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

Université Saad Dahlab Blida

N° D'ordre :…….

Faculté des sciences

Département d'informatique

Mémoire Présenté par :

DAAOU Zineddine

KERIOUI Abderrahim

En vue d'obtenir le diplôme de Master
Domaine : Mathématique et informatique

Filière : Informatique
Spécialité : Informatique
Option : Système informatique et réseaux

Thème

Extractions des motifs fréquents orientés besoins du décideur.

Soutenu le :

M.	OUKID Lamia	Président
M.	LAHIANI Nesrine	Examinateur
Mme	ZAHRA FATMA ZOHRA	Promotrice
Mme	ZERF NADJET	Encadreuse

Promotion : 2018 / 2019

Résumé

L'accès à une information pertinente, adaptée aux besoins et au contexte du décideur est un challenge dans un environnement de l'organisation, caractérisé par une prolifération des ressources hétérogènes et distribuées. En effet, les entreprises analysent leurs données pour avoir un avantage concurrentiel stratégique.

Cette analyse peut se faire par un des outils tel que l'extraction des motifs fréquents qui consiste à découvrir les structures de données qui se répètent fréquemment dans un ensemble de données. Cette technique d'extraction délivre, dans des temps de plus en plus longs, des résultats massifs en réponse aux requêtes des décideurs, générant ainsi une surcharge informationnelle dans laquelle il est souvent difficile de distinguer l'information pertinente d'une information secondaire.

C'est dans ce contexte que s'intègre notre travail qui consiste à intégrer au processus d'extraction des motifs fréquents une phase de spécification des besoins du décideur basée sur le profil, et utiliser ces données pour raffiner les motifs fréquents extraits.

Les résultats obtenus montrent l'efficacité et l'utilité de la solution proposée à condition que le modèle de connaissances du décideur soit bien appris.

Mots-clés : Extractions des motifs fréquents, Personnalisation, profil du décideur, connaissances orientés besoins du décideur.

Abstract

Access to relevant information, tailored to the needs and context of the decision maker is a challenge in an organizational environment, characterized by a proliferation of heterogeneous and distributed resources. Indeed, companies analyze their data to have a strategic competitive advantage.

This analysis can be done by one of the tools such as the extraction of frequent patterns which consists in discovering data structures that are repeated frequently in a set of data. This extraction technique delivers, in longer and longer times, massive results in response to requests from decision makers, thus generating information overload in which it is often difficult to distinguish the relevant information from secondary information.

It is in this context that our work, which consists in integrating a phase of specification of the decision-maker's needs based on the profile, into the extraction process of frequent motives, and using this data to refine the frequently extracted patterns.

The results obtained show the effectiveness and usefulness of the proposed solution provided the decision maker's knowledge model is well learned.

Keywords: Extractions of frequent motives, Personalization, decision maker profile, decision-oriented knowledge of the decision maker.

.

ملخص

يمثل الوصول إلى المعلومات ذات الصلة، المصممة خصيصًا لاحتياجات وسياق صانع القرار، تحديًا في بيئة تنظيمية، تتميز بانتشار الموارد غير المتجانسة والموزعة. في الواقع ، تقوم الشركات بتحليل بياناتها للحصول على ميزة إستراتيجية.

يمكن إجراء هذا التحليل من خلال إحدى الأدوات مثل استخراج الأنماط المتكررة والتي تتمثل في اكتشاف هياكل البيانات التي تتكرر كثيرًا في مجموعة من البيانات. توفر تقنية الاستخراج هذه، في أوقات أطول وأطول ، نتائج هائلة في الاستجابة لطلبات صانعي القرار، مما يؤدي إلى زيادة الحمل على المعلومات التي يصعب فيها تمييز المعلومات ذات الصلة عن المعلومات الثانوية.

في هذا السياق ، عملنا ، الذي يتمثل في دمج مرحلة من تحديد احتياجات صانع القرار على أساس الملف الشخصي ، في عملية استخراج الدوافع المتكررة ، واستخدام هذه البيانات لتحسين الأنماط التي يتم استخراجها بشكل متكرر.

توضح النتائج التي تم الحصول عليها فعالية وفائدة الحل المقترح شريطة أن يكون نموذج المعرفة لدى صانع القرار جيدًا.

الكلمات المفتاحية: استخلاص الدوافع المتكررة ، التخصيص ، ملف تعريف صانع القرار ، المعرفة الموجهة نحو اتخاذ القرار لصانع القرار.

Remerciement

En préambule à ce mémoire, nous remercions ALLAH qui nous a aidé et donné la patience et le courage durant cette longue année d'étude.

Nous souhaitons adresser nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire.

Ces remerciements vont tout d'abord à notre promotrice Mme ZAHRA FATIMA ZAHRA Pour sa disponibilité tout en long de la réalisation de ce Mémoire, Ainsi pour son inspiration, aide et son suivi.

Nous remercions très chaleureusement aussi, Mme ZERF NADJAT, Notre encadreuse, pour sa confiance et ses encouragements.

Nos remerciements iront également vers tous ceux qui ont accepté avec bienveillance de participer au jury de ce mémoire.

On n'oublie pas nos parents pour leur contribution, leur soutien et leur patience.
Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours encouragées au cours de la réalisation de ce mémoire

Merci à tous et à toutes

Dédicaces

Je dédie ce modeste travail

À mes parents qui depuis mon plus jeune âge ont toujours fait leur

Maximum, en consacrant temps et argent, pour m'éveiller et

M'encourager dans mes passions. C'est grâce à vous et pour vous que

J'ai fait mon mémoire. Aucun mot sur cette page ne saurait exprimer

Ce que je vous dois, ni combien je vous aime. Qu'Allah vous bénisse,

Vous assiste, vous vienne en aide

A ma promotrice Mme ZAHRA FATMA ZAHRA

A mon encadreuse Mr ZERF NADJAT

A mon binôme ZINEDINNE

A mes chers frères, ABDERRAOUF ET DHIA

A tous mes collègues et toute la section Master2 SIR

A tous ceux qui m'ont soutenu, qu'ils trouvent ici l'expression de mon

Amour et ma profonde

Gratitude.

KERIOUI ABDERRAHIM

Dédicaces

Je dédie ce modeste travail

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.

A mes chères frères et sœurs pour leurs encouragements permanents, leur soutien moral et leur appui et leur encouragement.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire.

A mon binôme ABDERRAHIM.

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible.

Merci d'être toujours là pour moi.

J'exprime toute ma reconnaissance et gratitude à l'administration et à l'ensemble du corps enseignant de l'Université de (SAAD DAHLAB) pour leurs efforts à nous garantir la continuité et l'aboutissement de ce programme de Master.

Je tiens à remercier aussi et chaleureusement mes encadreuses Mme ZAHRA FATMA ZOHRA et Mme ZERF NADJAT de m'avoir permis de mener ce travail, pour leur engagement et leur soutien ainsi que pour la pertinence de leur remarques et de leur feed-back.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Enfin, Je tiens également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail, à mes amis, ma famille, Merci.

