait l'état de l'art de l'emploi des ontologies dans le processus de la fouille de données pour filtrer les règles utiles. Ainsi, une synthèse sur les différentes étapes proposées est définie. Une proposition est faite comme contribution dans ce processus.
Le dernier chapitre clôture ce mémoire avec une conclusion générale.

Chapitre1

Ingénierie Ontologique

Plan du chapitre

1.Introduction

Pour les systèmes d'information des entreprises, l'objectif principal est d'extraire à partir de leurs bases de données, des connaissances vitales et utiles pour un emploi éventuel dans leur système d'aide à la décision (Business intelligence).

Après avoir bien choisi puis appliquer la méthode et la technique appropriée aux bases de données et à la nature du champ d'application étudié, un volume important d'informations est généré. Le traitement et l'analyse de ce volume d'informations deviennent un vrai problème à résoudre, car d'une part le volume est important pour l'analyser manuellement, et d'autre part il nécessite l'intervention d'un expert du domaine pour le choix des informations intéressantes et utiles.

Comme exemple, la technique des règles d'association permet de détecter tous les liens et corrélations possibles entre les données. C'est une technique non supervisée qui peut être descriptive ou prédictive, d'où le nombre très important des règles générées. Certaines de ces règles peuvent être intéressantes, mais certaines autres peuvent être triviales ou inutiles car elles proviennent de la particularité de l'ensemble d'apprentissage. Le nombre important des règles générées nécessite un post-traitement pour choisir manuellement celles qui sont utiles. Ceci nécessite beaucoup de temps en plus de l'expertise demandée. L'utilisateur final se retrouve noyé dans ce nombre important de règles, et donc peut ne pas bénéficier des résultats obtenues.

Les dernières recherches ont proposé le concept des ontologies pour compléter le processus du Data Mining, et en servir comme post-traitement pour une fouille ciblé des informations, tout en se basant sur les connaissances acquises des experts dans les différents domaines de métiers.

L'emploi des ontologies est devenu ainsi une solution efficace pour résoudre plusieurs problématiques dans le web sémantique, le commerce électronique (E-commerce), l'intelligence artificielle, ainsi que la fouille de données.
Dans ce chapitre, nous allons présenter la notion d'ontologie, ses objectifs, les langages et les outils utilisés ainsi que les concepts liés.

2. Définitions d'une ontologie

Il existe plusieurs définitions d'ontologies, nous citons ici certaines d'elles.

Définition 1 : [FG & al-10]

Une ontologie est un ensemble de connaissances statiques concernant un domaine qui peut être utilisé efficacement par les agents (personne ou agent virtuel). Une ontologie est une spécification explicite et formelle d'une conceptualisation partagée.

Définition 2 : [WP]

En philosophie, le mot grec ontologie est l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe. Par analogie, le terme est repris en informatique et en science de l'information, où une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que les relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné.

Définition 3 : (choisie)

Une ontologie représente les connaissances d'un expert pour un domaine donné. Une ontologie permet donc de définir une base conceptuelle concise pour le partage des connaissances des experts. Cette base conçue offre aux utilisateurs

la possibilité de comprendre, définir, structurer et standardiser la sémantique des termes et concepts représentés dans les données du domaine traité.
L'acquisition des connaissances d'un domaine pour construire une ontologie n'est pas une tache facile, elle est très coûteuse en temps et d'effort. C'est pour cela que plusieurs méthodes et techniques ont été développées pour réduire cet effort. Pour une entreprise, l'ontologie représente sa mémoire pour les domaines d'intérêts, car elle définit un **vocabulaire** de référence pour cette entreprise.

Définition 4 : (formelle) [JF & al-07]

Soit C un ensemble de concepts, T un ensemble de termes, Rc un ensemble de relations entre concepts, Rt un ensemble de relations entre termes.

L'ontologie O est définie par le tuple O = {C, T, Rc, Rt} tel que :

Rc : C × C est la relation d'ordre partiel sur C définissant la hiérarchie entre les concepts,

Rc (c1, c2) signifie c1 est plus général que c2.

Rt : C → T est la fonction d'association d'un terme préféré à un concept.

Pour désigner un concept de l'ontologie, on peut utiliser l'un de ses termes associés. Ce terme sera alors le terme préféré de ce concept.

3. Graphe d'ontologie

D'une façon formelle, l'ontologie est un ensemble de concepts reliés entre eux par des relations de subsomption ou de conceptualisation : le concept parent est une généralisation du concept fils englobé, tandis que les concepts fils sont des spécialisations de leur concept parent. L'hiérarchie ainsi obtenue est une représentation conceptuelle des connaissances.

Les concepts sont organisés dans un **graphe** dont les relations peuvent être :
• des relations sémantiques

• des relations de subsomption (inclusion).

Les ontologies sont employées comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde. Les ontologies décrivent généralement :

• Individus : les objets de base.

• Classes : ensembles, collections, ou types d'objets.

• Attributs : propriétés, fonctionnalités, caractéristiques ou paramètres que les objets peuvent posséder et partager.

• Relations : les liens que les objets peuvent avoir entre eux.

• Evénements : changements subis par des attributs ou des relations.

4. Objectifs d'une ontologie

Le but d'une ontologie est de définir un vocabulaire pour décrire un domaine, si possible de manière complète ; ni plus, ni moins. Contrairement aux bases de connaissances par exemple, on n'attend pas d'une ontologie qu'elle soit en mesure de fournir systématiquement une réponse à une question arbitraire sur le domaine. Une ontologie est la théorie la plus faible couvrant un domaine ; elle ne définit que les termes nécessaires pour partager la connaissance liée à ce domaine.
Donc si on résume, les principaux objectifs d'une ontologie sont :

- Modéliser un ensemble de connaissances des experts du domaine donné qui peut être réel ou imaginaire, cette modélisation sert pour une utilisation future par d'autres personnes.

- Fournir un vocabulaire commun pour les chercheurs qui ont besoin de partager les informations dans un domaine.

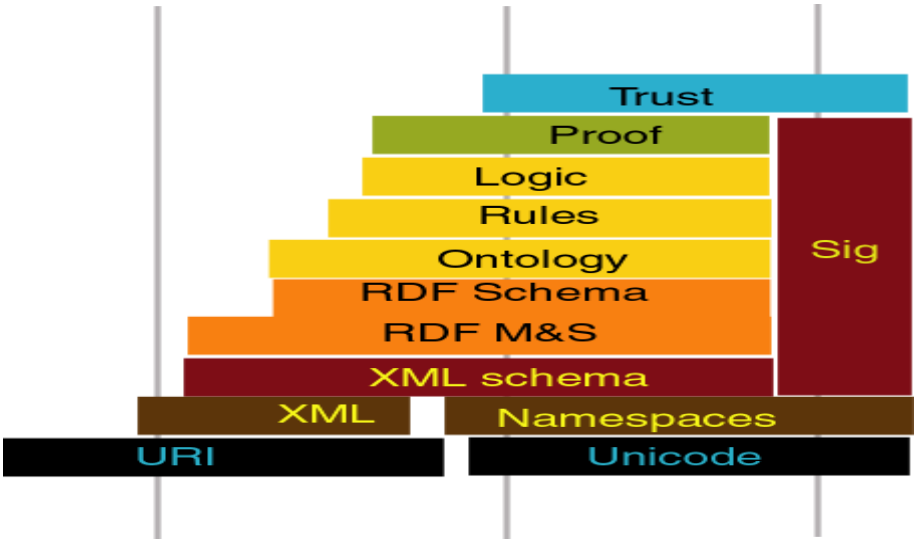
- Offrir à l'utilisateur la possibilité de comprendre, définir, structurer et standardiser la sémantique des termes et concepts représentés dans les domaines traités.
- Fournir un moyen efficace pour filtrer toute nouvelle connaissance à vérifier ou à valider dans les domaines d'intérêts.

Ce dernier point est important, car il fait appel à l'ontologie dans le processus du datamining pour sélectionner et filtrer les connaissances collectées (règles associations, patterns, etc.) intéressantes, vu le volume important des données générés du processus du datamining.

5. Les domaines d'applications des ontologies

Les ontologies sont employées dans

- L'intelligence Artificielle.
- Le Web sémantique.
- Le génie logiciel.
- L'informatique Biomédicale.
- L'architecture du système d'information.
- Le domaine du Data Mining et gestion des connaissances (Knowledge Management).



- Figure I.1 **:** Pyramide du Web sémantique.

6.Composantes d'une ontologie

Comme nous l'avons abordé, les ontologies fournissent un vocabulaire commun d'un domaine et définissent la signification des termes et des relations entre elles. La connaissance dans les ontologies est principalement formalisée en utilisant les cinq types de composants [12] à savoir : **concepts** (ou classes), **relations** (ou propriétés), **fonctions, axiomes** (ou règles) et **instances** (ou individus).

- **Les concepts :** aussi appelés termes ou classe de l'ontologie, correspondent aux abstractions pertinentes d'un segment de la réalité (le domaine du problème) retenus en fonction des objectifs qu'on se donne et de l'application envisagée pour l'ontologie.
- **Les relations :** traduisent les associations (pertinentes) existant entre les concepts présents dans le segment analysé de la réalité. Ces relations incluent les associations suivantes :
  - Sous classes de (généralisation-spécialisation)
  - Partie de (agrégation ou composition)
  - Associe à
  - Instance de, etc.

  Ces relations nous permettent d'apercevoir la structuration et l'interrelation des concepts, les uns par rapport aux autres.

- **Les fonctions :** constituent des cas particuliers de relations, dans laquelle un élément de la relation, (le nième) est défini en fonction des N-1 éléments précédents.
- **Les axiomes :** constituent des assertions, acceptées comme vraies, à propos des abstractions du domaine traduites par l'ontologie.
- **Les instances :** constituant la définition extensionnelle de l'ontologie ; ces objets véhiculent les connaissances (statiques, factuelles) à propos du domaine du problème.

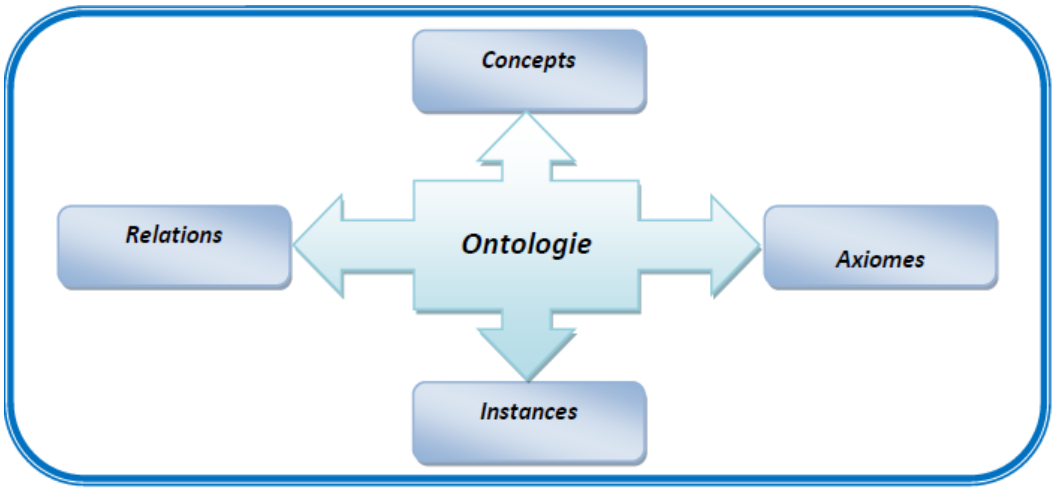


Figure I.2: Les Composants de l'ontologie

7. Cycle de vie d'une ontologie [B-04]

La construction d'une ontologie nécessite tout d'abord la définition des objectifs et utilisation de scénarios identiques à ceux utilisés en génie logiciel, puis la collecte des données par des entretiens informels et structurés avec les experts. Ensuite, l'étude sémantique pour normaliser les termes et éviter les définitions circulaires, ceci permet de résoudre les ambiguïtés. Par la suite, viendra la création des concepts et d'une taxonomie (relation entre ces concepts) qui est le cœur du travail réel. En fin la formalisation ciblée.

Plusieurs propositions ont été faites pour modéliser le processus d'évolution d'une ontologie.

Voici une proposition (faite par l'équipe ACACIA de l'INRIA) comportant les étapes Suivantes :

- Planification

- Spécification

- Acquisition des connaissances

- Conceptualisation

- Formalisation

- Intégration

- Implémentation

- Evaluation et Maintenance

Il est à noter que :

- ✓ si une ontologie est devenue importante, son processus de création doit être considéré comme étant un projet séparé.
- ✓ une ontologie appliquée à un domaine qui évolue, est appelée aussi à évoluer et sa maintenance devient une tâche vitale [UG-96]. Ceci est nécessaire pour que l'ontologie soit toujours à jour.

Certains chercheurs proposent même l'utilisation des techniques du Data Mining pour enrichir les ontologies. Ceci permet dans les textes (web, blogs, …) de mettre en évidence des schémas fréquents sous forme de motifs séquentiels et détecter les relations sémantiques qui peuvent exister entre eux [JF & al-07].

Cette technique est très bénéfique pour la construction des moteurs de recherche. C'est exactement l'inverse de notre travail qui consiste à utiliser les ontologies dans le Data Mining.

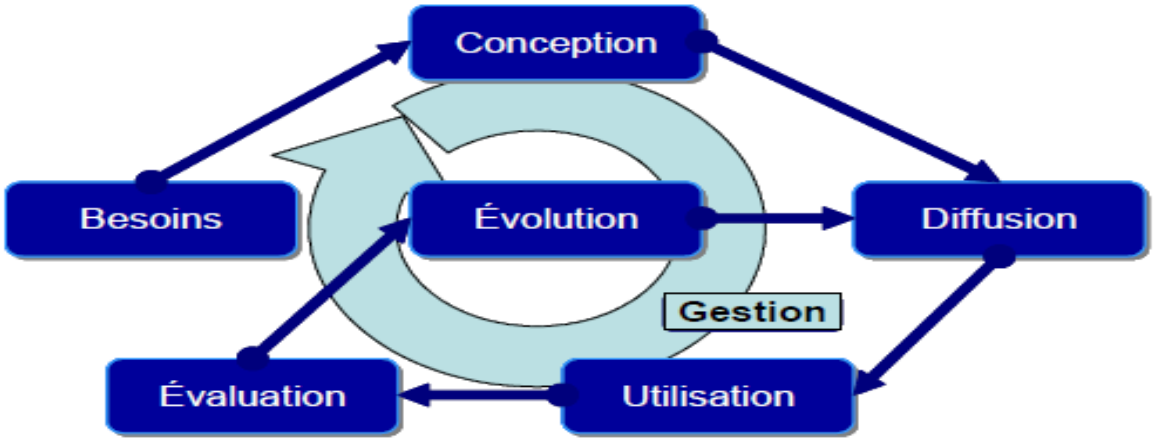


Figure I.3: Cycle de vie d'une ontologie

8. Critères d'évaluation d'une ontologie

D'après Thomas Gruber [G-93], cinq critères permettent de mettre en évidence des aspects importants d'une ontologie :

•**La clarté :** La définition d'un concept doit faire passer le sens voulu du terme, de manière aussi objective que possible (indépendante du contexte). Une définition doit de plus être complète (c'est-à-dire définie par des conditions à la fois nécessaires et suffisantes) et documentée en langage naturel.

• **La cohérence :** Rien qui ne puisse être inféré de l'ontologie ne doit entrer en contradiction avec les définitions des concepts (y compris celles qui sont exprimées en langage naturel).

•**L'extensibilité :** Les extensions qui pourront être ajoutées à l'ontologie doivent être anticipées. Il doit être possible d'ajouter de nouveaux concepts sans avoir à toucher aux fondations de l'ontologie.

Une déformation d'encodage minimale : Une déformation d'encodage a lieu lorsque la spécification influe la conceptualisation (un concept donné peut être plus simple à définir d'une certaine façon pour un langage d'ontologie donné,

bien que cette définition ne corresponde pas exactement au sens initial). Ces déformations doivent être évitées autant que possible.

9. Langages des ontologies

Le langage de spécification est l'élément central sur lequel repose l'ontologie. La plupart de ces langages se basent sur la logique du premier ordre, et représentent donc les
connaissances sous forme d'assertion (sujet, prédicat, objet). Parmi les formalismes les plus employés se basant sur la logique des prédicats, on retrouve des langages comme N3 ou N-Triple. On peut aussi évoquer le langage DEF-*.

Bien que développé pour la représentation des vocabulaires contrôlés et structurés (thésaurus), SKOS peut être utilisé pour élaborer et gérer des ontologies légères multilingues.