DAAOU ZINEDINNE

Sommaire

Chapitre 2: Personnalisation basée sur les profils

Chapitre 3 : Algorithmes d'extraction des motifs fréquents

Liste des figures

Liste des tableaux

Introduction Générale

1. Contexte du travail

Aujourd'hui les systèmes d'information donnent accès à un grand nombre de sources de données hétérogènes et distribuées. A fur et à mesure que les sources se multiplient et que le volume de données disponibles s'accroît, cela mène les entreprises à analyser leurs données pour avoir un avantage concurrentiel stratégique. Cette analyse peut se faire par un des outils d'aide à la décision telle que le processus d'extraction de connaissances (KnowledgeDiscovery).

L'aide à la décision facilite les tâches de prise de décision en simplifiant et en raccourcissant le chemin cognitif suivi par l'homme qui mène à la recherche d'informations pertinentes **[1].** Actuellement, et avec l'explosion de volume des données dans le monde entier, le décideur se voit confronté à une surcharge informationnelle dans laquelle il est difficile de distinguer l'information pertinente de l'information secondaire extraite à partir du processus d'extraction des connaissances.

2. problématique

Le problème de base est de disposer de supports d'aide à la décision pour améliorer le processus cognitif des décideurs et la connaissance de la situation décisionnelle. Il est devenu difficile de prendre une décision, cette difficulté a tendance à l'accroissement de plusieurs facteurs :

- Le volume de données est très grand.
- Les connaissances extraites n'expriment pas les besoins du décideur :
- Le temps de réponse élevé.
- le coût des erreurs de décision est plus grand, en raison de la complexité et de l'importance des conséquences engendrées par une décision.
- Le nombre des informations par les méthodes d'extraction de connaissance est parfois très important.

L'extraction des motifs fréquents est une technique de Data Mining (noyau du processus d'extraction de connaissances) qui consiste à découvrir les structures de données (appelées motifs) qui se répètent fréquemment dans un ensemble de données. Ces motifs constituent en

eux même une forme de connaissance comme ils peuvent être exploités comme une entrée pour d'autres méthodes de Data Mining et d'extraction de connaissances. Le nombre des motifs générés par les méthodes d'extraction de motifs est parfois très important et peut même être supérieur au nombre de données en entrée. Par conséquent, l'interprétation et l'exploitation de ces motifs devient une tâche difficile au décideur.

3. Objectifs

Les objectifs de ce travail sont de deux ordres : la modélisation des connaissances décrivant le décideur et son environnement d'une part, et l'exploitation et application de ce modèle décideur dans l'extraction des motifs fréquents d'autre part.

Notre travail se focalise justement sur le décideur. En effet, on s'intéresse à l'orientation de processus d'extraction de connaissance afin qu'il génère des connaissances utiles et pertinentes en prenant en considération les besoins du décideur. Ces connaissances sont en forme des motifs fréquents.

Plus précisément, l'objectif de ce travail est d'orienter le processus d'extraction des motifs fréquents afin de proposer des motifs fréquents pertinents et intéressants aux décideurs, en prenant en considération les points suivants :

- Modélisation du décideur : aider le décideur à mieux expliciter son besoin et formuler sa requête en élaborant le profil, le contexte et les préférences du décideur.
- Affinement et augmentation de la pertinence des d'informations que le processus d'extraction peut fournir pour mettre à la disposition du décideur uniquement la connaissance dont il a besoin.
- Définition des connaissances du décideur est dérivée par des algorithmes d'apprentissage à partir des actions passées.

4. Organisation du mémoire

A part l'introduction le mémoire se reparti en cinq chapitres

Chapitre 1 : « l'extraction des connaissances à partir de données»

Ce chapitre est consacré à la présentation des différents types d'algorithme pour l'extraction des items fréquents les plus connue.

Chapitre 2 : « La personnalisation basée sur les profils»

Ce chapitre a permis de cerner la thématique de la personnalisation de l'information dans sa globalité**.**

Chapitre 3 : « Algorithmes d'extractions des motifs fréquents»

Ce chapitre est consacré à la présentation des différents types d'algorithme pour l'extraction des motifsfréquents fondamentaux.

Chapitre 4 : « Approche proposé »

Ce chapitre, sera réservé pour exposer notre solution proposée.

Chapitre 5 : « Tests et Expérimentation »

Le dernier chapitre concrétise et valide la méthode adoptée**.**

Extraction des connaissances à partir des données

1. Introduction

L'informatique est devenu un outil important dans tous les domaines (l'industrie, le finance, la médecine, etc.), on le utilise pour facilite les taches et diminue le budget et son but c'est de stocker et traiter les données mais ses donnes à augmenter d'une façon très rapide et sa qu'a inspirer les informaticiens de profiter de cette quantité de données et on inventer l'extraction des connaissances à partir de données(ECD).

En effet, ces données volumineuses causent beaucoup de problèmes dans leur analyse comme la représentation, la recherche, l'exploration, etc. L'ECD (knowledgediscovery in databases (KDD)) utilise les données collectées par des experts, pour proposer des connaissances nouvelles qui enrichissent les interprétations du champ d'application, tout en fournissant des méthodes automatiques qui exploitent cette information.

Dans ce chapitre nous définissons processus d'extractions des connaissances à partir des données et aussi détaillé chaque étapes de ECD. Nous commençons par définir ECD et leur étape après on va mentionner les méthodes et les techniques de datamining.

2. L'Extraction de connaissances à partir des données

L'extraction de connaissances à partir des données (ECD) se définit comme un processus non trivial d'identification de modèles valides, nouveaux, potentiellement utiles, compréhensibles à partir d'une base de données (Fayyad et al., 1996) **[2].** En fait, on cherche surtout à isoler des traits structuraux (patterns) qui soient valides, non triviaux, nouveaux, utilisables et si possible compréhensibles ou explicables.

En analysant les termes utilisés dans cette définition, nous obtenons une définition plus pertinente :

- Modèle : c'est une représentation d'un phénomène réel, le plus souvent constitués d'objets mathématiques comme par exemple des ensembles de données, des tables, des matrices, des fonctions, des relations, des listes de règles, des systèmes d'équ²ations, des arbres, des graphes, des hypergraphes, des réseaux, des opérateurs fonctionnels linéaires et non-linéaires, etc.
- Processus : c'est l'exécution d'un ensemble de plusieurs tâches qui peut être exécuté d'une façon itérative.

- Processus Non-trivial : c'est un processus qui exécute des tâches dans un ordre spécifique
- valide : un modèle valide est un modèle appliqué à une base de test doit renvoyer un degré de certitude.

Discipline émergente et multifocale, rassemblant les travaux des chercheurs en statistiques, intelligence matricielle, apprentissage automatique, reconnaissance de formes, bases de données, visualisation des données et linguistique, ECD génère des techniques et des outils permettant la révélation de connaissances enfouies dans d'énormes quantités de données hétérogènes et protéiformes. **[2]**

3. Les étapes du processus ECD

ECD est un processus anthropocentré, les connaissances extraites doivent être les plus intelligibles possibles à l'usager. Elles doivent être validées, mises en forme et agencées. Nous allons détailler toutes ces notions et les situer dans le processus général de ECD, se processus et schématisé dans la figure 1.1.