Par ailleurs, dans le cadre de ses travaux sur le Web sémantique, le W3C (World Wide Web Consortium) a mis en place en 2002 un groupe de travail dédié au développement de langages standards pour modéliser des ontologies utilisables et échangeables sur le Web. S'inspirant de langages précédents comme DAML+OIL et des fondements théoriques des logiques de description, ce groupe a publié en 2004 une recommandation définissant le langage **OWL** (Web Ontology Language), fondé sur le standard RDF et en spécifiant une syntaxe XML. Plus expressif que son prédécesseur RDFS, OWL a rapidement pris une place prépondérante dans le paysage des ontologies et est désormais, de facto, le standard le plus utilisé.

Le langage OWL est le dernier standard développé pour les langages d'ontologie. Une ontologie OWL peut contenir des descriptions de classes, propriétés et leurs instances. La sémantique formelle d'OWL pour une telle ontologie, spécifie comment se déroule la logique de ses conséquences. Les faits ne sont pas présentés littéralement dans cette description mais ils sont déduits par les sémantiques.

Il existe dans la littérature, des outils qui permet de faire la transition directe entre le formalisme bases de données et OWL par l'import des structures et données (tables) à partir des bases de données relationnelles et les transformes en concepts OWL. Nous citons comme exemple : Data Master (plug-in de Protégé créé par Nyulas 2007), et KAON2 (développé par l'université de Karlsruhe), RDBToOnto (développé en 2008 dans le cadre du projet européen TAO Transitioning Applications to Ontologies). [KN & al-07]

10. Outils de travail avec des ontologies

Les éditeurs d'ontologie suivants sont gratuits et téléchargeables

• **Protégé** est le plus connu et le plus utilisé des éditeurs d'ontologie. Un plus de détail sur Protégé est présenté par la suite.

•**SWOOP** est un éditeur d'ontologie développé par l'Université du Maryland dans le cadre du projet MINDSWAP. Contrairement à Protégé, il a été développé de façon native sur les standards RDF et OWL, qu'il prend en charge dans leurs différentes syntaxes (pas seulement XML). C'est une application plus légère que Protégé, moins évoluée en termes d'interface, mais qui intègre aussi des outils de raisonnement.

• **KMgen** est un éditeur d'ontologie pour le langage KM (Knowledge Machine).

Avec l'émergence du marché des technologies du Web sémantique, on peut noter l'apparition depuis 2005 d'outils logiciels proposés par des éditeurs commerciaux. On peut citer:

•**SemanticWorks** fait partie de la suite d'outils XML développée par Altova. Il supporte le langage OWL à travers sa syntaxe XML.

•**TopBraid** Composer est développé par Top_Quadrant. Son interface et ses fonctionnalités ressemblent beaucoup à celles de Protégé (le développeur principal de TopBraid étant l'ancien développeur des extensions OWL de Protégé).

• **Ontology Craft Workbench** développé par l'équipe Condillac "Ingénierie des Connaissances" de l'Université de Savoie. Les ontologies sont disponibles aux formats XML et OWL. OCW est utilisé par la société Ontologia.

Il existe d'autre part des outils informatiques permettant de construire une ontologie à partir d'un corpus de textes. Ces outils parcourent le texte à la recherche de termes récurrents ou définis par l'utilisateur, puis analysent la manière dont ces termes sont mis en relation dans le texte (par la grammaire, et par les concepts qu'ils recouvrent et dont une définition peut être trouvée dans un lexique fourni par l'utilisateur). Le résultat est une ontologie qui représente la connaissance globale que contient le corpus de texte sur le domaine d'application qu'il couvre. Le projet WordNet est l'exemple le plus important donné de ces outils.

11.L’outil Protégé [PS]

Protégé est un éditeur d’environnement intégré (EDI), gratuit et open-source, développé par l'université Stanford, il est devenu actuellement l’outil le plus utilisé des éditeurs d'ontologie et framework des connaissances.

Il a été évolué depuis ses premières versions (Protégé-2000) pour intégrer à partir les standards du Web sémantique et notamment OWL. Il offre de nombreux composants optionnels : raisonneurs, interfaces graphiques

La plateforme Protégé supporte deux façons principales pour modéliser les ontologies en utilisant les éditeurs Protégé-Frames et Protégé-OWL. Les ontologies générées de Protégé peuvent être exportées sous plusieurs formats tels que RDF(S), OWL, et schémas XML.

Protégé est basé sur langage Java, il est extensible, et fourni des environnent Plug-and--play qui le rend flexible et donc bien adapté au prototypage et le développement rapide d’application(RAD).

Protégé est supporté par une forte communauté de développeurs et académiques, gouvernements et utilisateurs, qui utilisent Protégé pour leurs solutions de connaissance dans les différents domaines tels que la biomédicale, la collecte intelligence, et la modélisation.

Il existe actuellement plusieurs versions de Protégé, les plus utilisées sont :

- Les séries Protégé 3.x développé pendant 3 années.
- Protégé 4.0 (dernière release 4.0.2 apparu en Juin 2009).
-Protégé 5.0 (dernière version 5.5 apparu en 14 Mars 2019)

Protégé peut être installé dans toutes les plateformes système : MacOSX, Microsoft Windows

(32/64 bits), Linux, Unix, Sun Solaris, HPUX, IBM-AIX, ou toute autre plateforme supportant Java.

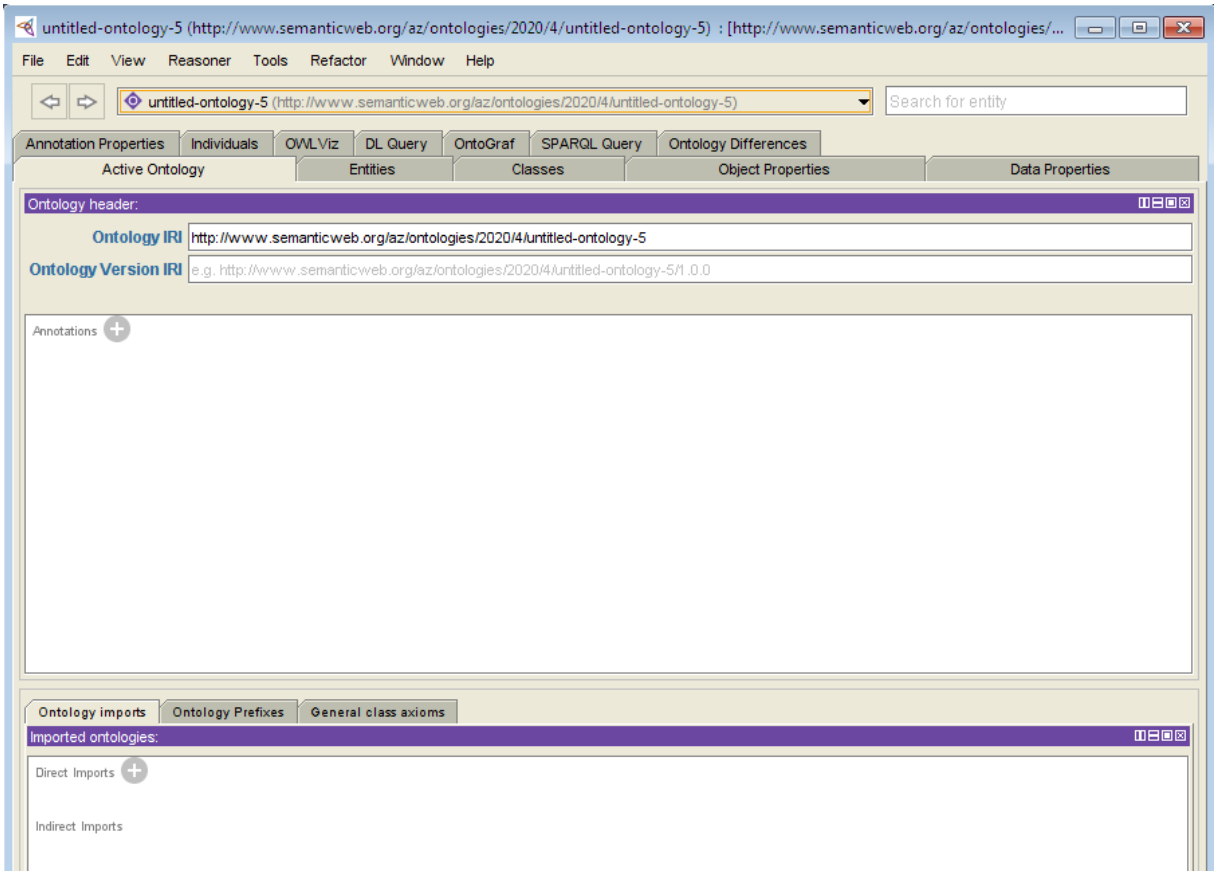


Figure I.4 : L'éditeur d'environnement intégré Protégé

11.1. L'éditeur protégé-owl [PS]

L'éditeur Protégé-OWL est une extension de Protégé qui supporte le langage OWL. Cet éditeur permet aux utilisateurs de :

• Charger et sauvegarder des ontologies OWL et RDF.

• Editer et visualiser les classes, propriétés, et les règles SWRL.

•Définir les caractéristiques logiques de class par les expressions OWL.

•Exécuter le raisonneur (reasoner) tel que le classificateur de description logique.

•Editer les individus OWL pour un marquage Web Sémantique.

L'architecture flexible de Protégé-OWL l'a rendu facile à configurer, extensible et intégré avec d'autres outils. Il a une API Java open-source pour le développement des composants d'interface utilisateur ou des services Web sémantique.

12. Conclusion

Les ontologies permettent de définir une base conceptuelle concise pour le partage de connaissances expertes. Elles offrent la possibilité de définir, structurer et standardiser la sémantique des termes et concepts d'un domaine d'intérêt. Les ontologies sont largement utilisées afin de fournir des annotations sémantiques de n'importe quelle source d'information, ce qui les rend très efficace dans le domaine du web sémantique, les systèmes d'informations médicales et la fouille de données.

Dans le chapitre suivant, nous allons voir les règles d'association générées dans le processus du Data Mining fournissent un grand nombre de règles, l'analyse manuelle de cet ensemble est très fastidieux, de ce fait l'appel à l'ontologie dans cette étape est très bénéfique afin d'effectuer une analyse automatique en se servant des connaissances incluses dans l'ontologie

Chapitre 2

Data Mining et règles d'association

Plan du chapitre2

1. Le Data Mining

1.1. Introduction

Durant les dernières années on assiste à une croissance importante des moyens de génération et de collection des données. Ceci est dû principalement à l'évolution de la technologie des supports de stockage, leurs capacités, et la réduction considérable de leurs coûts. Du fait de l'informatisation rapide des entreprises, des administrations, du commerce, des télécommunications, la quantité de données disponibles augmente très rapidement. Cependant, l'analyse et l'exploitation de ces données restent très difficiles.

Les modèles classiques de recherche d'informations ne sont pas adaptés pour traiter des masses gigantesques de données, souvent hétérogènes « Beaucoup de données et aucune information en retour ». C'est ce constat qui a permis au concept de Data Mining d'émerger et de vulgariser les méthodes d'analyse. [C-06]

Il s'est ainsi créé un besoin d'acquisition de nouvelles techniques et méthodes intelligentes de gestion qui permettent d'extraire des données, des informations utiles appelées connaissances. C'est ainsi que l'on a commencé à parler de découverte de connaissances à partir de données (KDD) ou encore de Data Mining ou de fouille de données. Ces masses de données contiennent des connaissances d'une grande valeur commerciale ou scientifique.

Le processus d'extraction de connaissances ne se limite donc pas à une extraction automatique, il comporte plusieurs étapes pendant lesquelles l'expert humain a un rôle important. Il faut tout d'abord récupérer les données qui peuvent être issues de plusieurs sources différentes et les mettre dans un format commun pour pouvoir les fusionner. Il faut ensuite les préparer, par exemple résoudre le problème posé par les valeurs manquantes ou aberrantes, et sélectionner les données sur lesquelles va être appliqué l'algorithme d'extraction de connaissances proprement dit. Ensuite, il faut post-traiter les résultats de cet

algorithme et les interpréter. Ce post-traitement peut consister à sélectionner les résultats les plus prometteurs, à les trier ou à vérifier leur pertinence à l'aide d'outils statistiques. [K-07]
Dans ce chapitre, on va présentée le processus complet d'extraction de connaissances à partir de données ainsi que les domaines d'applications, les tâches et les techniques du Data Mining.Ensuite on va focaliser sur une technique très répandue qui est les règles d'association. Un plus détail est donné sur cette technique et leurs différents algorithmes.

La technique d'extraction des règles d'association a attiré le plus l'attention des chercheurs qui l'ont adopté pour effectuer leurs travaux parmi les autres méthodes de fouille de données. Ceci est justifié par le fait que cette technique permet la découverte de règles intelligibles et exploitables dans un ensemble de données, tout en exprimant des associations entre les données.
L'objectif du processus d'extraction est de découvrir des règles (relations) significatives entre attributs extraits des bases de données, dont le support et la confiance sont au moins égaux à des seuils minimaux définis par l'utilisateur.
Dans cette section, nous allons décrire les domaines d'applications, le processus de l'extraction des règles d'association. Par la suite, nous présenterons l'algorithme générique de l'extraction des règles d'association tout en s'attardant sur l'algorithme de base APRIORI, puis un survol sur les autres algorithmes, et enfin une conclusion.

1.2. Définition du Data Mining

Le Data Mining (en français fouille de données) est un néologisme américain qui signifie "recherche de pépites d'information utile dans un grand ensemble de données". [K-07]

Le terme de Data Mining signifie littéralement forage de données. Comme dans tout forage, son but est de pouvoir extraire un élément : la connaissance. Ces concepts s'appuient sur le constat qu'il existe au sein de chaque entreprise des

informations cachées dans le gisement de données. Ils permettent, grâce à un certain nombre de techniques spécifiques, de faire apparaître des connaissances. [K-07]

Donc, le Data Mining a pour objet l'extraction d'un savoir à partit de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Le Data Mining repose sur un ensemble de fonctions mais aussi sur une méthodologie de travail (voir le point 1.4).

On pourrait définir le Data Mining comme une démarche ayant pour objet de découvrir des

relations et des faits, à la fois nouveaux et significatifs, sur de grands ensembles de données. On devrait ajouter que la pertinence et l'intérêt du Data Mining sont conditionnés par les enjeux attachés à la démarche entreprise, qui doit être guidée par des objectifs directeurs clairement explicités (améliorer la performance commerciale, fidéliser la clientèle, ... etc.).

Nous appellerons Data Mining l'ensemble des techniques qui permettent de transformer les données en connaissances. Les techniques de Data Mining permettent de découvrir des informations importantes cachées dans les données. Elles permettent par exemple de trouver les tendances qui se dégagent dans les ventes d'un supermarché. La connaissance de telles informations peut permettre au supermarché d'élaborer des stratégies commerciales ou de marketing en direction de ses clients.

Le Data Mining fait appel à un lot de méthodes issues de la statistique, de l'analyse des données, de la reconnaissance de formes, de l'intelligence artificielle (connexionisme, réseaux de neurones) ou de l'apprentissage automatique.