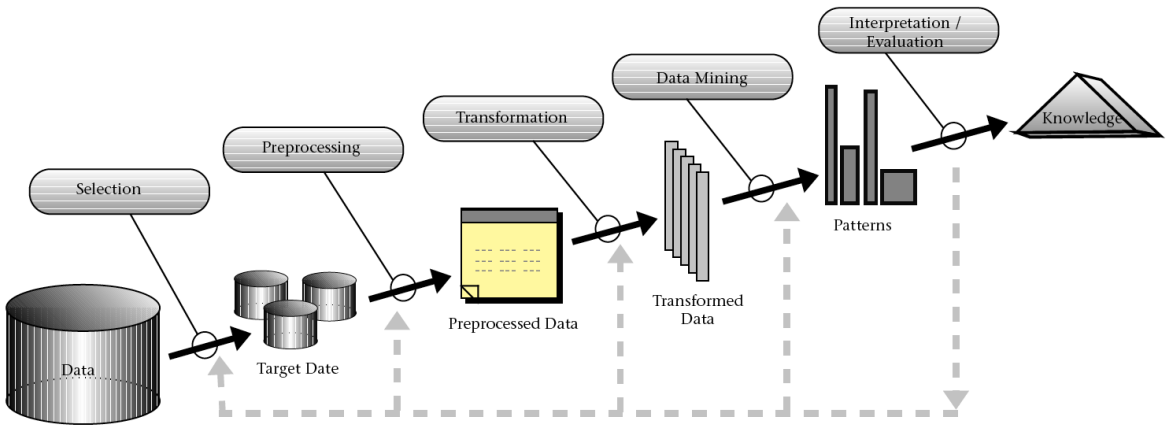


Figure 1.1: Processus ECD. **[2]**

Le processus ECD se décompose en plusieurs étapes, les différentes phases sont les suivant :

3.1. Sélection

Les données nécessaires au processus d'exploration de données peuvent être obtenues à partir de nombreuses sources de données différentes et hétérogènes. Cette première étape permet d'obtenir les données de diverses bases de données, fichiers et sources non électroniques. **[4]**

Cette phase ne se limite pas à la seule sélection des données qui vont être exploitées par le système ECD. Elle comprend également l'analyse du problème à résoudre**,** ce qui permet d'en déduire le ou les types de données qui sont exploitées, ainsi que les méthodes qui pourraient être utilisées pour résoudre ce problème. **[5]**

Un système ECD idéal est un système qui nécessite l'intervention d'aucune entité, c'est-à-dire un système automatisé qui va extraire de nouvelles connaissances à partir de grandes bases de données mises à sa disposition sans l'intervention de l'utilisateur. Actuellement, ce type de système présente de nombreux inconvénients. Le premier de ceux-ci est la perte de temps et de ressources nécessaires à l'exploitation de l'ensemble des données disponible au système**.[5]**

Parmi les conséquences de ces inconvénients, on peut citer :

– Les recherches lancées par le système peuvent toucher divers domaines ou thèmes qui n'ont aucun rapport avec l'objectif défini par l'utilisateur.

– Le système peut fournir des connaissances qui ne présentent aucun intérêt ou sont incompréhensibles pour l'utilisateur

– L'utilisateur submergé de nouvelles connaissances, ne peut distinguer des connaissances proposées celles qui lui sont réellement intéressantes.

Ceci implique que l'utilisateur doit avoir la possibilité de communiquer avec le système afin d'orienter la recherche selon ses objectifs. Pour faciliter la communication de l'utilisateur avec le système ECD, un ensemble de primitives (data mining primitives) a été conçu. Ces primitives incluent :

- ✓ Spécification des données
- ✓ Spécification du type de connaissances à extraire
- ✓ Spécification des connaissances préalables

- ✓ Spécification de la mesure
- ✓ Représentation de la connaissance extraite

3.2. Prétraitement

Les données à utiliser par le processus peuvent contenir des données incorrectes ou manquantes. Il peut y avoir des données anormales provenant de plusieurs sources impliquant différents types de données et mesures. Il peut y avoir aussi beaucoup d'activités différentes effectuées à ce moment. Les données erronées peuvent être corrigées ou supprimées, tandis que les données manquantes doivent être fournies ou prévues (souvent à l'aide d'outils d'exploration de données). **[4]**

Les données à analyser par les méthodes de data mining sont parfois incomplètes, inconsistantes, erronées, incompatibles entre elles, inadaptées ou encombrantes. Ces types de données sont courants et se retrouvent régulièrement dans les bases de données et d'entrepôts de données. **[5]**

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

La procédure de nettoyage des données : elle se compose des tâches de traitement des données manquantes, de traitement des données erronées et inconsistantes.

- Tâche de traitement des données manquantes : Plusieurs méthodes permettent d'accomplir cette tâche. Le choix de la méthode dépend des données et de l'objectif de l'étude.
  1. méthode consiste à ignorer les instances incomplètes
  2. méthode consiste à compléter les données manuellement
  3. méthodes qui consistent à compléter les données incomplètes à l'aide de constantes globales
  4. méthodes qui consistent à remplacer la valeur manquante d'un attribut par la valeur moyenne de cet attribut
  5. méthodes qui remplacent la valeur manquante par la valeur la plus probable

- Tâche de traitement des données de type bruit
  1. méthode de groupement

2. méthode combinant une solution algorithmique à l'utilisation d'un expert
3. méthode de régression.

3.3. Transformation

Les données provenant de différentes sources doivent être converties en un format commun pour le traitement. Certaines données peuvent être encodées ou transformées en formats plus utilisables. La réduction des données peut être utilisée pour réduire le nombre de valeurs de données possibles considérées. **[4]**

Permet de modeler les données sous une forme exploitable par les méthodes de data mining.

1. méthode d'agrégation
2. méthode de généralisation des données
3. méthode de normalisation
4. méthode d'ajout d'attributs

La procédure de réduction des données permet de réduire la taille des données tout en gardant leur intégrité.

Les méthodes de réduction les plus connues sont :

1. Agrégation des données cibles
2. Réduction dimensionnelle
3. Compression des données
4. Discrétisation et génération de concept hiérarchique

3.4. Data Mining (Exploration de données)

En fonction de la tâche d'exploration de données en cours d'exécution, cette étape applique des algorithmes aux données transformées pour générer les résultats souhaités. **[4]**

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

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

Les méthodes de fouille de donnée qui sont les plus couramment utilisées dans les systèmes ECD sont les méthodes de type classification, régression, structuration et association

– Méthodes de classification et de structuration (algorithme des k-moyennes (k-means), algorithme du plus proche voisin),

– Méthodes d'explication et de prédiction (arbre de décision, réseaux de neurones, réseaux bayésiens, règles d'associations),

– Méthodes de visualisation et de description.

3.5. Interprétation / évaluation

La manière dont les résultats de l'exploration de données sont présentés aux utilisateurs est extrêmement importante, car leur utilité en dépend. Diverses stratégies de visualisation et d'interface graphique sont utilisées à cette dernière étape. **[4]**

4. Les niveaux d'extraction des connaissances à partir des données

L'introduction de l'ECD dans les entreprises est récente. Le rattachement des activités liées à l'ECD n'est pas toujours clair. Selon les cas, elle peut être intégrée au service ou la direction : informatique, organisation, études, statistique, marketing, etc. Comme le montre le schéma de la figure 1.2, il convient de distinguer le niveau opérationnel et le niveau « analyse » que nous allons décrire. **[6]**

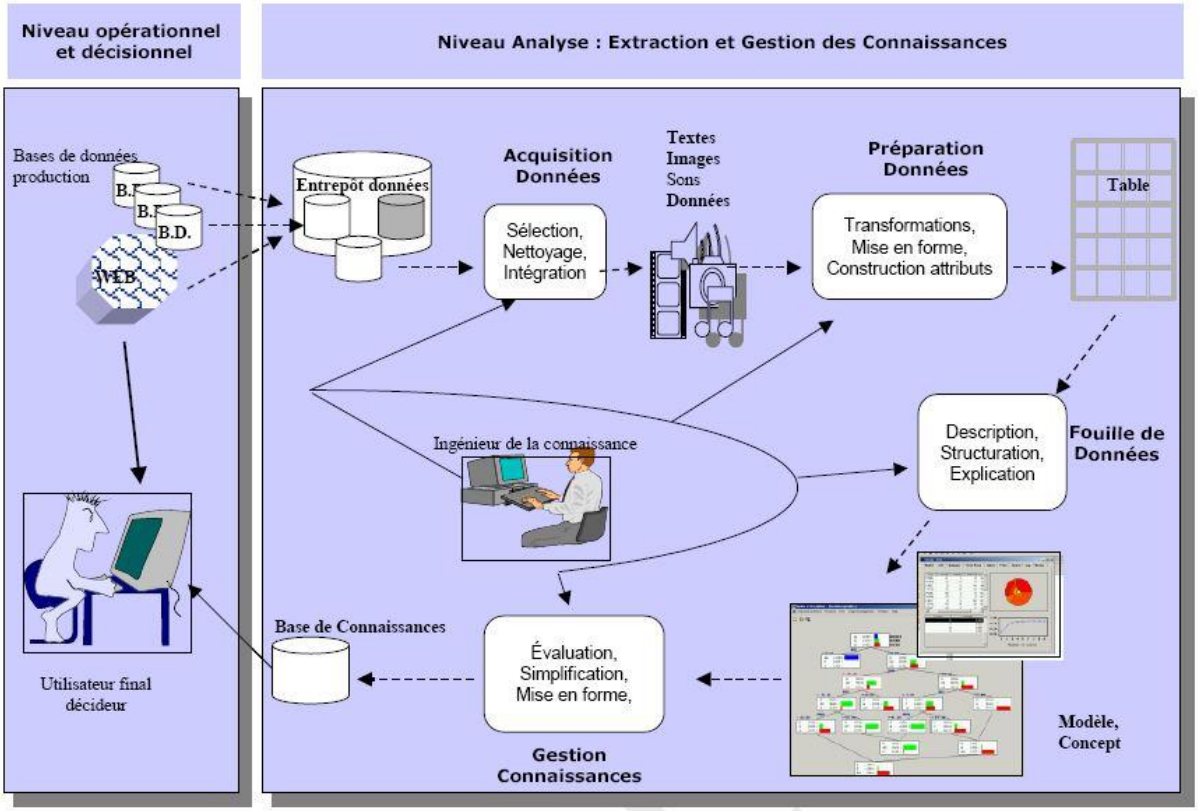


Figure 1.2: Niveaux de processus ECD**[6].**

4.1. Niveau opérationnel et décisionnel

Toutes les actions sont très souvent le résultat d'une décision prise pour répondre à une demande de l'environnement. Ces décisions ou ces actions ne sont bien sûr pas toutes de même importance. Elles peuvent être stratégiques ou de simples actions de routine. Les décisions importantes nécessitent une évaluation qui repose sur des connaissances ou des modèles préétablis. A ce niveau, l'utilisateur cherche à répondre au mieux aux sollicitations de l'environnement**.** **[6]**

4.2. Niveau analyse

C'est le centre des opérations d'extraction des connaissances à partir des données. Les données issues des bases de données de production, en service en front office, alimentent les entrepôts de données qui seront utilisées en ECD. Généralement, le processus d'ECD, sous la supervision d'un spécialiste, se déroule en quatre phases : acquisition des données, prétraitement et mise en forme, fouille de données (data mining dans un sens restrictif) et analyse, validation et mise en forme des connaissances**.[6]**

5. Les techniques de Data Mining (Exploration de données)

Alsagheer R.H.A. et al [7] a dit que il existe un grand nombre de bonnes techniques d'extraction et d'extraction de données. Ces techniques impliquent l'association, le regroupement, la régression et la classification, comme indiqué sur la figure 1.3.

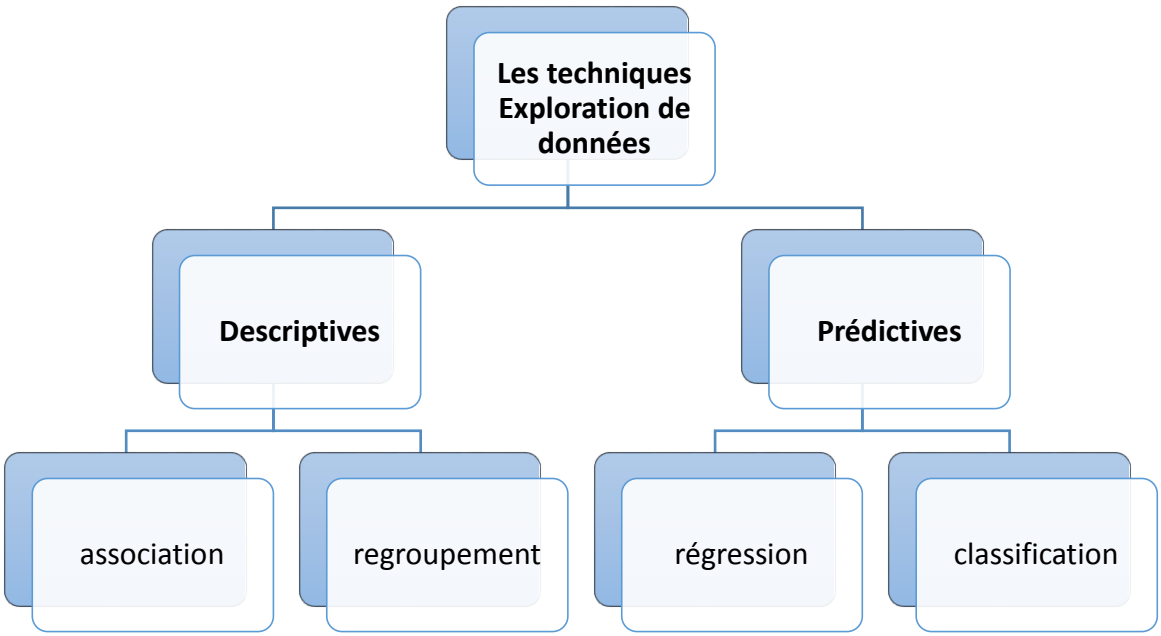


Figure 1.3: Les techniques de Data Mining. [7]

5.1. Règles d'Association

L'association permet de trouver des liens cachés entre des variables différentes dans des bases de données. Il expose des modèles ambigus dans les données, pour tirer le meilleur des règles à partir d'autres règles choisit différentes mesures d'importance. La meilleure mesure est constituée des seuils les plus bas en matière de support et de confiance.