Il faut noter que le Data Mining a une approche très différente de la méthode statistique qui exige qu'on se fixe une hypothèse qui sera confirmée ou non par l'analyse des données. Rien de tel avec le Data Mining qui, au contraire, fait émerger à partir des données brutes des réponses à des questions. [C-06]

Le consortium CRISP-DM qui regroupe des membres spécialisés du Data Mining ont défini ce concept comme étant tout sauf une opération improvisée, et ils ont tous adopté la méthodologie de développement présentée dans le point 1.4 de ce chapitre. [CRISP]

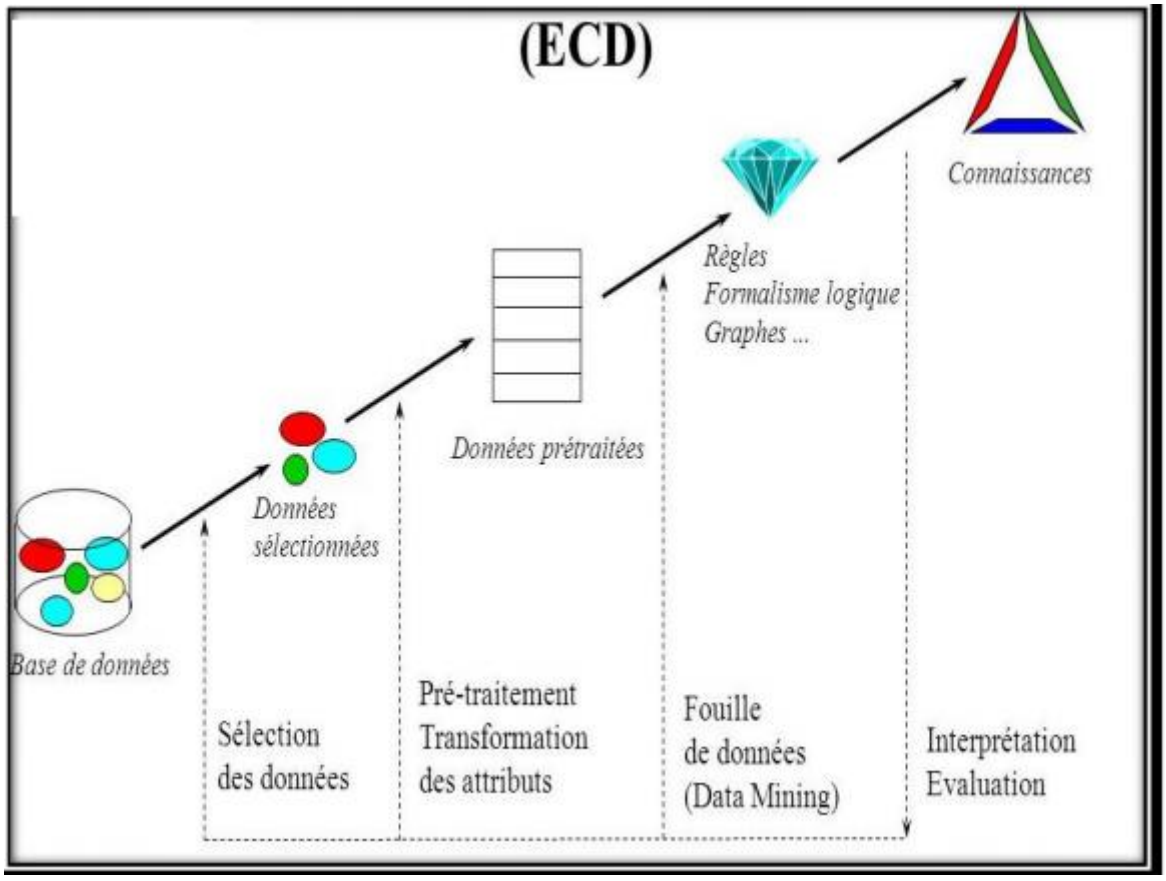


Figure II.1 : Etapes du processus d'ECD

1.3. Les domaines d'applications du Data Mining

Avec la performance des systèmes informatiques actuels et la maturité des méthodes d'apprentissage automatique, le Data Mining est devenu très attrayant dans de nombreux domaines d'applications : médecine, génétique, astronomie, processus industriels, agriculture ou encore la gestion de la relation client, et la production industrielle, … etc.

Les entreprises ont mis en œuvre ces outils pour améliorer leur connaissance afin de mieux les servir et augmenter leur satisfaction et leur fidélité, pour enfin augmenter leur rentabilité. Les principaux secteurs économiques utilisant ces techniques sont le secteur financier (banques et assurances), les

télécommunications ainsi que les entreprises de grande distribution. Dans ces secteurs, massivement informatisés depuis longtemps, les données sont disponibles au sein d'entrepôts de données. [K-07]

On peut résumer les champs d'application les plus importants du Data Mining dans les domaines suivants :

- La gestion de relation client (CRM Customer Relationship Management)

C'est le domaine principal où le Data Mining a prouvé son efficacité. En effet, dans ce cas, le Data Mining permet d'accroître les ventes par une meilleure connaissance de la clientèle. Dans un contexte concurrentiel de plus en plus soutenu, la capacité à conquérir et à retenir les clients repose sur une connaissance fine de leurs besoins et de leur comportement. Les objectifs des analyses en Data Mining sont multiples, tels que la fidélisation, les ventes additionnelles et croisées, l'efficacité de la force de vente, la personnalisation de l'offre, le contact client, l'enquête de satisfaction des clients, …etc.

- L'aide à la décision (Business intelligence)

C'est l'un des meilleurs facteurs d'augmentation de la productivité. Le Data Mining est incorporé dans cette activité pour mieux analyser les données, rechercher des facteurs expliquant les défauts de la production et leur qualité, et anticiper d'éventuelles réactions.

1.4. Les tâches du Data Mining

Pour le Data Mining, plusieurs types de problèmes peuvent être regroupés, suivant leur
formalisation, dans l'une des tâches suivantes :

- Classification
- Estimation
- Prédiction

●Groupement par similitudes

● Segmentation (clustering)

● Description.

Aucun des outils de Data Mining exposés plus bas ne peut résoudre tous les types de tâches. Chaque outil a ses spécificités et usages propres.

●Classification

La classification consiste à examiner les caractéristiques d'un objet et lui attribuer une classe prédéfinie. La classe est un champ particulier à valeurs discrètes. Comme exemple de classification, on peut citer l'attribution ou non un prêt à un client, l'acceptation ou le refus d'un retrait dans un distributeur, ...etc.
A noter que la classification se rapporte à des événements discrets (le prêt a été accepté ou non, le patient a été ou non hospitalisé).
Les techniques des arbres de décision, les plus proches voisins, et les réseaux de neurones sont les plus appropriées à cette tâche de classification.

●Estimation

Contrairement à la classification qui se rapporte à des événements discrets, l'estimation se fait sur des variables continues (durée du prêt, la durée d'hospitalisation). Elle consiste à estimer la valeur d'un champ à partir des caractéristiques d'un objet. Un des intérêts de l'estimation est de pouvoir ordonner les résultats pour ne retenir que les N meilleures valeurs. Cette technique sera souvent utilisée en marketing pour proposer des offres aux meilleurs clients potentiels.
L'estimation peut être utilisée dans un but de classification, en attribuant une classe particulière à un intervalle de valeurs du champ estimé. Par exemple, on peut estimer le revenu d'un ménage selon divers critères (type de véhicule, catégorie socioprofessionnelle, type d'habitation, ...etc.). Il sera ensuite possible de définir des tranches de revenus pour classifier les individus.

La technique des réseaux de neurones est la plus appropriée à l'estimation.

● Groupement par similitudes (selon les affinités)

Cette tâche est assimilée à la technique des règles d'associations. Il s'agit de déterminer quels produits sont achetés ensemble lors d'une même transaction. La technique des règles d'association est la plus adéquate à cette tâche.

● Segmentation (clustering)

Elle permet le fractionnement d'une population hétérogène d'individus en un ensemble de sous-groupes (clusters) plus homogènes. La différence avec la classification réside dans le fait que la segmentation ne se base pas sur des classes prédéfinies. Des outils de détection des clusters sont employés à cet effet.
Les techniques ad hoc à la segmentation sont l'analyse des clusters et les réseaux de neurones.

●Description
Cette tâche permet simplement de décrire ce qui se passe au sein d'une base de données complexe dans l'objectif d'augmenter la compréhension de celle-ci.
Les techniques convenables à la description sont les règles d'association et les arbres de décisions.

1.5. Les outils (techniques de modélisation) de Data Mining

Il existe plusieurs techniques de Data Mining, et il faut insister sur le fait que chacune a une utilité propre par rapport aux tâches présentées ci-dessus. Ces techniques proviennent en fait de diverses disciplines.
Cependant il est impossible de détailler chacune de ces techniques vu l'amplitude du domaine du Data Mining. On présente ici un survol sur certaines approches à savoir:
> Les règles d'association.

> La méthode du plus proche voisin ou RBC

> La détection de clusters.

> Les arbres de décision.

> Les réseaux neuronaux artificiels.

> Les algorithmes génétiques.

1.5.1. Les règles d'association

Les règles d'association sont une des méthodes de Data Mining les plus répandus dans le domaine du marketing et de la distribution.
Les règles d'association générées sont de la forme "Si action1 ou condition alors action2". Elles peuvent se situer dans le temps : "Si action1 ou condition à l'instant t1 alors action2 à l'instant t2" c'est les règles d'association séquentielles.
Leur principale application est « l'analyse du panier de la ménagère », qui consiste, comme l'indique son nom, en la recherche d'associations entre produits sur les tickets de caisse et l'étude de ce que les clients achètent. La méthode recherche quels produits tendent à être achetés ensemble. Elles peuvent être appliquées à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services.
Voici quelques exemples de règles:

- Si un client achète du lait alors il achète du pain (90%)

- Si un client achète une télévision, il achètera un récepteur satellite dans un mois (50%)

- Si maladie X et traitement Y alors guérison (95%)

- Si maladie X et traitement Y alors guérison dans Z années (97%)

- Si présence et travail alors réussite à l'examen (99%)

Ces règles sont intuitivement faciles à interpréter car elles montrent comment des produits ou des services se situent les uns par rapport aux autres. Elles sont particulièrement utiles en marketing et peuvent être facilement utilisées dans le système d'information de l'entreprise. Le but principal de cette technique est donc descriptif. Dans la mesure où les résultats peuvent être situés dans le temps, cette technique peut être considérée comme prédictive. Cependant, il faut noter que cette méthode, si elle peut produire des règles intéressantes, peut aussi produire des règles triviales ou inutiles (provenant de particularités de l'ensemble d'apprentissage). La recherche de règles d'association est une méthode non supervisée car on ne dispose en entrée que de la description des achats.

Puisque la technique des règles d’association sera utilisée pour réaliser notre travail dans ce mémoire, un plus de détail est exposé plus bas dans la deuxième partie de ce chapitre.

1.5.2. La méthode du plus proche voisin ou RBC

La méthode des plus proches voisins PPV (ou nearest neighbor) est une méthode dédiée à la classification qui peut être étendue à des tâches d'estimation. PPV est une méthode de raisonnement à base de cas RBC (ou Case Based Reasoning, CBR). L’être humain base la plupart de ses décisions sur ses expériences passées. Lorsqu’un individu réagit face à une situation quelconque, son processus de décision consiste à faire appel à sa mémoire et dans celle-ci, aux cas vécus précédemment, à identifier certains de ceux-ci comme étant voisins ou similaires à la situation vécue au moment présent et à prendre une décision en fonction de ces voisins. Lorsque l'on présente un nouvel enregistrement, le RBC trouve les voisins les plus proches et positionne ce nouvel élément. Le RBC s'adapte bien aux bases de données relationnelles, qui sont les plus courantes dans le domaine de gestion. On peut l'utiliser pour estimer des éléments manquants, détecter des fraudes, déterminer le meilleur traitement d'un malade,

prédire si un client sera intéressé ou non par telle offre, ou pour classifier les réponses en texte libre, …etc.

De façon similaire, cette technique de Data Mining, permet de classer ou de prédire des données inconnues à partir d'instances connues. En conservant une base de données de cas passés, cette méthode recherche, au sein de celle-ci, des cas similaires ou voisins au cas à analyser soit pour le classer, lorsque le but est la classification, soit pour en faire une prédiction.

1.5.3. La détection de clusters

En classification non-supervisée (ou clustering), les classes possibles et leur nombre ne sont pas connues à l'avance et les exemples disponibles sont non étiquetés. Le but est donc de découvrir des relations intéressantes qui peuvent exister implicitement entre les données et qui permettront de regrouper dans un même groupe (ou cluster) les objets considérés comme similaires, pour constituer les classes.

On distingue trois grandes familles de clustering :

● Clustering hiérarchique: son objectif est de former une hiérarchie de clusters, telle que plus on descend dans la hiérarchie, plus les clusters sont spécifiques à un certain nombre d'objets considérés comme similaires.

●Clustering par partition: son objectif est de former une partition de l'espace des objets, chaque partition représentant alors un cluster; dans cette famille, plusieurs méthodes se distinguent fortement (Clustering basé sur les K-means, basé sur la densité, basé sur l'utilisation de grilles, et basé sur les réseaux de neurones).

● Clustering par sous espace (subspace) : son objectif est de cibler les clusters denses existant dans des sous-espaces de l'espace original.

1.5.4. Les arbres de décision

Les arbres de décisions est un outil puissant utilisé tant pour la classification que pour la prédiction. Ces arbres permettent de distinguer différentes classes et de leur associer une ou plusieurs règles. L'atout principal des arbres de décision par rapport aux réseaux neuronaux artificiels, réside dans le fait qu'elles représentent des règles faciles à comprendre. Il existe un certain nombre d'algorithmes qui partagent cette qualité de compréhension. Les arbres de classification et de régression (CART), et les «CHisquared Automatic Interaction Detection» (CHAID) sont parmi les plus répandus.

1.5.5. Les réseaux neuronaux artificiels

Les réseaux neuronaux sont des outils très évolués. Ces réseaux issus de la biologie, sont en pleine expansion, et représentent de façon simplifiée, les interconnections du cerveau humain. Les réseaux neuronaux artificiels utilisent des méthodes statistiques telles que les probabilités et les distributions. Cependant, les résultats obtenus sont très faibles, et les modèles qui en découlent sont difficiles à comprendre et très sensibles au format des données traitées.

Il existe deux types de réseaux :

- Réseaux à apprentissage supervisé où la réponse est connue à l'avance.
- Réseaux à apprentissage non supervisé où le résultat n'est pas connu à l'avance.

Ces outils sont généralement utilisés pour la classification, l'estimation, la prédiction et la segmentation. Ceux-ci obtiennent de bonnes performances, en particulier, pour la reconnaissance de formes. Donc, ils sont bien adaptés pour des problèmes comprenant des variables continues éventuellement bruitées. Le principal désavantage est qu'un réseau est défini par une architecture et un grand ensemble de paramètres réels (les coefficients synaptiques), le pouvoir explicatif est faible : on parle parfois de « boîte noire ».

2. Les règles d'association

2.1. Domaines d'application

Etant un outil efficace de fouille de données, la recherche des règles d'associations est appliquée dans tous les domaines du Data Mining. Vu ses avantages offerts, cette technique est devenue un sujet attractif et actif appliqué à un large champ d'applications dans divers domaines.

Nous citons ici une liste non exhaustive des applications dont les résultats ont pu être améliorés par l'analyse des règles d'association extraites.

Marketing et Planification commerciale: placement des articles achetés fréquemment ensemble (étagère ou une page de catalogue), organisation des catalogues, choix des articles en promotion, …etc.

Réseaux de télécommunication: filtrage des alarmes non informatives, identification des causes d'anomalies, prédiction des anomalies, …etc.

Recherche médicale: aide au diagnostic et définition de traitement, identification de population à risque vis-à-vis de certaines maladies, prédiction de résultats d'analyses par combinaison de caractéristiques des patients et de résultats d'autres analyses.

Analyse de données spatiales: détection des relations entre caractéristiques des données, prédiction d'évènements, etc.

Internet: Amélioration des modes d'accès aux informations, Modification de la structure des pages et des liens, Personnalisation des pages suivant le profil utilisateur, l'intelligence économique dans les sites E-commerce, suggestion aux clients (comme c'est la cas du site Amazon.com), etc.

Le domaine industriel : prévision des ventes, surveillance des unités de production, diagnostic et analyse des pannes, contrôle de qualité, etc.

Multimédia: analyse d'imagerie, prévision météorologique, aide aux enquêtes, etc.

2.3. Extraction de RA :

L'extraction de règles d'association est un processus itératif et interactif constitué de plusieurs phases. Avant d'évoquer ce processus, on défini certains concepts de bases.

2.3.1. Quelques définitions

Item:

Un item est tout article, attribut, littéral appartenant à un ensemble fini d'éléments distincts X= {x1, x2, ..., xn}.Par exemple, dans les applications de type analyse du panier de la ménagère, les articles en vente dans un magasin sont des items. L'ensemble X peut contenir les items A, B, C et D correspondant aux articles lait, beurre, pain et confiture par exemple.

ItemSet:

Un itemset ou motif est tout sous-ensemble d'items de X. Un itemset constitué de k-items sera appelé un k-itemset. Par exemple, l'ite
mset {A, B, C} est un 3-itemset noté ABC.

Contexte d'extraction de règles :

Un contexte d'extraction de règles d'association est un triplet D= (O, I, R) dans lequel O et I sont respectivement des ensembles finis d'objets et d'items, et R inclut dans O * I est une relation binaire entre les objets et les items. Un couple (o, i) appartient à R dénote le fait que l'objet o inclut dans O est en relation avec l'item i inclut I. Donc, un contexte d'extraction de taille m est une partie d'une base de données dans laquelle s'effectué le traitement. Ensemble des itemsets **fréquents :**

Soit un contexte d'extraction D = (O, I, R). Etant donné un seuil minimal de support minsup, l'ensemble F des itemsets fréquents ou motifs fréquents dans Dans le cas d'un ensemble d'items I de taille m, le nombre d'itemsets potentiels est de 2m-1. Ces items forment le treillis des parties de I, également appelé treillis (voir défi. Treillis ci-dessous) des itemsets de I, dont la hauteur est de m+1.