5.2. Le clustering

Les ont deux buts c'est de minimiser la distance intra-classe (grappes d'éléments homogènes) et de maximiser la distance inter-classe afin d'obtenir des sous-ensembles le plus distincts

Possible. La mesure des distances est un élément prépondérant pour la qualité de l'algorithme de clustering.

Ils existent plusieurs algorithmes de classification, l'algorithme de K-mean est parmi les

Algorithmes de clustering les plus répandu dans la littérature.

5.2.1 Algorithme K-mean

L'algorithme k-means mis au point par McQueen, (1967) **[8]** est l'un des plus simples algorithmes d'apprentissage non supervisé qui permettent de résoudre le problème de clustering bien connu.

Maitrise, (2014)**[9]** a donné l'idée principale c'est de définir k centres de gravité, un pour chaque cluster. Ces centres de gravité qui devraient être placés d'une manière rusée raison de l'emplacement différent causes résultat différent. L'étape suivante consiste à prendre chaque point appartenant à un ensemble de données et de l'associer au centre de gravité le plus proche. Lorsqu'aucun point n'est en cours, le groupage début est fait.

A ce stade, nous avons besoin de recalculer k nouveaux centres de gravité comme barycentres des groupes issus de l'étape précédente. Après nous avons ces nouveaux centres de gravité k, a une nouvelle liaison à faire entre les mêmes points de consigne de données et le nouveau centre de gravité le plus proche. Une boucle a été générée. À la suite de cette boucle, nous pouvons remarquer que les k centroïdes changent leur étape de localisation jusqu'à ce qu'il n'y plus de changements sont effectués.

L'algorithme est résumé dans les étapes suivantes:

1. Placez K des points dans l'espace représenté par les objets qui sont regroupés. Ces points représentent des centroïdes de groupe initial.
2. Attribuez à chaque objet pour le groupe qui a le centre de gravité le plus proche.
3. Lorsque tous les objets ont été assignés, recalculer les positions des centres de gravité de K.
4. Répétez les étapes 2 et 3 jusqu'à ce que les centres de gravité ne bougent plus. On obtient ainsi une séparation des objets dans des groupes dont la métrique à être réduite au minimum peut être calculée.

Des exemples de résultats de clustering sont montrés dans la figure 1.4.

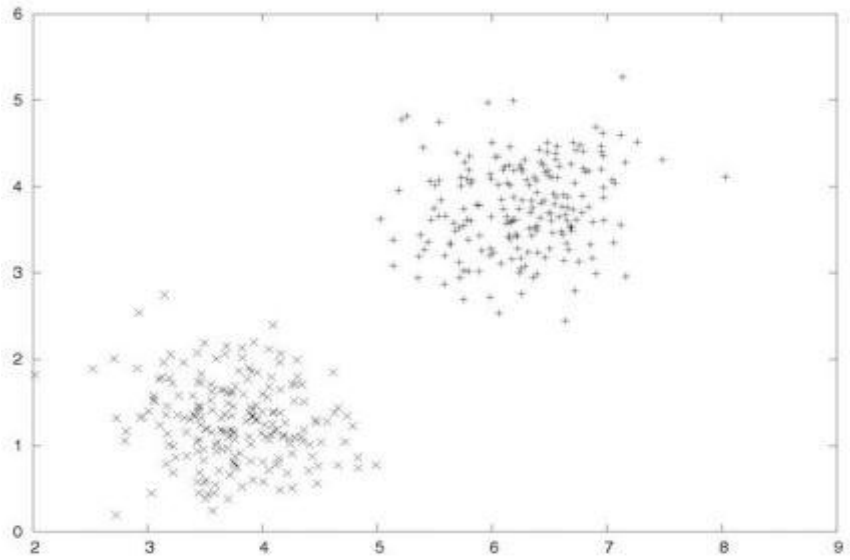


Figure 1.4: Résultats de clustering de donnée, deux clusters sont affichés. **[10]**

5.3. La classification

La classification permette à partir des modèles entrés de prédire les valeurs nominales et numériques. Ces outils permettent d'obtenir une valeur chiffrée résultante des données et expériences à partir d'algorithme qui selon leur performance répondent à des indications précises sur l'attribut étudié. **[9]**

Ils existent plusieurs méthodes de classification, on va présenter dans la section suivante deux importantes méthodes qui sont les arbres de décision et les réseaux bayésiens.

5.3.1 Arbre de décision

La construction des arbres de décision à partir de données est une discipline déjà ancienne. Les statisticiens en attribuent la paternité à Morgan et Sonquist (1963) **[11]** qui son les premiers qui ont utilisé les arbres de régression dans un processus de prédiction et d'explication (AID – Automatic Interaction Detection).

Caron, (2011)**[12]** à bien parler sur l'arbre de décision, a dit qu'elle modélise une hiérarchie de tests sur les valeurs d'un ensemble de variables appelées attributs. À l'issue de ces tests, le prédicteur produit une valeur numérique ou choisit un élément dans un ensemble discret de conclusions. On parle de régression dans le premier cas et de classification dans le second. Par

exemple, l'arbre de la figure 1.5 décide une réponse booléenne (classification dans l'ensemble {oui, non}) en fonction des valeurs discrètes des attributs {difficile, durée, motivation, surprenant}.

Un ensemble de valeurs pour les différents attributs est appelé une « instance », que l'on note généralement (x, y) où y est la valeur de l'attribut que l'on souhaite prédire etx = x1, . . . ,xm désignent les valeurs des m autres attributs. L'apprentissage d'un arbre de décision se fait sur un ensemble d'instances T = {(x, y)} appelé « ensemble d'entraînement », l'algorithme est résumer dans la figure 1.6.

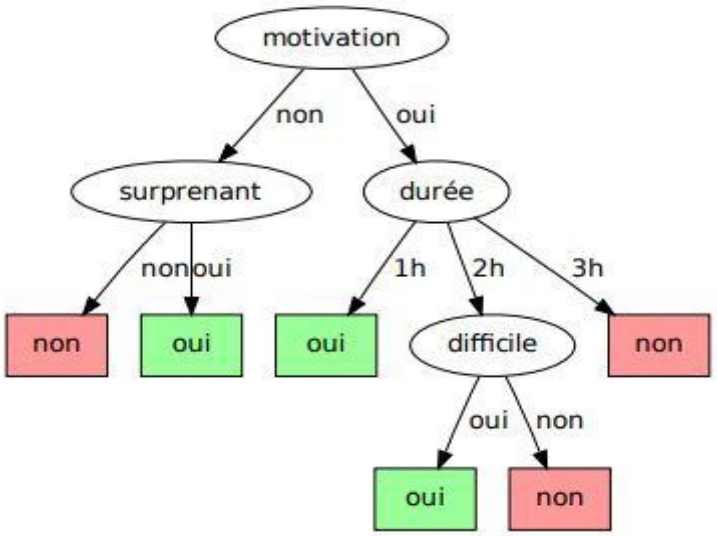


Figure 1.5: Arbre de décision. [7]

```
1:  ArbreDecision(T)
2:      si "condition d'arret"
3:          retourner feuille(T)
4:      sinon
5:          choisir le "meilleur" attribut i entre 1 et m
6:          pour chaque valeur v de l'attribut i
7:              T[v] = {(x, y) de T tels que x_i = v}
8:              t[v] = ArbreDecision(T[v])
9:          fin pour
10:         retourner noeud(i, {v -> t[v]})
11:     fin si
```

Figure 1.6: Algorithme de construction d'un arbre décisionnel. [12]

5.3.2 Réseaux bayésiens

Les réseaux bayésiens **[5]** sont basés sur les probabilités conditionnelles. Ils permettent en quelque sorte de prédire le futur à partir du passé, en supposant la reproductibilité des probabilités. Cette technique dérive du théorème de Bayes qui permet de calculer la probabilité d'une hypothèse A sachant que E a été observé. Ainsi, la construction d'un réseau bayésien permet de trouver des règles avec des probabilités transitionnelles.

Un réseau bayésien **[13]** est un graphe orienté acyclique, dans lequel chaque sommet correspond à une variable aléatoire du domaine. Un arc X -> Y, décrit une relation père fils dans laquelle X est le père et Y le fils. De plus, à chaque sommet est associée une table de probabilités conditionnelles, spécifiant la probabilité de chaque état du sommet étant donné la combinaison d'états de ses parents, exemple de ce graphe et montre dans la figure 1.7.

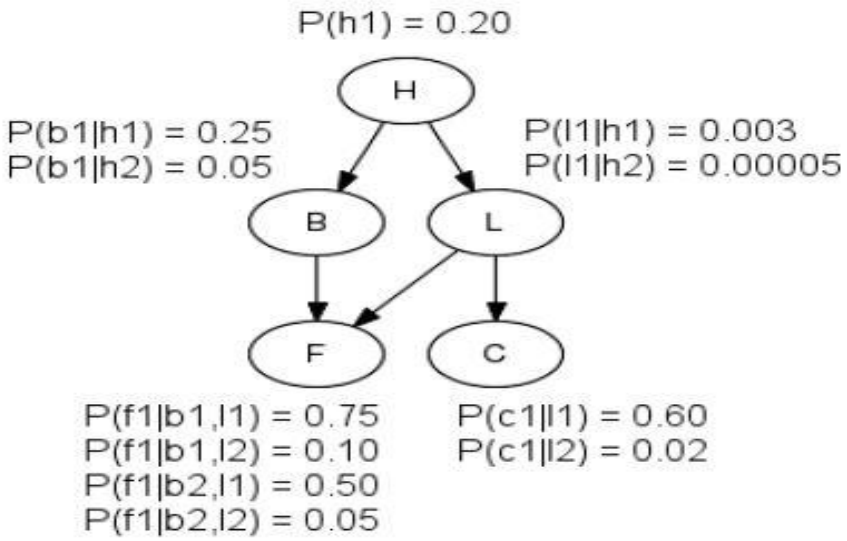


Figure 1.7: Exemple de reseaux baysien.**[13]**

5.4. La régression

La régression est une technologie qui permet à l'analyse de données de caractériser les liens entre variables. Il utilise pour obtenir de nouvelles valeurs repose sur la présentation des valeurs. Il utilise la régression linéaire pour les cas simples mais les cas complexes difficiles à prédire utilisent le déclin relatif car il repose sur des interactions complexes de plusieurs variables. **[14]**

6. Extractions des connaissances à partir de données dans le domaine médical

Les algorithmes de data mining et les cadres de ECD ont été appliqués avec succès dans un certain nombre de domaines d'application, notamment les télécommunications, le commerce, astronomie et sécurité.

John F.Roddick et al. (2003) **[14]** disent que pour un certain nombre de raisons, l'application de la technologie à de grands ensembles de données médicales a nécessairement été plus circonspecte. Cependant, alors que l'exploration de données à partir de données médicales et cliniques peut être problématique, Par exemple, les ensembles de données disponibles vont de ceux qui sont pratiques, précis, indexés et bien gérés à ceux qui sont incomplets, incohérents, potentiellement inexacts et extrêmement volumineux. De plus, contrairement à d'autres domaines (sans doute plus simples), le domaine médical La discipline elle-même est diversifiée, complexe et, pour un étranger, relativement opaque. Elle nécessite donc une collaboration active entre le spécialiste du domaine et le data miner.et il promet des récompenses substantielles et a suscité un certain intérêt.

6.1. Les problèmes dans extraction des données médicales

John F.Roddick et al. (2003) **[14]** a parlé sur quelques-uns des problèmes majeurs qui compliquent l'exploration de données médicales exploratoires par rapport aux divers autres domaines dans lesquels elle s'avère utile.

6.1.1. La méthode d'investigation

Par exemple, le paradigme dominant de la recherche médicale, la méthode expérimentale à hypothèse nulle, utilisée dans la recherche médicale / bioscientifique, diffère de celle adoptée par de nombreux autres domaines et nécessite souvent une modification du cadre de découverte des connaissances utilisé. Par exemple, nous pouvons rechercher une absence de données en conflit ou générer un ensemble de règles attendues, qui sont ensuite comparées à un ensemble généré à partir de l'hypothèse nulle et pour lesquelles nous recherchons un manque d'intersection entre elles.

6.1.2. Soutien longitudinal, temporel et spatial

L'incidence de la maladie et les méthodes de traitement sont étroitement liées à leur ordre et à leur présence temporelle et spatiale par rapport à d'autres épisodes, intervalles ou lieux.

L’utilisation de techniques statiques simplifie donc excessivement les relations possibles et prend ainsi en charge la sémantique longitudinale, temporelle et spatiale dans le processus d’exploitation minière est hautement souhaitable. Les données épisodiques sont souvent la clé pour bonne exploration des données.

6.1.3. Ensembles de données externes

La santé d’une personne étant souvent liée à son environnement, la prise en compte des ensembles de données médicales standard est essentielle pour une interprétation raisonnable.

6.1.4. Interprétation des règles et utilisation d'une base de connaissances considérable

L'interprétation des résultats de l'extraction sur des jeux de données médicales nécessite une expertise de domaine importante. Il est un peu plus simple pour les développeurs de routines d’exploration de données de comprendre les résultats des routines répétées. La nécessité d’une coopération étroite entre le chercheur en informatique et le ou les experts du domaine médical est donc plus pressante.

6.1.5. Disponibilité et précision des données

Les données générées par la pratique médicale diffèrent de ceux d'un d'autres domaines dans lesquels l'exploration de données a été appliquée.

- **Problèmes d'encodage**

Bien que de nombreux jeux de données médicales soient généralement de grande qualité et que de plus en plus de détails soient enregistrés électroniquement, beaucoup ne sont pas codés de manière à pouvoir être utilisés immédiatement.