Borne inférieure (Minorant) et borne supérieure (Majorant) :

Soit E un ensemble ordonné et F une partie de E.

Un minorant de F est un élément x de E tel que tous les éléments de F sont supérieurs ou égaux à x. De même, un majorant de F est un élément x de E tel que tous les éléments de F sont inférieurs ou égaux à x.

Treillis :

Formellement, un ensemble ordonné (E,≤) non vide est un treillis si pour tout couple d'éléments x,y appartient à E l'ensemble {x,y} possède un plus petit majorant et un plus grand minorant. Un ensemble ordonné (E,≤) est un treillis complet si tout sous-ensemble S□ E admet un plus petit majorant et un plus grand minorant.

Donc, un treillis est un graphe des sous ensemble des itemset de taille 0 à i contenant toutes les combinaisons possibles, où i est le nombre des items dans le contexte D. A noter que pour un ensemble d'items I de taille i, on aurait 2i-1 itemsets potentiels. Et donc le treillis contient 2i (si on considère aussi l'itemset vide).

Pour l'exemple, voir le treillis généré dans la phase 3.2.2 (Recherche d'itemsets fréquents).

Support et Confiance :

Chaque règle est évaluée par deux mesures (facteurs) : le support et la confiance.

Le support est une mesure d'importance statistique (statistical significance). Pour qu'une règle x→y vérifie un facteur de support S si et seulement si au moins S % des transactions dans la base de données x et y vérifie.

La confiance est une mesure de la force de la règle (strength of the rule). Pour qu'une règle x→ y vérifie un facteur de confiance C si au moins C % des transactions dans la base de données qui vérifient x vérifie aussi y.

Par exemple , dans une base de données d'enseignement, si on considère la règle suivante :

« 75 % des étudiants qui suivent le cours "Linux/ Unix ", suivent également le cours de
"Programmation C ", et 30 % de tous les étudiants ont en fait suivis les deux cours ».
On peut dire que cette règle est vérifiée avec une certitude supérieure à 75% (confiance de la règle), et que la règle est supportée par au moins 30% des étudiants (support de la règle). Pour être acceptable, il faut que le support de cette règle (30%) soit supérieur à une autre valeur définie à l'avance par l'utilisateur (support minimum), et que la confiance de cette règle (75%) soit supérieure à une autre valeur définie à l'avance (confiance minimale).

Ensemble de règles d'association générées :

Soit un ensemble F d'itemsets fréquents dans un contexte d'extraction D pour un seuil minimal de support minsup. Etant donné un seuil minimal de confiance minconf, l'ensemble AR des règles d'association valides dans D est:

$$AR = \{r: l2 \rightarrow (l1 - l2) \mid l1,l2 \in F \text{ et confiance}(r) \geq \text{minconf} \}$$

$$\text{Avec Confiance}(r)= \sup (l1 \cup l2) / \sup (l2)$$

Donc, pour chaque itemset fréquent l1 dans F, tous les sous-ensembles l2 de l1 sont déterminés et la valeur de la confiance(r) est calculée. Si cette valeur est supérieure ou égale au seuil minimal de confiance alors la règle d'association $l2 \rightarrow (l1 - l2)$ est générée.

2.3.2. Processus d'extraction de règles d'association

Le processus d'extraction de règles d'association est constitué de plusieurs phases allant de la
sélection et la préparation des données jusqu'à l'interprétation des résultats, en passant par
la phase de recherche des connaissances (extraction des ensembles fréquents d'attributs et génération des règles d'association).

Ci-dessous une description de différentes phases de ce processus.

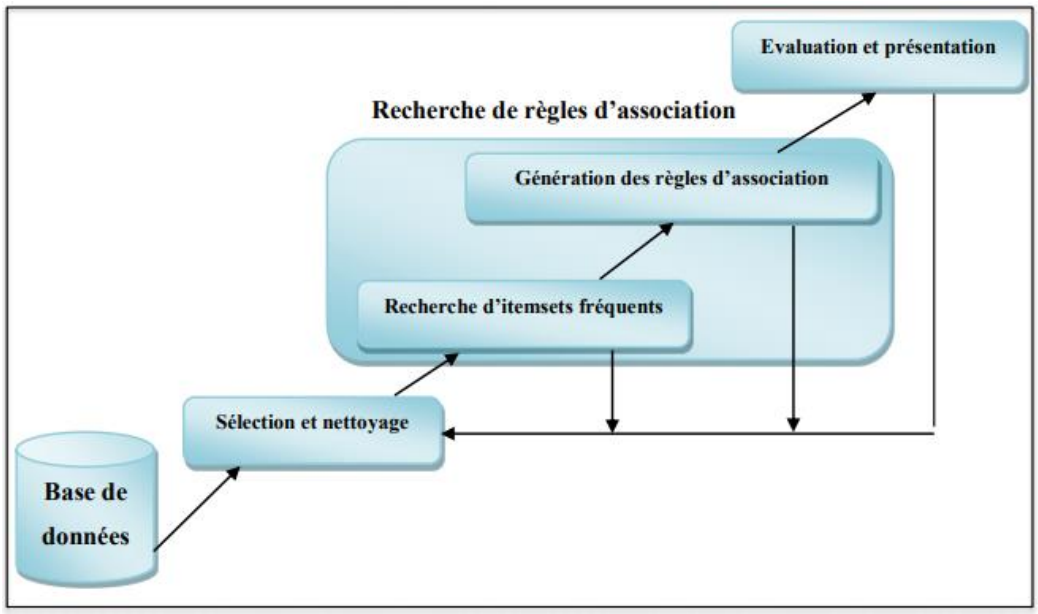


Figure II. 2 : Processus d'extraction des règles d'association

2.3.2.1. Sélection et préparation des données (nettoyage)

Cette phase consiste à sélectionner les données (attributs et objets) de la base de données utiles à l'extraction des règles d'association et transformer ces données en un contexte d'extraction. L'extraction de règles d'association peut être effectuée à partir des bases de données de divers types, comme des données spatiales, temporelles, orientées objets, multi-média, etc. Cette première phase est très importante car à partir de la qualité des données en entrées dépendent la qualité des résultats.

Cette phase est nécessaire pour pouvoir appliquer les algorithmes d'extraction des règles sur des données de natures différentes provenant de sources différentes, de concentrer la recherche sur les données utiles pour l'application et de minimiser le temps d'extraction. A noter que le problème des données incomplètes (valeurs manquantes), et les données erronées ou incertaines et la taille du jeu de données doivent être pris en considération dans cette phase.

Par exemple, le tableau ci-dessous 2.3.2.1 suivant représente un contexte d'extraction D constitué de 6 objets, chacun représenté par son identifiant et de

quatre items. Ce contexte sera utilisé comme exemple dans tout le reste de ce chapitre.

ITEM ID	Items
1	ACD
2	BCE
3	ABCE
4	BE
5	ABCE
6	BCE

Tableau II.1-Contexte d'extraction de règles d'associations D.

Algorithmes de recherche de règles d'association (Algorithme APRIORI)

2.3.2.2. Recherche d'itemsets fréquents

Cette phase consiste à extraire du contexte D tous les itemsets qui sont fréquents. La recherche des itemsets fréquents est un problème non trivial car le nombre d'itemsets fréquents potentiels est exponentiel en fonction du nombre d'items du contexte D.
Dans le cas d'un ensemble d'items I de taille m, le nombre d'itemsets potentiels est de 2m-1.

Ces items forment le treillis des itemsets de I, dont la hauteur est de m+1.

Les balayages du contexte doivent être réalisés lors de cette phase et il est donc nécessaire de développer des méthodes efficaces d'exploration de cet espace de recherche exponentiel.

La phase découverte des items fréquents constitue la phase la plus coûteuse en temps d'exécution et en espace. L'espace de recherche est de taille exponentielle par rapport au nombre d'items. Plusieurs méthodes ont été proposées dans le but de réduire l'espace de recherche de cette phase ainsi que le nombre de balayages du contexte réalisé.

Voici un exemple d'un treillis des itemsets du contexte D donné dans le tableau 2.3.2.1 précédent.

2.3.2.3. Génération des règles d'association

La génération des règles d'association s'effectue à partir des itemsets fréquents générés précédemment.

En général, la génération des règles d'association est réalisée de manière directe, sans accéder au contexte d'extraction, et le coût de cette phase en temps d'exécution est donc faible par rapport au coût de l'extraction des itemsets fréquents.

La génération de règles d'association est assez simple suivant le principe donné dans la définition donnée pour l'ensemble de règles d'association générées (point 2.3.1).Pour chaque itemset fréquent l1 dans F, tous les sous-ensembles l2 de l1 sont déterminés et la valeur de la confiance(r) est calculée. Si cette valeur est supérieure ou égale au seuil minimal de confiance alors la règle d'association l2 → (l1 - l2) est générée.

2.3.2.4. Visualisation et interprétation

C'est la phase finale du processus d'ECD. Cette phase consiste en la visualisation par l'utilisateur des règles d'association extraites du contexte et leur interprétation afin d'en déduire des connaissances utiles pour l'amélioration de l'activité concernée. Ainsi l'expert du domaine peut juger de leurs pertinences et utilités. Mais le nombre important des règles d'association extraites impose le

développement d'outils de classification de règles selon leurs propriétés, de sélection de sous-ensembles de règles selon des critères définis par l'utilisateur, et de visualisation de ces règles sous une forme intelligible. Cette nouvelle problématique est également appelée « Knowledge Mining ».

La forme de présentation de règles peut être textuelle, graphique ou bien une combinaison de ces deux formes intelligibles. Ceci va donner naissance à un nouveau domaine de recherche: la fouille visuelle de données « Visual Data Mining » afin d'améliorer le processus d'extraction de connaissances en proposant des outils de visualisation adaptés à différentes problématiques.

Les connaissances de l'utilisateur concernant le domaine d'application sont nécessaires lors des phases de pré-traitement afin d'assister la sélection et la préparation des données et de post-traitement, pour l'interprétation et l'évaluation des règles extraites. En fonction de l'évaluation des règles extraites, les paramètres utilisés lors des précédentes phases (critères de sélection et préparation des données et seuils minimaux de support et de confiance) peuvent être modifiés avant d'effectuer à nouveau l'extraction des règles d'association, ceci afin d'améliorer la qualité du résultat.

Il ressort de la grande majorité de ces applications qu'au final, beaucoup de règles sont générées par les algorithmes et qu'il est parfois difficile aux experts du domaine de les exploiter dans leur intégralité, car cela engendre un travail cognitif très important. Devant cette tâche, leur premier souhait est souvent de réduire cet ensemble pour ainsi diminuer le temps d'expertise correspondant. En effet, dans le domaine industriel, les experts n'ont pas forcément beaucoup de temps à consacrer à l'analyse des résultats.
Dans le chapitre suivant, nous allons décrire l'implémentation de cette approche sur ce domaine.

1.6. Conclusion

Le Data Mining est le résultat de la combinaison de nombreux facteurs technologiques et économiques. Il peut être vu comme une nécessité imposée par le besoin des entreprises de valoriser les données qu'elles accumulent dans leurs entrepôts de données. Le développement des capacités de stockage et les vitesses de transmission des réseaux ont conduit les utilisateurs à accumuler de plus en plus de données. Le Data Mining répond au besoin d'exploitation pour bénéficier de ces données collectées.

D'une façon générale, le Data Mining est l'art d'extraire des connaissances à partir de données qui peuvent être stockées dans des entrepôts, des bases de données distribuées ou sur Internet. Le Data Mining ne se limite pas au traitement des données structurées sous forme de tables numériques; Il offre des moyens pour aborder les données textuelles exprimées en langage naturel (Text Mining), les images (Image Mining), le son (Sound Mining) ou la vidéo et dans ce cas, on parle alors plus généralement de Multimedia Mining et sur Internet le Web Mining.

Ce chapitre a fait un tour d'horizon sur le concept du Data Mining, sa définition et ses domaines d'applications, ensuite le processus de l'ECD a été décrit, ainsi que les tâches et les techniques du Data Mining.

Parmi les techniques citées ci-dessus, la technique des règles d'association est la mieux adaptée à notre étude de cas qui consiste à fouiller les données de contribution dans le cadre de la méthode des règles d'association par l'utilisation d'une ontologie.

3

Chapitre 3

Conception de l'ontologie AR

Plan du chapitre3

1. ***Introduction***
2. **Spécification (Détermination du domaine et de la portée) de nos ontologies (OntoAR)**
3. **Méthodes et méthodologies d'ingénierie ontologique**
4. **METHONTOLOGY**
5. **Conclusion**

1.Introduction

Dans ce chapitre on va entamer la partie réalisation et implémentation dans laquelle on s'assure que l'application est prête pour être exploitée par les utilisateurs finaux.

Le but de cette phase est de créée un document contenant la spécification de nos ontologies, écrit en langage naturel, en utilisant une représentation intermédiaire. Le domaine (Analyse de domaine incluant les questions de compétences) et la portée de notre ontologie (OntoAR) a été déterminés.

2.Spécification (Détermination du domaine et de la portée) de nos ontologies (OntoAR)

Dans cette première étape de spécification, nous devons répondre à quelques questions :

- Quel est le domaine que va couvrir l'ontologie ?
- Dans quel but utiliserons-nous l'ontologie ?
- A quels types de questions l'ontologie devra-t-elle fournir des réponses ?
- Qui va utiliser l'ontologie ?

Domaine	OntoAR : Règles d'Association
Questions de compétences	1. Quelles sont les taches de RA ? 2. Quels sont les concepts de base qui doivent être compris par un débutant en RA ? 3. Quels sont les types des RA ? 4. Quels sont les étapes de génération des RA ? 5. Quels sont les algorithmes les plus utilisés en RA ? 6. Quels sont les mesures de base en RA ? 7. Y a-t-il un logiciel qui permet à un débutant de commencer par RA ? 8. Quels sont les domaines d'application de RA ?
Objectifs opérationnels	-Décrire les entités de base de la méthode des règles d'association ; -Unifier les termes de cette méthode -Partager la compréhension de la méthode de RA en explicitant ce qui est considéré comme implicite.
Utilisateurs futures	-Chercheurs de la méthode des règles d'association en général ; -Débutant de la méthode des règles d'association.
Degré de formalisme	Formel
Granularité	Fine
Sources de connaissances	-Documents technique relatifs à la méthode des RA (livres, thèses doctorats, articles scientifique ([Agrawal et al. 2004],…) ; Experts de la méthode des règles d'association.

Tableau III.1: Spécification de l'ontologie OntoANN

Dans la littérature, il existe plusieurs méthodologies pour le développement d'une ontologie comme METHONTOLOGY [Fernandez et al., 1997], La méthodologie METHONTOLOGY a été choisie et utilisée pour développer nos ontologies. Le choix de cette méthodologie est dû au fait que ses phases sont distinctes et bien documentées.

3.Méthodes et méthodologies d'ingénierie ontologique :

Le processus de développement d'une ontologie est un processus complexe où plusieurs acteurs interviennent dans les différentes étapes du processus. Il s'agit donc d'une équipe

Pluridisciplinaire. Pour cela, il est nécessaire d'utiliser des méthodes ou méthodologies pour

seconder le processus de construction des ontologies. Cependant, selon (Corcho, et al., 2003), il n'existe pas une méthodologie parmi celles proposées dans la littérature qui est complètement maturée par rapport du génie logiciel ou de l'ingénierie des connaissances.

Les méthodes et les méthodologies recensées permettent la construction d'ontologies à partir de zéro (from scratch) c.-à-d. à partir des données brutes ou par réutilisation d'autres ontologies, la ré-ingénierie, l'intégration ou fusion avec d'autres ontologies, la construction collaborative ainsi que l'évolution des ontologies construites. Jusqu'en 1995, les premières ontologies ont été développées de façon complètement artisanale, sans suivre de méthode prédéfinie. Des premiers projets sont issus des listes de recommandations constituant des ébauches de méthodes, ou cadres méthodologiques. Depuis 1998, on assiste à la naissance de cadres méthodologiques plus élaborés inspirés des méthodes de l'Ingénierie des connaissances (ex : METHONTOLOGY) ou fondés sur la linguistique (ex : TERMINAE),…….etc

4.1.Les activités oriente au développement d'ontologie

Elles sont divisées en trois groupes d'activités :

a- Les activités de développement : Permettent les activités suivantes :

- Etude de l'environnement : Pour savoir ou l'ontologie sera utilisée (plateforme), et les application ou l'ontologie sera intégrée.
- Etude de faisabilité : Pour savoir est ce qu'il est possible de construire une telle ontologie et est-il utile de la construire.

b- Les activités de développent : Ces activités contiennent les phrases suivantes :

- Spécification : Définir pourquoi l'ontologie doit être construire, quels sont ces cas d'utilisation dans le futur et quels sont leurs futures utilisateurs.
- Conceptualisation : Construire un modèle conceptuel sur la connaissance de domaine.
- Formalisation : Transformer le modèle conceptuel en modèle formel.
- Implémentation : coder l'ontologie avec un langage approprié.

c- Les activités post-développement : Permettent les activités suivantes :

- Maintenance : Assure la mise à jour et la correction de l'ontologie.
- Utilisation : Possibilité d'utilisation par les autres ontologies.

4.2. Les activités de support d'ontologie

Ces activités comportent les phases suivantes :

- Acquisition de connaissance : Acquérir la connaissance de domaine.
- Evaluation : Evaluer l'ontologie, son environnement et la documentation et prendre des jugements par rapport à une référence.
- Intégration : Besoin d'autres ontologies pour les utiliser dans l''ontologie en cours de construction.
- Documentation : Produire des documents clairs pour chaque étape et chaque produit.

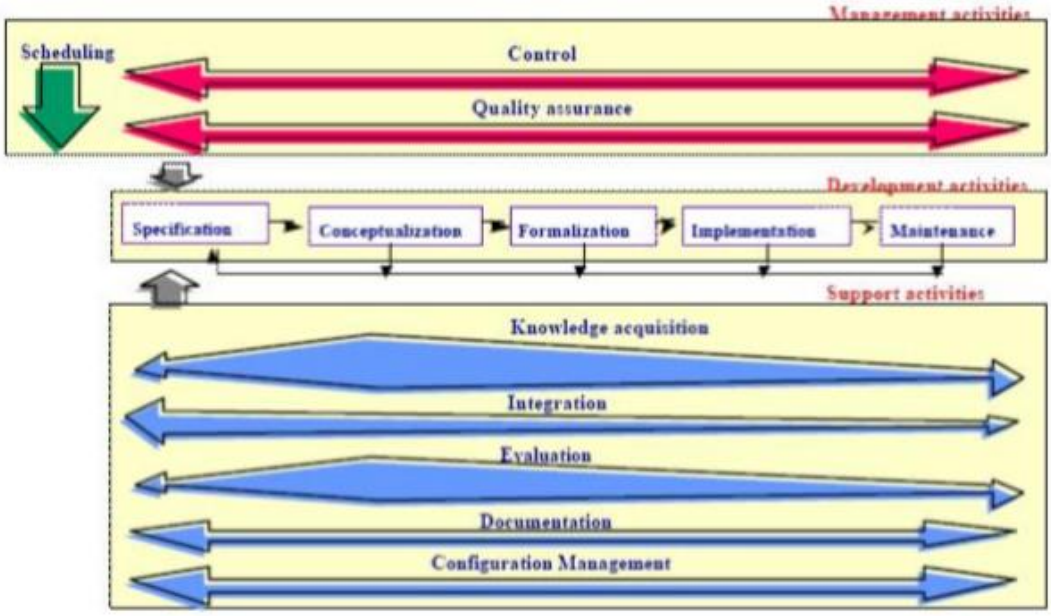


Figure III.1 -processus de développement et le cycle de vie de METHONTOLOG

. Règles d'association guidées par des ontologies et des schémas de règles

En s'inspirant des travaux menés sur les règles d'association généralisées [Srikant et al., 1995] et les schémas de règles, [Marinica, 2010] propose une nouvelle approche de fouille de règles d'association qui intègre explicitement les connaissances de l'utilisateur. Elle propose de modéliser les connaissances du domaine du décideur à l'aide d'ontologies associées aux données et de schémas de règles [Marinica, 2010].

Cette approche est basée sur trois éléments principaux (Figure III.11) : Une base de données dont on extrait des règles d'association. Une ontologie représentant des connaissances liées à la base de données. Un ensemble de schémas de règles, portant sur les concepts de l'ontologie afin de sélectionner les règles intéressantes. La base de données est constituée d'un ensemble de n transactions décrites à travers p attributs. Soit I = {I1, I2, ..., Ip} l'ensemble d'attributs appelés traits (items) et T = {t1, t2,…tn} l'ensemble de n transactions. Chaque transaction ti = {I1, I2, ..., Imi} est un sous-ensemble de l'ensemble d'attributs I. L'algorithme Apriori [Agrawal et al., 1993] permet l'extraction de règles

d'association de la forme X -> Y, où X et Y sont deux ensembles disjoints d'attributs.

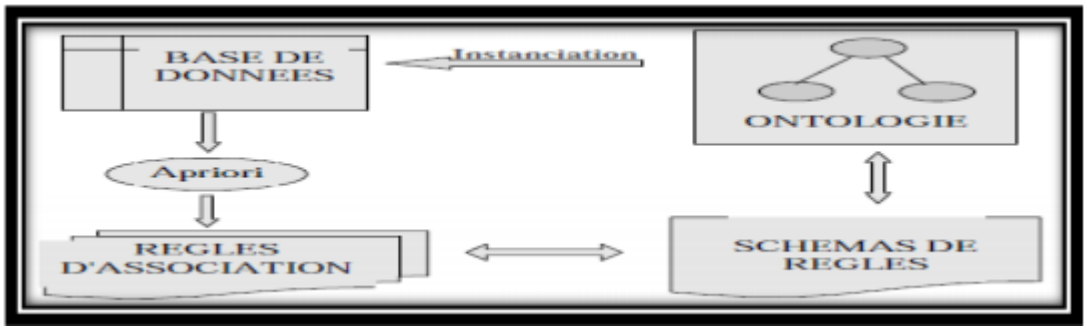


Figure III.2 : Ontologie et schéma de règle pour des RA [Marinica, 2010]

Une ontologie est définie par un ensemble de concepts C = {C1, C2, ..., Co} et un ensemble de relations/propriétés R = {R1, R2, ..., Rr}. Les concepts sont hiérarchisés par une relation de subsomption ⊂. On dit que C1 et C2 sont en relation de subsomption, C2 ⊂ C1, si le concept C1 subsume le concept C2 [Marinica, 2010]. Dans ce scénario, il est fondamental de pouvoir connecter l'ontologie à la base de données. Chaque concept de l'ontologie est instancié dans la base de données par un sous ensemble d'enregistrements. Un moyen simple de réaliser cette connexion consiste à associer directement un concept à un attribut de la base de données. D'autre possibilités sont également envisageables, notamment l'association d'un sousensemble d'attributs à un concept. Enfin, un "schéma de règles" [Marinica, 2010] permet d'exprimer des connaissances sur la forme des règles recherchées. Il constitue une extension sémantique de la notion d'"impression générale" en permettant de combiner dans les schémas de règles non seulement des contraintes sur les attributs, mais également sur les concepts décrits dans l'ontologie.

Terme	Description
Item	Un item est tout article, attribut, littéral appartenant à un ensemble fini d'éléments distincts X
ItemSet	On appelle itemset ou motif tout sous-ensemble d'items de X. Un itemset constitué de k-items sera appelé un k-itemset.
Apriori	le premier algorithme d'extraction des règles d'association dans les bases de données transactionnelles
Frequent-Itemsets	Un itemset est fréquent s'il apparaît fréquemment dans la base de données par rapport à un seuil fixé appelé seuil de support minimum
Support	La mesure de support définie la portée de la règle exp : céréales et sucre → lait le sup est la proportion de clients qui ont acheté les trois articles
Confidence	La mesure de confiance définie la précision de la règle

Tableau III.2 : Description de quelques termes importants dans notre ontologie OntoAR

5.Conclusion :

Au long de ce chapitre, nous avons essayé d'éclaircir la notion d'ontologie en présentant certaines définitions. Nous avons montré aussi leurs avantages, leurs domaines d'application et leurs principaux types

Dans ce chapitre, nous avons présenté les aspects techniques utilisés dans notre travail,

Nous avons présenté dans ce chapitre un état de l'art sur les approches les plus utilisées pour le guidage, la description et le partage de la compréhension de Data Mining en intégrant les ontologies.

Chapitre 4

Implémentation

Plan du chapitre4

IV.*1. Introduction*

Dans cette partie, nous allons présenter notre application d'un point de vue pratique.

IV.*2.2: Implémentation de l'ontologie OntoAR*

Définir l'Ontology IRI qui vous ressemble.

Dans le menu système File/Save, Indiquer le format de sauvegarde de l'ontologie, je suggère RDF/XML et l'endroit sur votre disque ou sera conservée l'ontologie.

En principe, votre écran devrait ressembler à celui-ci :

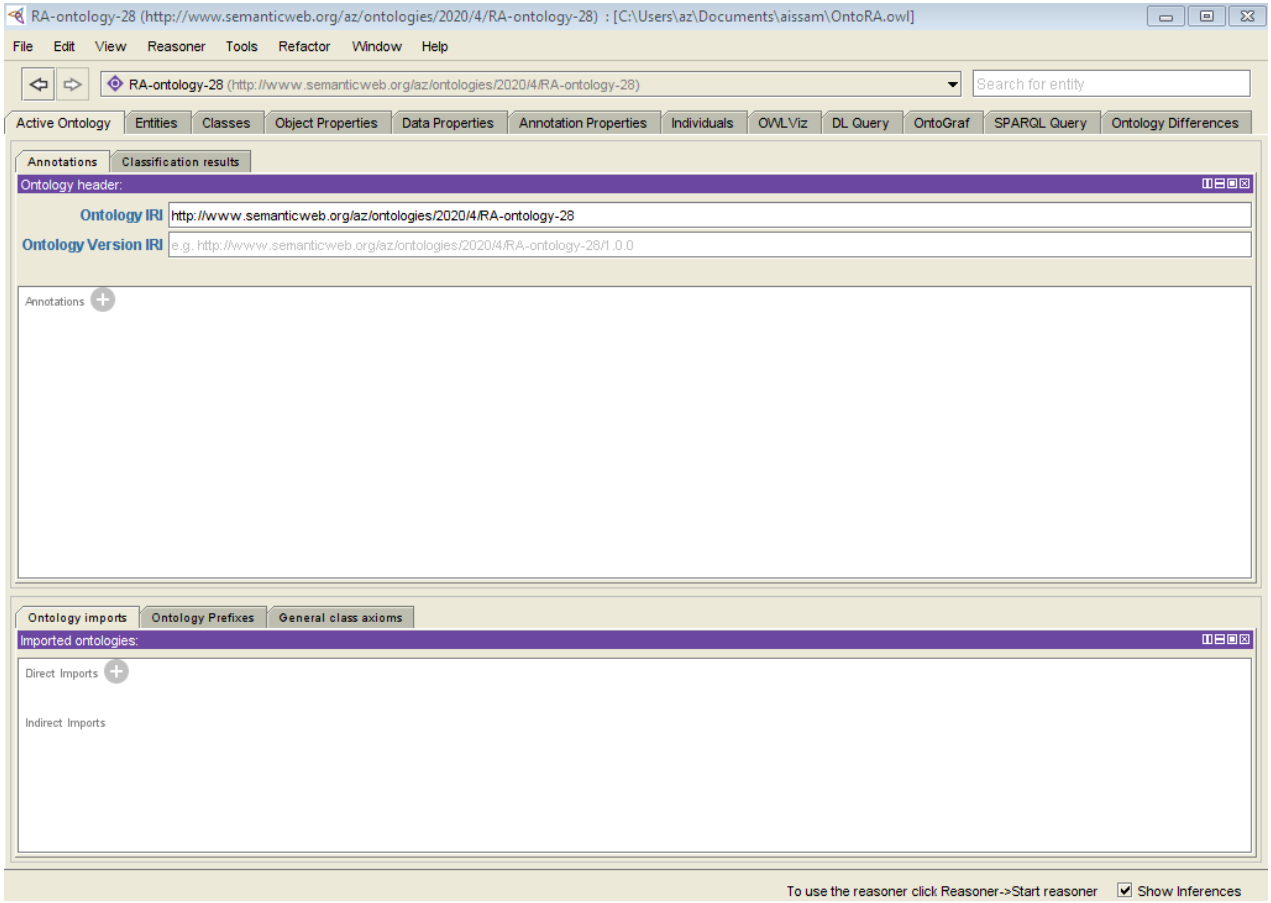


Figure IV.1 : Ontology IR

Etape 1.Définition des concepts :

La structure de notre ontologie « OntoAR.owl » se compose de huit parties principales (Algorithms, Application-Domains, Tasks, Soft, Types, Process, Measures et Basic-Concepts) présentées dans la figure III.2.2.

« owl :Thing » est une classe prédéfinie. Toute classe OWL est une sous-classe d'owl :Thing. Les figures ci-dessous sont une représentation graphique (des captures d'écran) de la hiérarchie des classes de notre ontologie, via l'onglet « OntoGraph ».

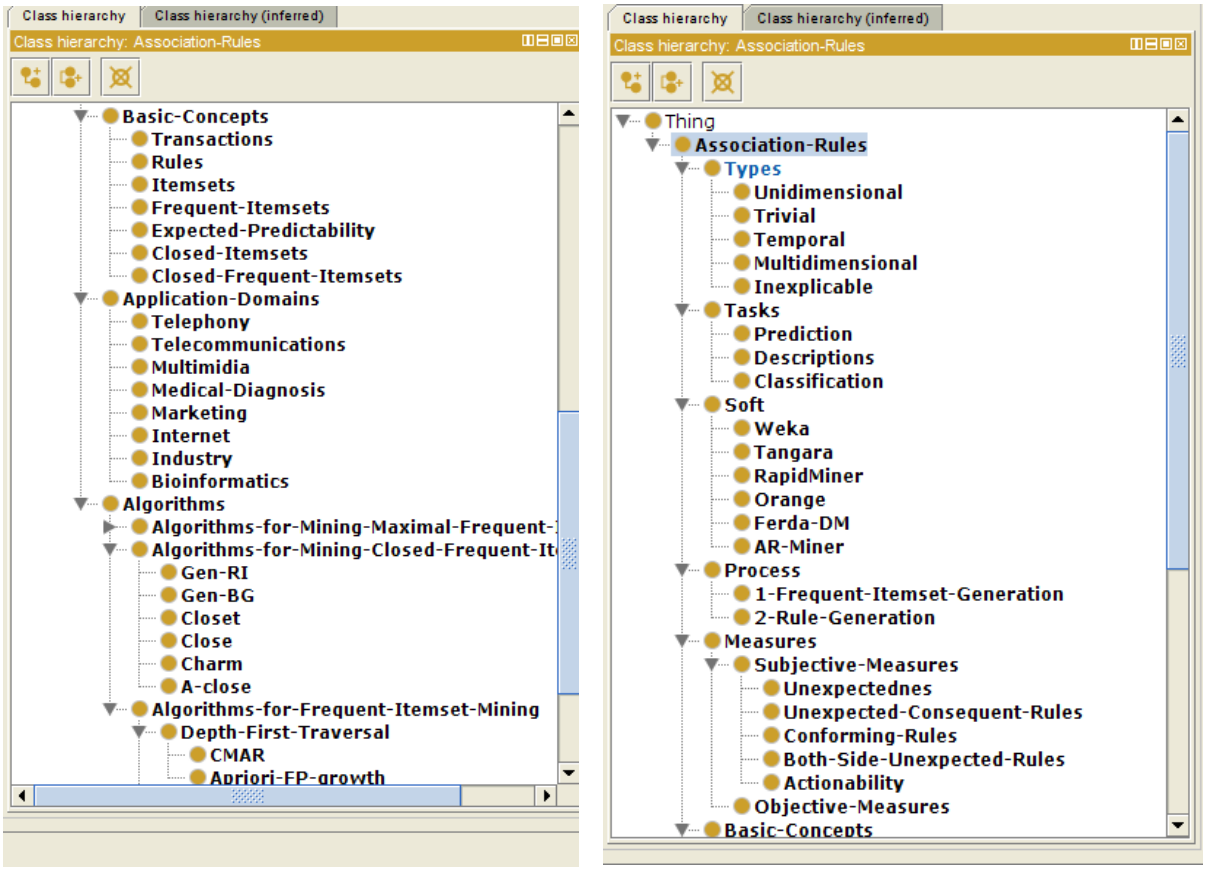


Figure IV.2 : La structure générale de notre ontologie OntoAR

Pour l'ontologisation de notre ontologie OntoAR, nous avons commencé par les classes les plus générales (qui sont les classes principales de notre ontologie

OntoAR), à savoir : Tasks, Basic-Concepts, Measures, Types, Process, Algorithms, Soft et Application-Domains (*Figure III.2.3).*Ces huit classes sont les reponses à nos quesions de compétences voir le(Tableau III.1).