- **Problèmes éthiques et juridiques**

Les exigences éthiques relatives au traitement de données médicales, même confidentielles, sont importantes. Peu de travaux ont été entrepris dans ce domaine, bien que la notion de compromis statistique soit bien comprise.

- **La possession**

La propriété des données peut facilement entraver les efforts pour obtenir les données nécessaires ou pour créer des liens entre des ensembles de données. L’obtention des autorisations nécessaires peut représenter une grande partie du temps de préparation.

7. Conclusion

L’extraction des connaissances a partir de donnée rassemble les recherches actuelles sur le problème passionnant de la découverte de connaissances utiles et intéressantes dans les bases de données.

Dans ce chapitre on a défini le processus ECD et parler sur chaque étape et on a motionnait quelque problèmes qu’on peut rencontrer dans chaque étape et aussi on a défini les méthodes global de data mining.

Pour avoir des résultats pertinents et couvre nos besoins et touche nos aspirations on doit avoir un bon déroulement de processus ECD on a besoin de trouver une solution a ses problèmes qu’on peut rencontrer.

Personnalisation basée sur les profils

1. Introduction

Les systèmes d'information actuels donnent accès à un grand nombre de sources de données hétérogènes et distribuées. Au fur et à mesure que les sources se multiplient et que le volume de données disponibles s'accroît, l'utilisateur se voit confronté à une surcharge informationnelle dans laquelle il est difficile de distinguer l'information pertinente de l'information secondaire **[1].** En outre, il est à remarquer que l'évaluation des requêtes se fait sans tenir compte de l'utilisateur qui les a émises ni du contexte d'où elles sont émises. La même requête, faite par deux utilisateurs différents, produit les mêmes résultats même si ces utilisateurs n'ont pas les mêmes attentes. **[15]**

Pour pallier à ce problème et pour pouvoir discriminer les utilisateurs en fonction de leurs besoins spécifiques, certains systèmes proposent des techniques de personnalisation basées sur le profil de l'utilisateur. Ces techniques tendent vers un objectif commun : délivrer, en un temps acceptable, une information pertinente et de qualité en fonction de caractères spécifiques de l'utilisateur**. [16]**

Le but de la personnalisation est de faciliter l'expression du besoin de l'utilisateur et de lui permettre d'obtenir des informations pertinentes lors de ses accès à un système d'information ; et cela se fait généralement par la prise en compte d'un ensemble de connaissances que l'on nomme « profil » qui permet de décrire les utilisateurs ainsi que leurs préférences dans l'environnement où ils se trouvent.

A cet effet nous allons aborder dans ce chapitre le concept de personnalisation et les grandes questions qui le sous-tendent. Lors de notre recherche bibliographique, nous avons pu lire de nombreux ouvrages et articles traitant tous de la personnalisation, mais chacun abordant le sujet sous un angle différent. L'objectif de ce chapitre est de présenter chaque concept rencontré dans la littérature comme une réponse possible à l'une des grandes questions suivantes : -*Qu'est-ce que la personnalisation* ? -*Que personnaliser* **?** et-*Comment personnalisé?*

2. Quelques notionssur la personnalisation

Le mot « personnalisation » figure d'une façon courante dans différents sujets de la littérature, On trouve beaucoup de description, par exemple :

Kostadinov (2003)**[17]** définit la personnalisation de l'information par un ensemble de préférences individuelles, par des ordonnancements de critères ou par des règles sémantiques spécifiques à chaque utilisateur ou communauté d'utilisateurs. Ces modes de spécification servent à décrire le centre d'intérêt de l'utilisateur, le niveau de qualité des données qu'il désire ou des modalités de présentation de ces données.

Le Gartner Group définit la personnalisation comme « *toute interaction avec l'utilisateur dans laquelle le message, l'offre ou le contenu a été taillé sur mesure pour un utilisateur ou groupe d'utilisateur spécifiques* » **[18].** Cette définition est très riche. Elle met l'accent sur le caractère bidirectionnel d'une démarche personnalisée et spécifie divers aspects : les messages adressés directement à l'utilisateur, l'offre de services ou de produits, et encore le contenu informationnel.

3. Les profils

La notion de profil utilisateur est apparue vers les années 80 avec les assistants et les agents d'interface due principalement au besoin de créer des applications personnalisées, capables de s'adapter à l'utilisateur.**[19]**

Un profil regroupe l'ensemble des connaissances nécessaire à une évaluation efficace des requêtes et à une production d'une information pertinente adaptée à chaque utilisateur. Un profil peut être défini comme un modèle personnalisé d'accès à l'information alors qu'une requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait en tenant compte de son profil. Un profil a un caractère plus invariant que les requêtes même si le centre d'intérêt et les préférences de l'utilisateur peuvent légitimement évoluer.

4. Construction d'un profil utilisateur

Un problème important qui se pose dans le cadre des applications personnalisées est la construction des profils utilisateurs. Plusieurs travaux ont adressé ce problème dans différents

domaines. De tels exemples sont les travaux de Mobasher et al. **[20], [21],** qui présentent la personnalisation comme une application de techniques de data mining pour l'apprentissage automatique des profils, ou encore le travail présenté dans **[22],** qui propose une description de techniques de construction de profils dans le domaine des bases de données. La grande diversité des approches existantes nécessite l'identification d'un ensemble de paramètres qui permettant de les classer, Parmi les caractéristiques d'un processus de construction des profils :

- Le moment où le profil est construit:Le processus de construction de profils utilise des techniques très couteuses basées sur l'apprentissage automatique **[23], [24], [25], [26], [27]** ou sur la théorie des graphes **[23], [28], [29].**
- L'implication de l'utilisateur: Selon ce critère, les approches se divisent en deux groupes : celle qui n'impliquent pas l'utilisateur dans la construction du profil **[23]** et celles qui impliquent l'utilisateur en lui demandant de remplir un formulaire (EX : ouverture d'une boite mail) ou de corriger manuellement ses préférences **[30].**
- Le type de sources de données utilisées pour la construction des profils: Différentes types de données peuvent être exploitées pour la construction des profils utilisateurs. Des exemples sont les données d'usage **[21],** les données de contenu **[28],** les données décrivant la structure et l'organisation des documents **[31]** ou encore les données provenant de l'utilisateur **[32].**
- La manière de mise à jour des profils: Lorsque de nouvelles données sont collectées sur l'utilisateur ou sur son comportement, le profil doit être mis à jour pour tenir compte de ces données. On distingue selon ce critère des systèmes qui mettent à jour uniquement les parties du profil utilisateur concernées par les nouvelles données et les approches qui construisent à nouveau le profil utilisateur

5. Domaines d'utilisation des profils

La personnalisation a été particulièrement abordée dans trois domaines technologiques : l'Interaction Homme-Machine (IHM), la Recherche d'Information (RI), et les Bases de Données(BD).