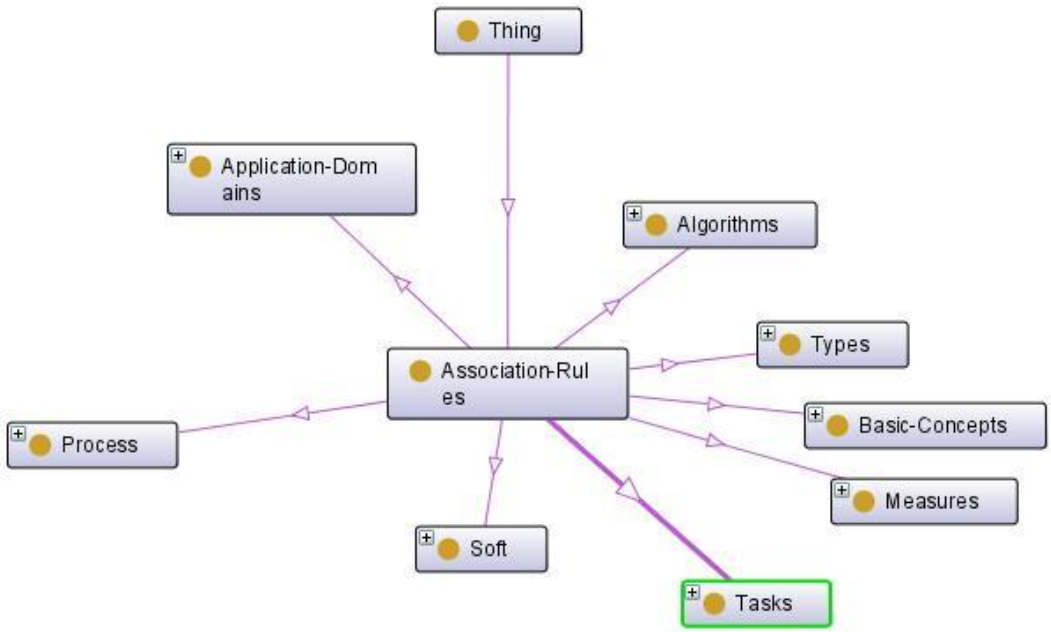


Figure IV.2.1 : Les classes principales de notre ontologie OntoAR

Ensuite, nous avons affiné chacune de ces classes pour repondre aux questions de compétences. Par exemple, la classe **Process** a été affinée en les concepts : **Frequent-Itemsets-Generation, Rule-Generation** (Figure III.2.4) pou répondre à la question de compétence n°4 : Quelles sont les étapes de génération des RA ?

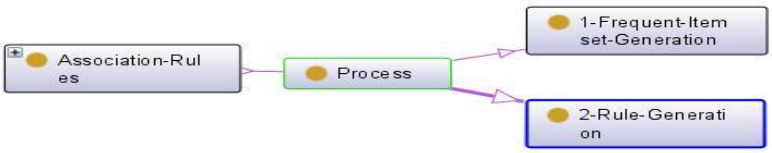


Figure IV.2.2 : Les étapes de la méthode des Règles d'Association

Pour répondre à la question de compétence n°6 : Quelles sont les measures de base en RA ?

La classe **Measures** a été affinée en les concepts : **Objective-Measures, Subjective-Measures** et la classe **Subjective-Measures** a été spécialisé en les sous-concepts suivants :

Conforming-Rules, Unexpected-Consequent-Rules,Both-Side-Unexpected-Rules,Actionaility,Unexpectednes pour indiquer les mesures subjectives de la méthode des Règles d'Association (Figure III.2.5)

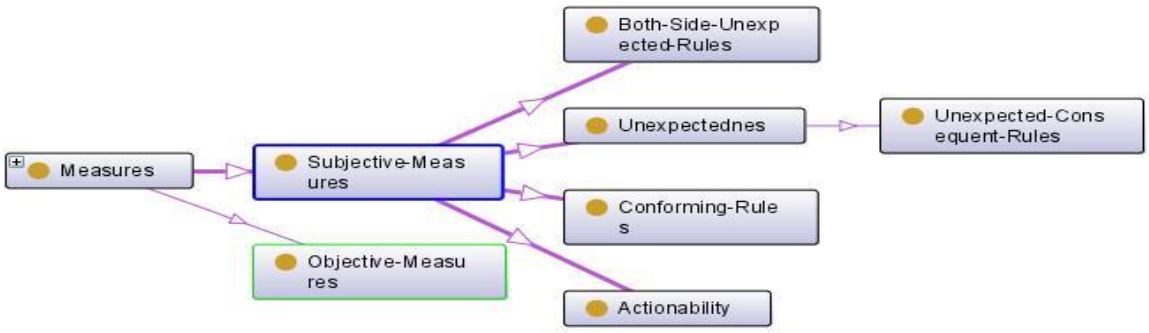


Figure IV.2.3: Les mesures subjectives de la méthode des Règles d'association

Parmi les logiciels qu'un débutant de la méthode des règles d'association peut l'utiliser pour appliquer cette technique nous trouvons : AR-Miner,Ferda-DM,Weka,Tanagra(*FigureIII.2.6*).

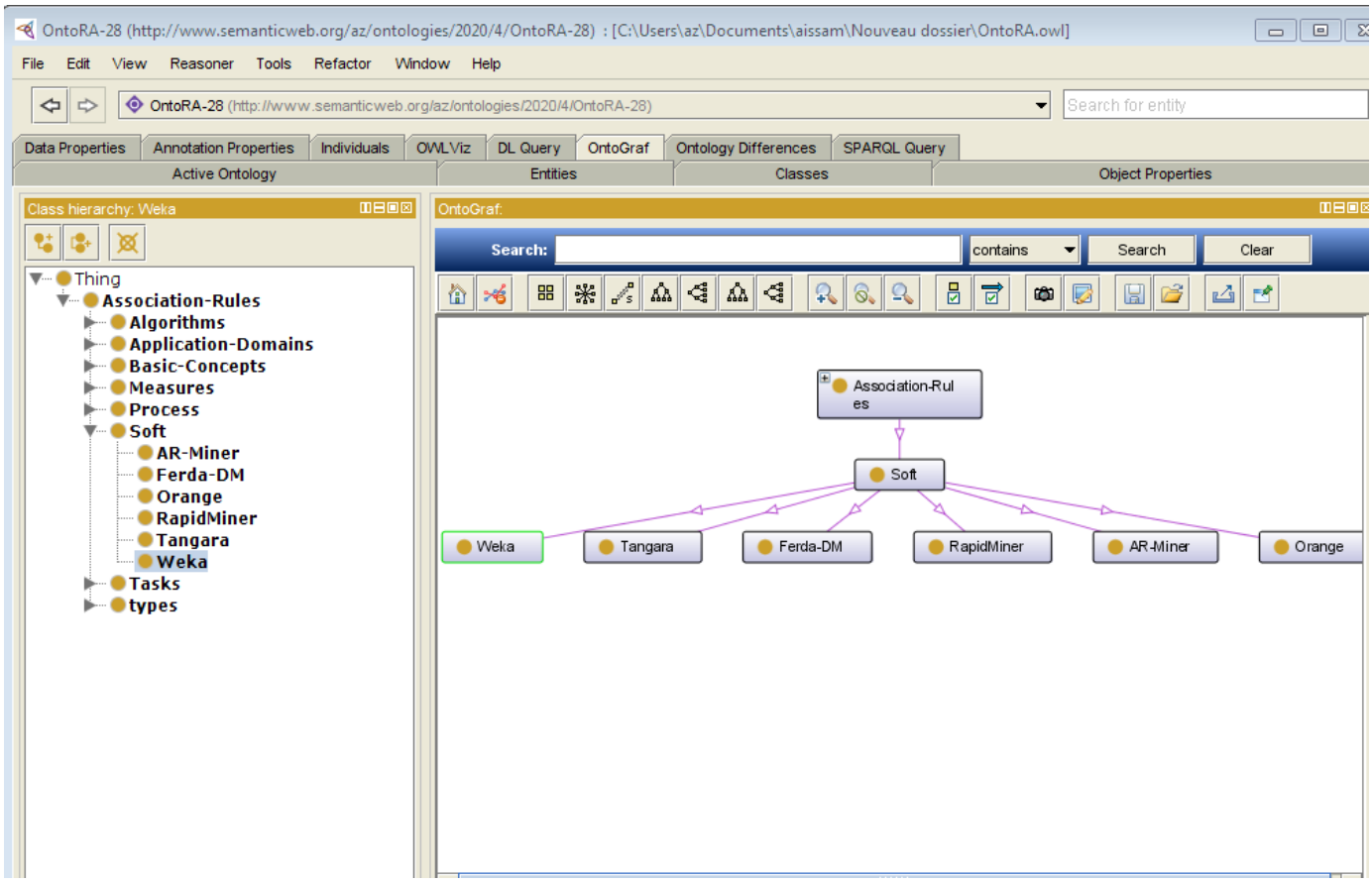


Figure IV.2.4: Les logiciels de la méthode des Règles d'Associatio

Nous avons classé les algorithmes de la méthode des règles d'association dans trois classes :

1. Algorithmes pour l'extraction des itemsets fréquent (Apriori,DIC,FP-Growth,…).
2. Algorithmes pour l'extraction des itemsets fréquent maximaux (Max-Miner, MAFIA).
3. Algorithmes pour l'extraction des itemsets fréquent fermés (Close,Closet,…).

La figure III.2.7 montre les algorithmes de base pour la fouille des itemsets fréquent dans la méthode des règles d'association tel que : Apriori,Apriori-DIC,CMAR,CBA-Algorithm…etc.

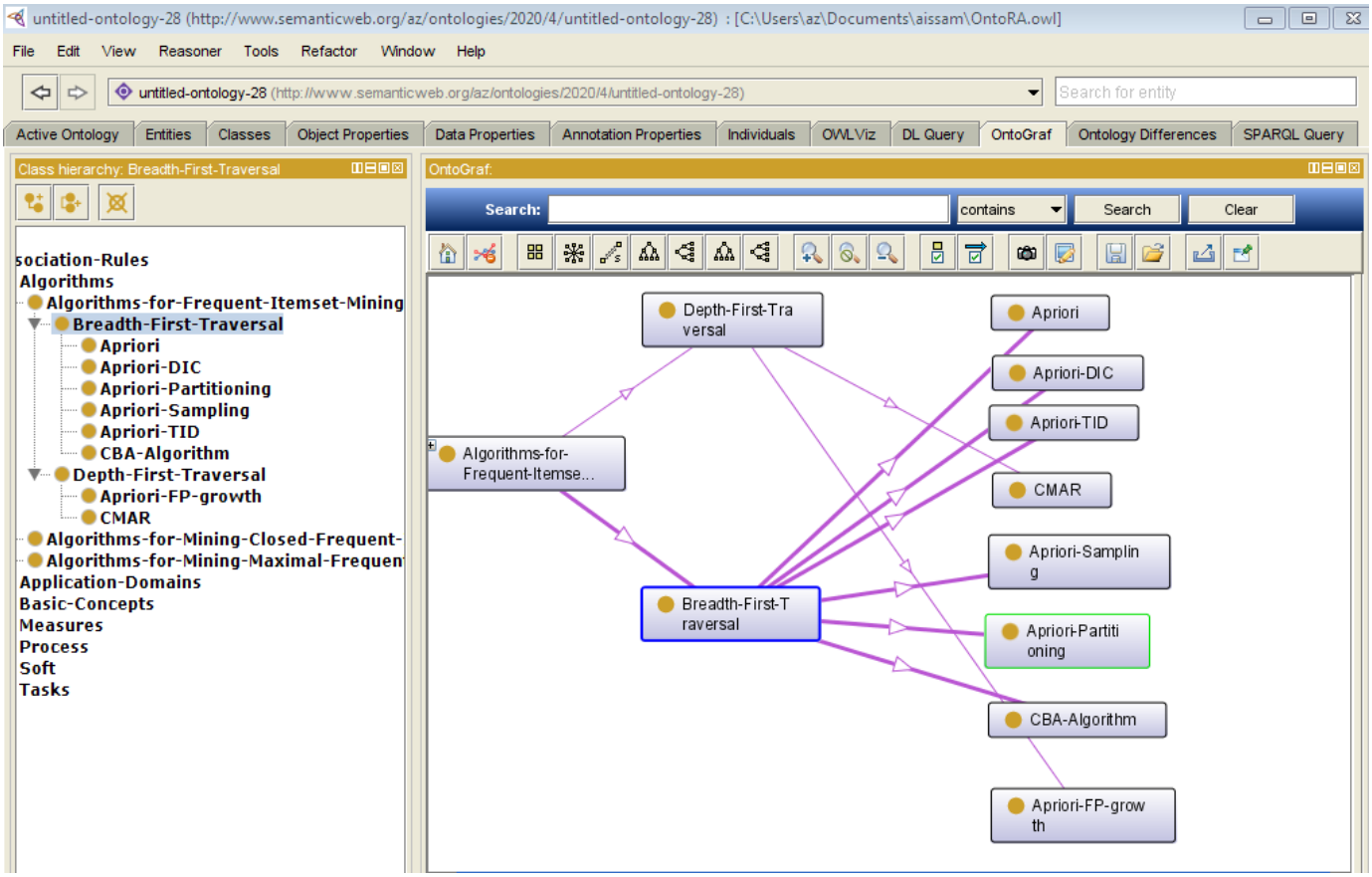


Figure IV.2.5: Les algorithmes de fouille des itemsets fréquent

Etape2.Définition des relations sémantiques

Les relations sémantiques sont des liaisons entre des concepts ou des instances, par exemple :

Le concept **Antecedent *is-part-of* (*est une partie*)** deconcepts***Rule***(FigureIII.2.8*)*

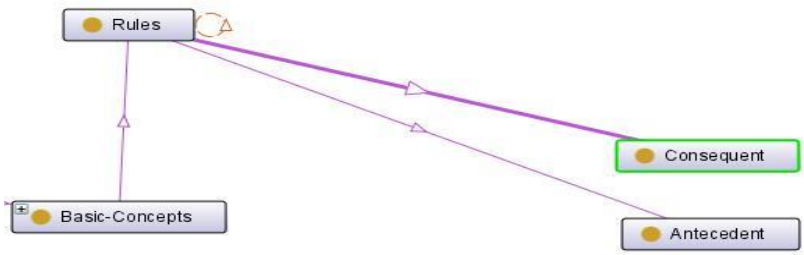


Figure IV.2.6: Exemple d'une relation sémantique dans OntoAR.owl

Le fragment de code OWL suivant montre la création de la relation sémantique **"use"** entre les concepts : **Apriori, Confidence et Support** où l'algorithme **apriori** utilise les deux mesures objectives : **le support et la confiance.**

<owl:ObjectProperty
rdf:about=http://www.semanticweb.org/ontologies/2020/4/OntoAR.owl#use
>
<rdfs:domain
rdf:resource=http://www.semanticweb.org/ontologies/2020/4/OntoAR.owl#Apriori"/>
<rdfs:range
rdf:resource=http://www.semanticweb.org/ontologies/2020/4/OntoAR.owl#Confidnce"/>
<rdfs:range
rdf:resource=http://www.semanticweb.org/ontologies/2020/4/OntoAR.owl#Support"/>
</owl:ObjectProperty>

Etape 3.Définition des propriétés des classes

La troisième étape consiste à associer des propriétés à chaque concept. Ces propriétés sont des attributs qui caractérisent chaque classe. Par exemple le concept Support a une propriété Minimum-Support de type réel (Figure III.2.9

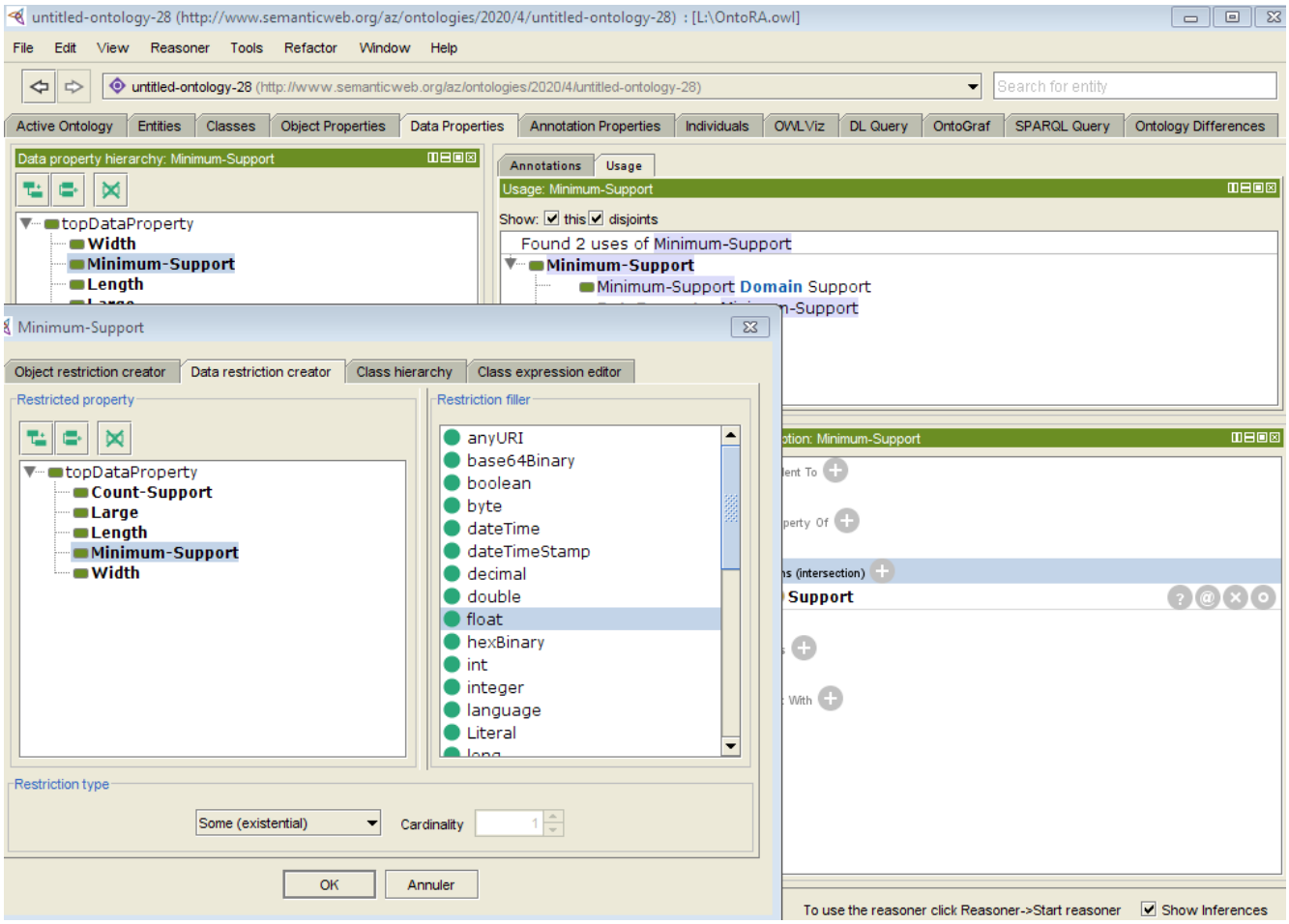


Figure IV.2.7: Les propriétés des classes d'OntoAR.owl

Etape 4.Création des instances

Cette étape consiste sur l'ajout des instances pour chaque classe si c'est nécessaire. Par

Exemple, l'instance **Filter** pour la classe Item (Figure III.2.10)

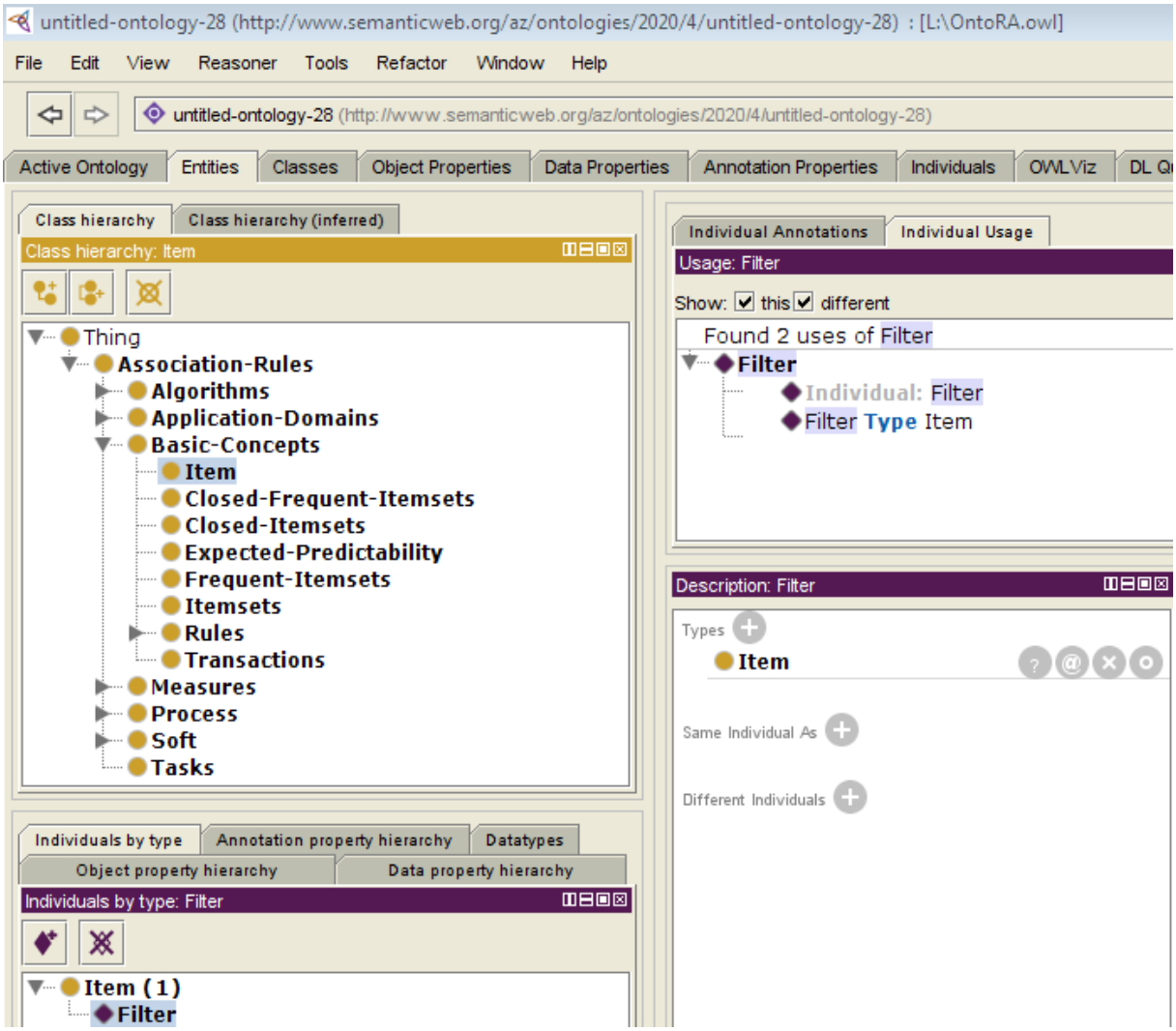


Figure IV.2.8: L'instance ***"Filter"*** *de la classe Item de l'ontologie OntoAR.owl*

Etape 5.Création des axiomes

Les «axiomes» sont souvent utilisés pour désigner des énoncés cohérent qui peuvent être réalisés dans RDFS / OWL, leur inclusion dans une ontologie peut avoir plusieurs objectifs impliqués dans la définition des significations des composants de l'ontologie, des contraintes sur les valeurs d'attributs ,des arguments relationnels et l'inférence de nouvelles informations

Par exemple :

-La propriété "Minimum-Confidence" est une propriété de la classe Confidence, de type :

Float avec un maximum de 1 et un minimum de 0 (restriction) (Figure III.2.11).

<DataPropertyDomain>

<DataPropertyDomain IRI="#Minimum-Confidence"/>

<DataMinCardinality cardinality="0">

<DataPropertyDomain IRI="#Minimum-Confidence"/>

<Datatype abbreviatedIRI="xsd :float"/>

</DataMinCardinality>

<DataMaxCardinality cardinality="1>

<DataPropertyDomain IRI="#Minimum-Confidence"/>

<Datatype abbreviatedIRI="xsd :float"/>

</DataMaxCardinality>

<DataPropertyDomain>

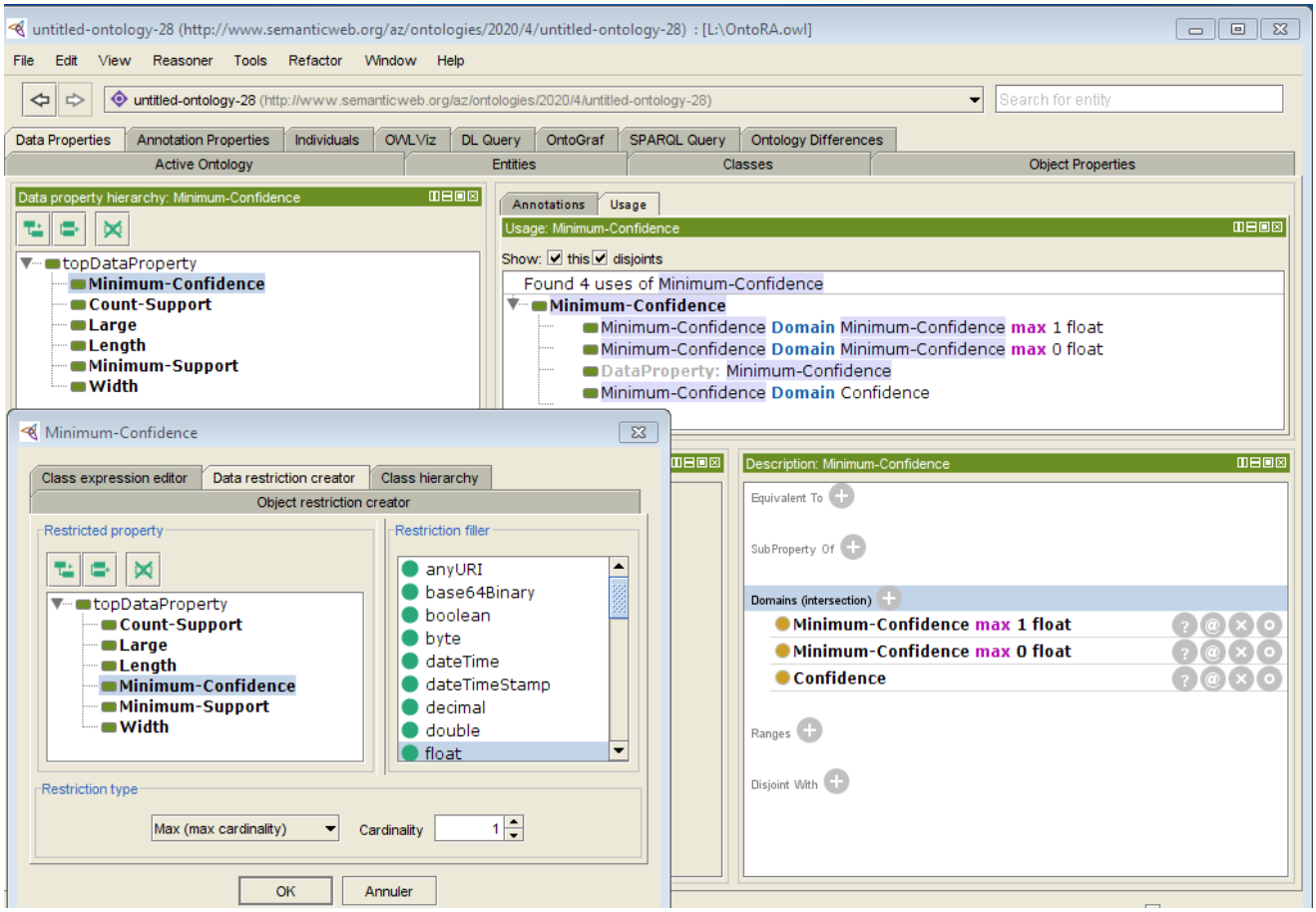


Figure IV.2.9: Exemple d'axiomes en OntoAR.owl

Finalement, voici un diagramme (Figure III.2.12) qui résume les différentes parties (concepts, instances et relations) de notre ontologie OntoAR :

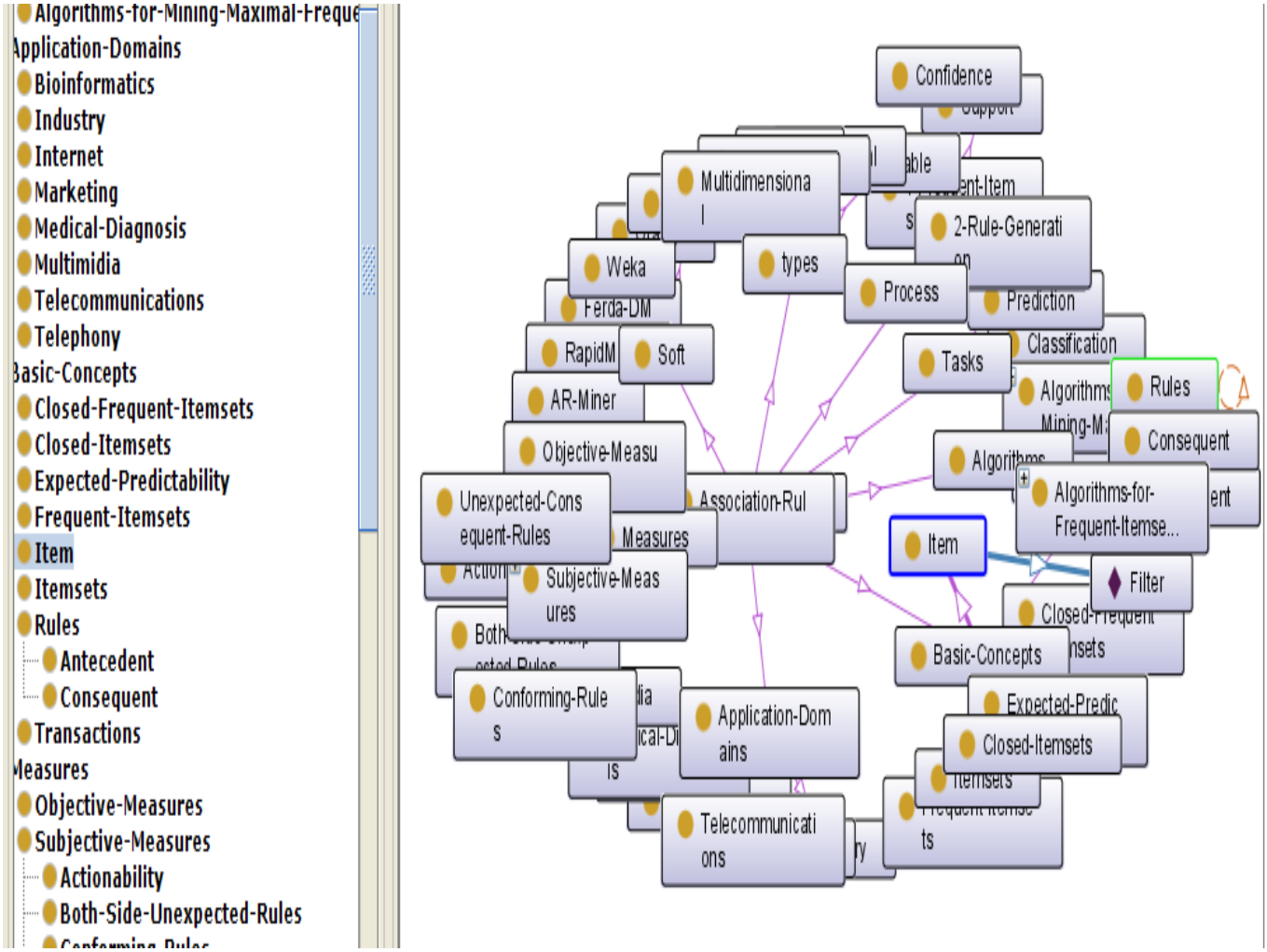


Figure IV.2.10: Diagramme d'OntoAR.

IV.3.Evaluation et validation d'OntoAR :

Selon la méthodologie METHONTOLOGY [Fernandez et al., 1997], adoptée pour le développement de nos ontologies : OntoAR, la phase d'évaluation de l'ontologie consiste à faire un jugement technique, y compris la validation et la vérification. La vérification se réfère au processus technique qui garantit que l'ontologie est correcte formellement, syntaxiquement et sémantiquement. Donc, la phase d'évaluation de nos ontologies a été divisée en trois étapes :

1. Evaluation formelle et syntaxique ;
2. Evaluation des métriques de chaque ontologie ;
3. Evaluation sémantique

Figure IV.3.1: Validation Syntaxique de l'ontologie OntoAR par le W3C RDF validateur

Description (DL) a été utilisée en profitant de plug-in DL Query et du raisonneur FaCT++ (Figure IV.4) dans l'outil Protégé.

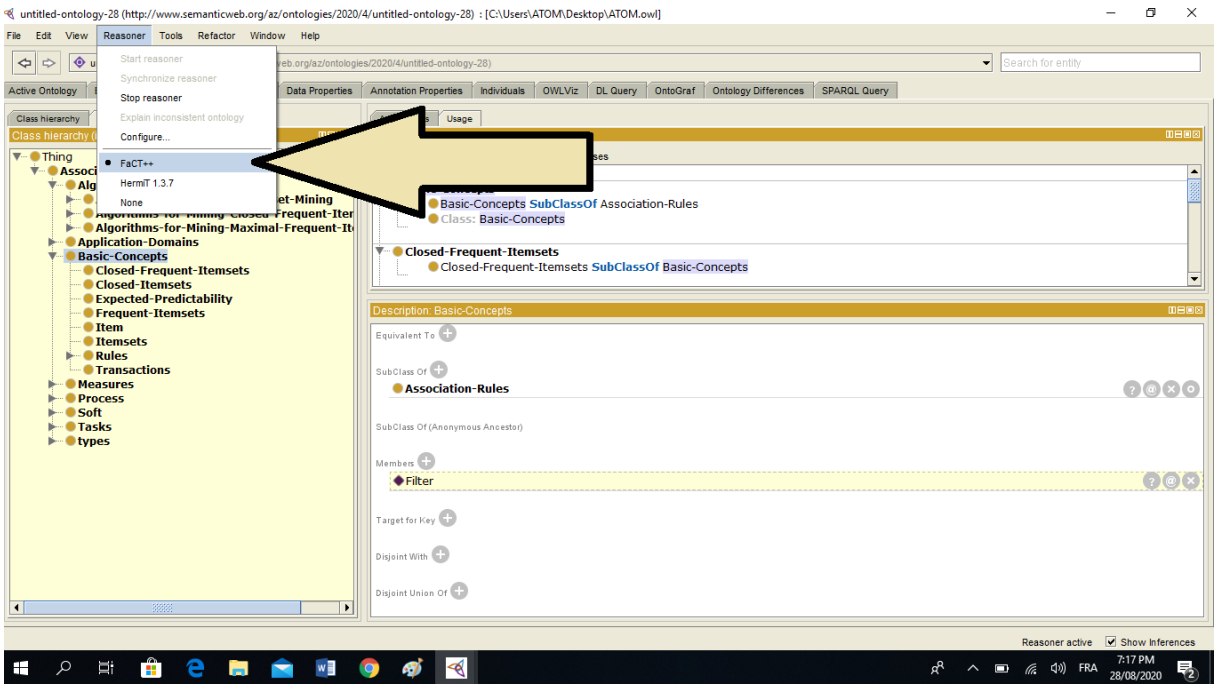


Figure *IV.3.2*: Lancement de raisonneur FaCT++

Par exemple, la Figure IV.5 montre les réponses aux questions de compétence n° 1 : Quelles sont les tâches de RA ? Et n° 3 : Quels sont les types des RA ? ,en posant les requêtes à l'ontologie OntoAR en utilisant le plug-in DL Query

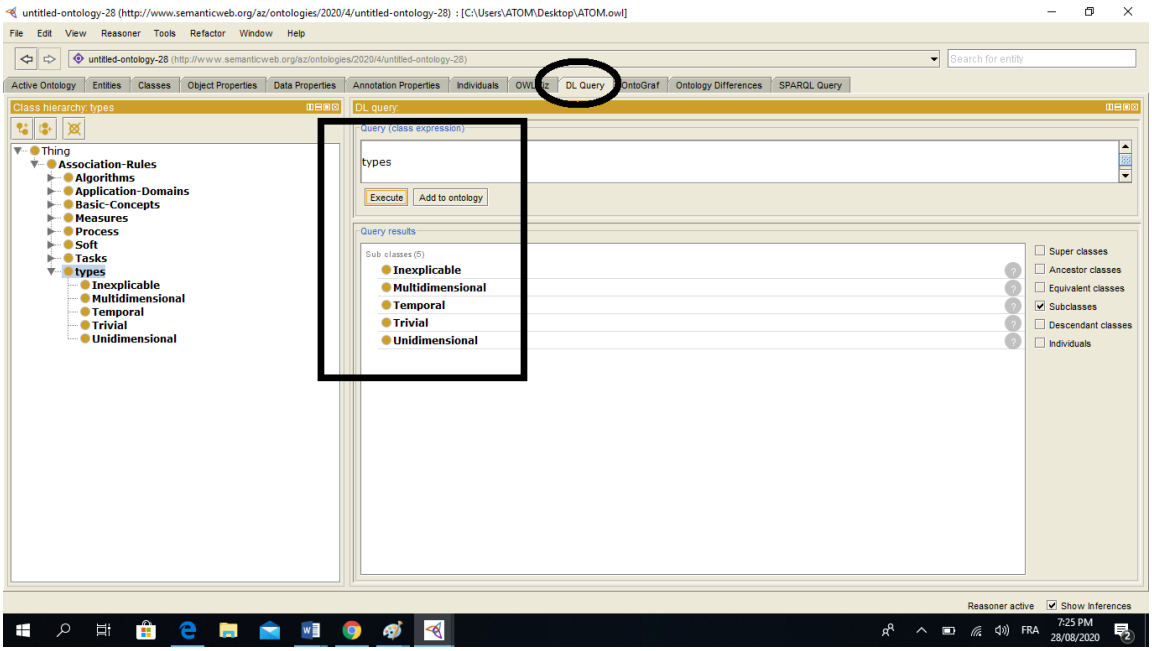


Figure IV.3.3: Réponses des requêtes DL aux questions de compétence 1 et 3 dans le domaine des RA

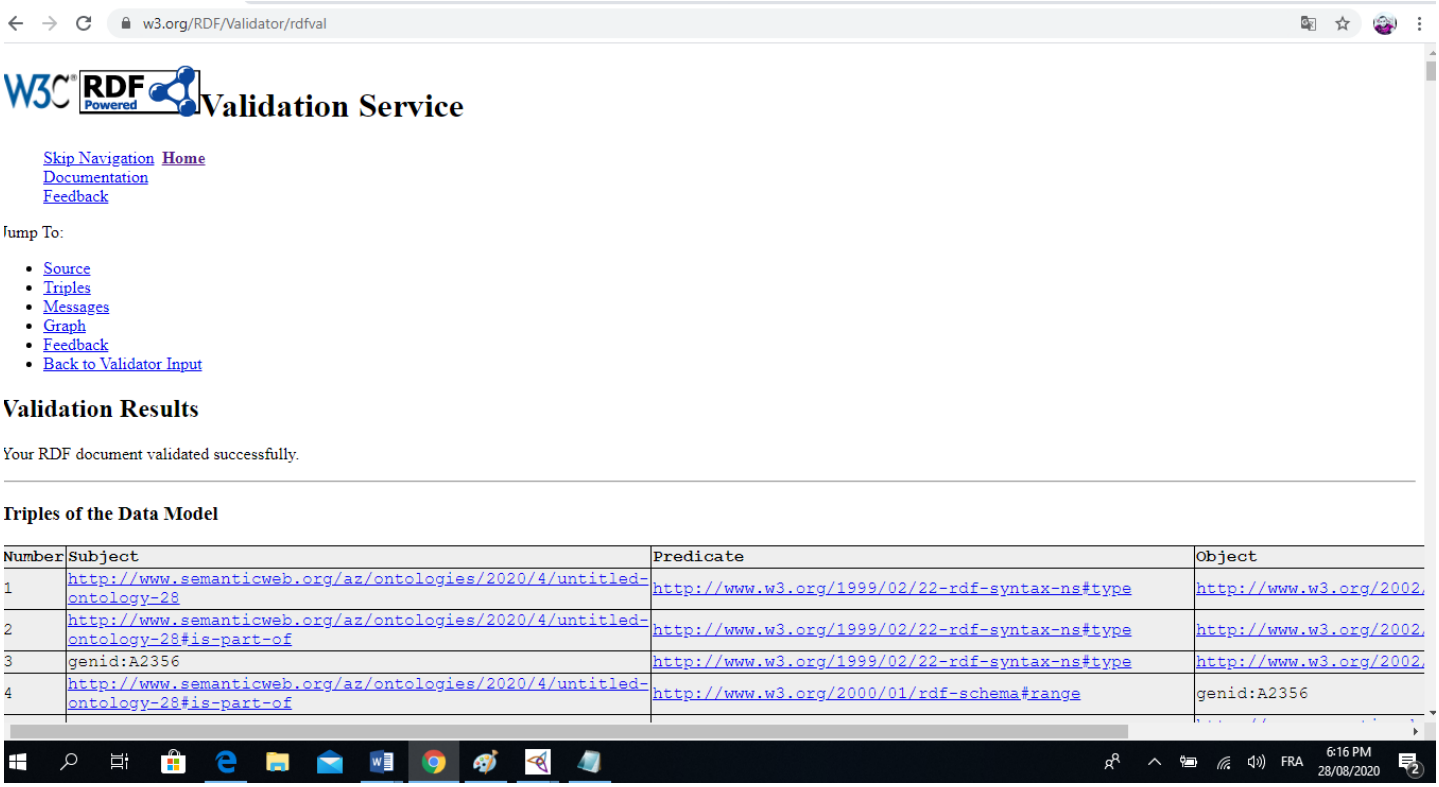


Figure *IV.3.4*: Validation Service

IV.*4 : Conclusion*

Tout au long de ce chapitre, nous avons détaillé la partie implémentation de notre ontologie de domaine OntoRA pour décrire respectivement la technique des Règles d'Association.Nous avons donné une description détaillée de notre ontologie à travers des fenêtres de captures qui représentent les interfaces, qui sont conçues de manière à être conviviales et simples d'utilisation.

A travers tous ce que nous avons présentes dans ce chapitre nous avons réussi à construire une ontologie de domaine de Data Mining qu'elle est prête à une future évaluation.

Conclusion Générale

Conclusion générale

L'intégration sémantique de données consiste à offrir une représentation conceptuelle des données afin d'éliminer les conflits entre ces données. Actuellement, les ontologies sont utilisées dans divers domaines pour représenter explicitement les connaissances d'un domaine particulier. Dans le cadre de cette mémoire nous nous sommes intéressés à la construction des ontologies do domaine pour décrire des techniques spécifiques en Data Mining.

Nous avons présenté à travers les chapitres de ce mémoire la notion d'ontologie qui a été introduite pour guider le processus, le langage et les outils utilisés exactement le langage Protégé , puis les notions de bases du Data Mining, ses taches et ses principales techniques ,ainsi nous avons défini la technique des règles d'associations, avec une description détaillée de l'algorithme de base *Apriori.* Par la suite, nous avons abordé dans le chapitre III les différentes étapes de construction d'ontologie avec des explications de la création des classes à la création des object properties et Data Property, puis la création des Individuals. Enfin la conception d'ontologie.

Afin de bien mener ce processus, nous avons proposé une contribution qui consistait à introduire les concepts de l'ontologie et des schémas de règles dés le début du processus du Data Mining. Cette contribution permet une amélioration considérable du processus dans ses deux phases.

La construction de notre ontologie OntoAR est une tache ardue ou le processus que nous avons suivi pour la construction de notre ontologie n'était pas linéaire. Plusieurs versions ont été construites avant la convergence à une première version plus ou moins complète.

Perspectives

Les résultats de ces travaux présentés dans cette mémoire offrent plusieurs perspectives de recherche ultérieurs au niveau théorique et au niveau pratique. Dans cette section, nous présentons succinctement quelques-unes de ces perspectives qui nous paraissent être les plus intéressantes.

- Il est peu probable que notre ontologie OntoAR soit suffisante pour représenter toutes les connaissances de la technique des RA qui évoluent sans cesse. Il s'agit donc d'étendre cette ontologie qui doit s'adapter à l'évolution des besoins des utilisateurs, ou encore de nouveaux algorithmes qui peuvent intervenir dans la technique de Data Mining etc…
- Compléter le processus suivi pour bâtir notre ontologie par l'ajout d'une phase d'acquisition de connaissances semi-automatique.
- Créer des outils de navigation et de visualisation de notre ontologie.
- Diffuser et héberger notre ontologie sur un site internet et effectuer des appels depuis les URIs correspondant.
- Intégrer notre ontologie dans un outil logiciel de Data Mining(comme Tanagra),afin d'améliorer la participation humaine à cet outil et nous allons connecter notre ontologie à d'autres proposition d'ontologie.
- Développer d'autres ontologies pour décrire et unifier les termes d'autres techniques de Data Mining (Les algorithmes Génétiques,le Text Mining…).

Bibliographie

[FG & al-10] **Javier Farreres**, Karina Gibert, Horacio Rodriguez,Charnyote Pluempitiwiriyawej.

[WP] **Site** Wikipédia http://wikipedia.org

[JF & al-07] **Lisa Di Jolio**, Lylia Abrouk, Céline Fiot,Danièle Hérin,Maguelonne Teisseire. «Enrichissement d'ontologie basé sur les motifs séquentiels».LIRMM 2007, Campus Saint-Priest, Montpelier.

[B-04] **Brisson Laurent.** «Mesures d'intérêt subjectif et représentation des connaissances »Laboratoire informatique,signaux et systèmes ,de,Sophia Antipolis. Project EWECO, Octobre 2004.

[UG-96**] Mike Usehold**, Michael Gruninger.«Ontlogies principes, methods and application» AIAI-TR,Knowledge Engineering Review,June 1996.

[G-93] **Thomas R.Gruber**. «Toward Principles for the Design of Ontologies Used for Knowledge Sharing in Formal Ontology in Conceptuel Analysis and Knowledge Representation» Kluwer Academic Publishers, 1993

[PS] **Site** Protégé Stanford University http://protege.stanford.edu/

[C-06**] Claire Noirault** ''Business Intelligence avec Oracle 10g (ETL,Data Warehouse,Data Mining Rapports)" Editions ENI-France,Novembre 2006.

[K-07**] Khiat Salim**. « Data Mining Industriel,application à la maintenance Aval/Sonatrach».

[CRISP] **Site** du Consortium CRISP-DM spécialise dans le Data Mining http://crisp-dm.org

[Fernandez et al., 1997] **M. Fernandez**, A. Gomez-Perez N. Juristo.

METHONTOLOGY, from Ontological art towards Ontological engineering. In Proceedings of the Springs Symposium Series on Ontological Engeneering (AAAI'97);AAAI Press ,1997.

Annexe

Dans cette annexe nous donnons un exemple d'une partie d'un document OWL qui concerne notre ontologie OntoAR et la déclaration de ses composants : les concepts, les propriétés, les relations sémantiques, les instances et les axiomes.

```
<?xml version="1.0"?>
<!DOCTYPE Ontology [
   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
   <!ENTITY xml "http://www.w3.org/XML/1998/namespace" >
   <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
   <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#"/ >
<Ontology xmlns=http://www.w3.org/2002/07/owl#/>
    xml:base=http://www.semanticweb.org/az/ontologies/2020/4/OntoRA-28/
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xml="http://www.w3.org/XML/1998/namespace"
    ontologyIRI=http://www.semanticweb.org/az/ontologies/2020/4/OntoRA-28/>
   <Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
   <Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#"/>
   <Prefix name="xsd" IRI="http://www.w3.org/2001/XMLSchema#"/>
   <Prefix name="owl" IRI="http://www.w3.org/2002/07/owl#"/>
   <Declaration>
      <Class
IRI="http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#1-Frequent-Itemset-Generation"/
```

IRI="http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#2-Rule-Generation"/>

</Declaration>

<Declaration>

<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#A-close/>

</Declaration>

<Declaration>

<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#AR-Miner/>

</Declaration>

<Declaration>

<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Actionability/>

</Declaration>

<Declaration>

<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Algorithms/>

</Declaration>

<Declaration>

<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Algorithms-for-Frequent-Itemset-Mining/>

</Declaration>

<Declaration>

<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Algorithms-for-Mining-Closed-Frequent-Itemsets/>

<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Algorithms-for-Mining-Maximal-Frequent-Itemsets_/>

</Declaration>

<Declaration>
<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Antecedent/>
</Declaration>
<DataPropertyDomain>
<DataProperty IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Minimum-Confidence/>
<DataMaxCardinality cardinality="1">
<DataProperty IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Minimum-Confidence/>
<Datatype abbreviatedIRI="xsd:float"/>
</DataMaxCardinality>
</DataPropertyDomain>
<DataPropertyDomain>
<DataProperty IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Minimum-Support/>
<Class IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Support/>
</DataPropertyDomain>
<DataPropertyDomain>
<DataProperty IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Minimum-Support/>
<DataSomeValuesFrom>

<DataProperty IRI=http://www.semanticweb.org/az/ontologies/2020/4/untitled-ontology-28#Minimum-Support/>

<Datatype abbreviatedIRI="xsd:float"/>

</DataSomeValuesFrom>

</DataPropertyDomain>

</Ontology>