5.1.Domaine de l'Interaction Homme-Machine

Dans le domaine des IHM **[33] [34],** la personnalisation se focalise principalement sur le niveau d'expertise et le métier de l'utilisateur afin de déterminer le type de dialogue que le système va avoir avec lui. La notion de requête n'existe pas sous forme langagière. Les systèmes utilisent des connaissances sur l'utilisateur (âge, niveau d'expertise, handicaps etc.) ou sur la technologie qu'il utilise (type du media, logiciels etc.) pour lui fournir une interface d'interaction adaptée. Un exemple de tel système est 'Apt Décision' qui représente un agent de recherche d'appartements **[35].**

Initialement l'utilisateur soumet un certain nombre de critères de recherche (nombre de pièces, surface etc.) et ensuite par le biais de l'interaction, le système guide l'utilisateur à travers les annonces disponibles. A chaque étape, le système analyse les actions que l'utilisateur effectue sur les annonces affichées pour lui proposer, dans la prochaine itération, des appartements conformes à ses préférences.

5.2. Domaine de la Recherche d'Information

Dans le domaine de la RI **[36], [37], [38], [39], [40],** l'utilisateur fait partie du processus de personnalisation qui est vu comme un processus incrémental et interactif dans lequel l'utilisateur décide à chaque pas quels sont les éléments qu'il aime et quels sont ceux qu'il n'apprécie pas. La personnalisation est ainsi définie comme un apprentissage réalisé à partir des préférences rendues par les utilisateurs à l'issue de la présentation des résultats successifs

Ici le profil est présenté comme un vecteur à N dimensions où les dimensions sont définies par les termes les plus significatifs pour les documents recherchés, le système calcule le matching entre le profil et les mots clés significatifs extraits des documents en utilisant une technique basée sur la distance entre vecteurs à N dimensions. Seuls les documents dont le matching dépasse un certain seuil (spécifié par l'utilisateur) sont inclus dans le résultat

5.3. Domaine des Bases de Données

Dans le domaine des BD **[41],[42],[43],[44],[45],[46],[47],[48**] l'utilisateur ne fait pas partie du processus de recherche d'informations. La requête contient en général l'ensemble des critères considérés nécessaires à produire des données pertinentes. Les profils sont alors

intégrés directement aux requêtes par les utilisateurs ou lors de la compilation de ces dernières; ils sont alors pris en compte en une seule fois durant l'exécution de la requête.

6. Les travaux sur la modélisation des profils

La modélisation de l'utilisateur consiste à désigner une structure pour stocker toutes les informations qui caractérisent l'utilisateur et qui décrivent principalement ses centres d'intérêts en plus d'autre informations relatives à ses préférences, le contexte dans lequel il travaille, le but et les objectifs de sa recherche, ses traits de personnalité, son style cognitif, son background et son expérience **[49].** Dans ce qui suit, nous allons passer en revue les travaux sur la modélisation de l'utilisateur.

6.1 Les standards P3P

Tout à commencer par l'élaboration des standards P3P (Platform for Privacy Préférences Project) **[50]** pour la sécurisation des profils des internautes. P3P propose les classes d'attributs suivantes : attributs démographiques, les attributs professionnels et les attributs de comportement (traces de navigation sur le web). Par la suite, plusieurs auteurs ont enrichi ces dimensions.

6.2 Travaux de Bouaka

N.Bouaka**[51]** a proposé le modèle MEPD(Modèle d'Explicitation du Problème Décisionnel) qui est un outil d'aide à l'analyse du besoin permettant au décideur d'obtenir la vision la plus claire possible du problème décisionnel et de faciliter sa traduction en termes d'enjeu relatif au contexte qui l'a engendré.

Le modèle regroupe des paramètres statiques (l'identité, le cursus et le style cognitif du décideur) qui sont persistants aux différents problèmes décisionnels et des paramètres dynamiques (les caractéristiques de l'environnement ou de l'organisation) dont les valeurs changent selon le contexte et les problèmes rencontrés. La figure 2.1 décrit le schéma global de ce modèle.

On peut résumer l'objectif et l'apport de cemodèle par les points suivants :

1. Parvenir à une définition claire et sans ambigüité du problème décisionnel.
2. Permettre le passage du problème décisionnel au problème de recherche d'information.
3. Contribuer à la sélection des informations pertinentes susceptibles d'aider le décideur dans son processus de décision

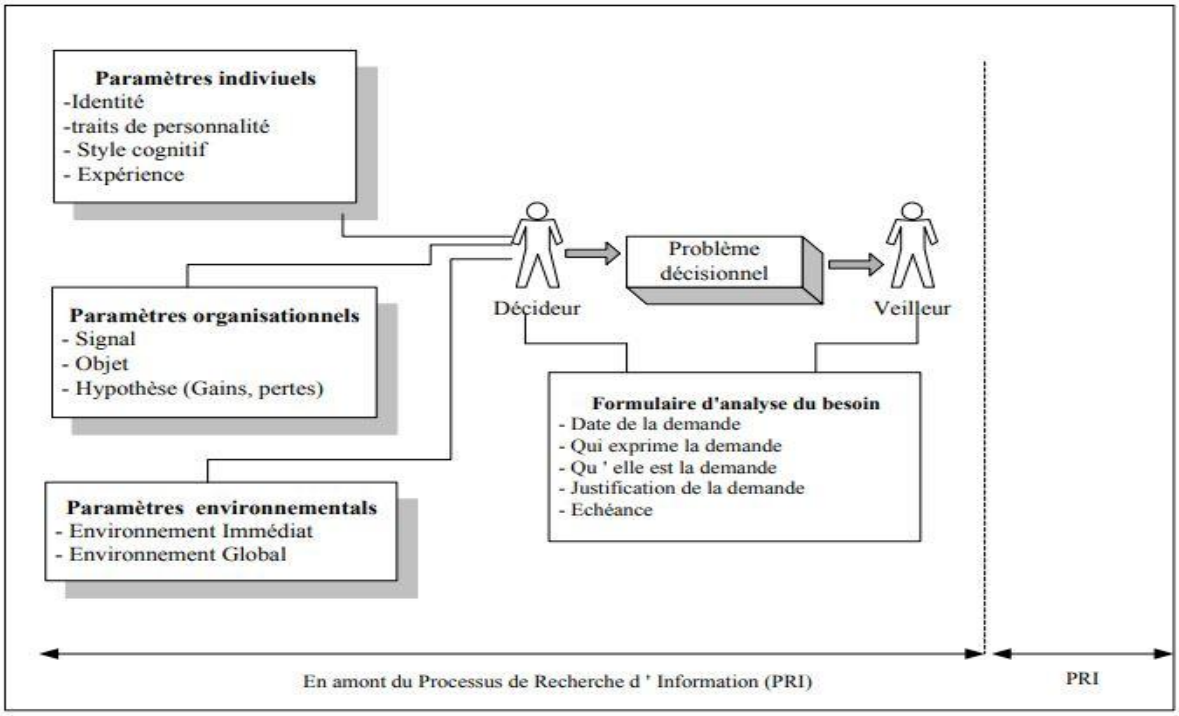


Figure 2.4: Le schéma général du modèle **[51**

6.3.Travaux de Kostadinov

Kostadinov**[17]** a identifié plusieurs dimensions : le centre d'intérêt, les données démographiques, l'ontologie du domaine, la qualité, la customisation, la sécurité, le retour depréférences et divers. La figure suivante illustre ces différentes dimensions :

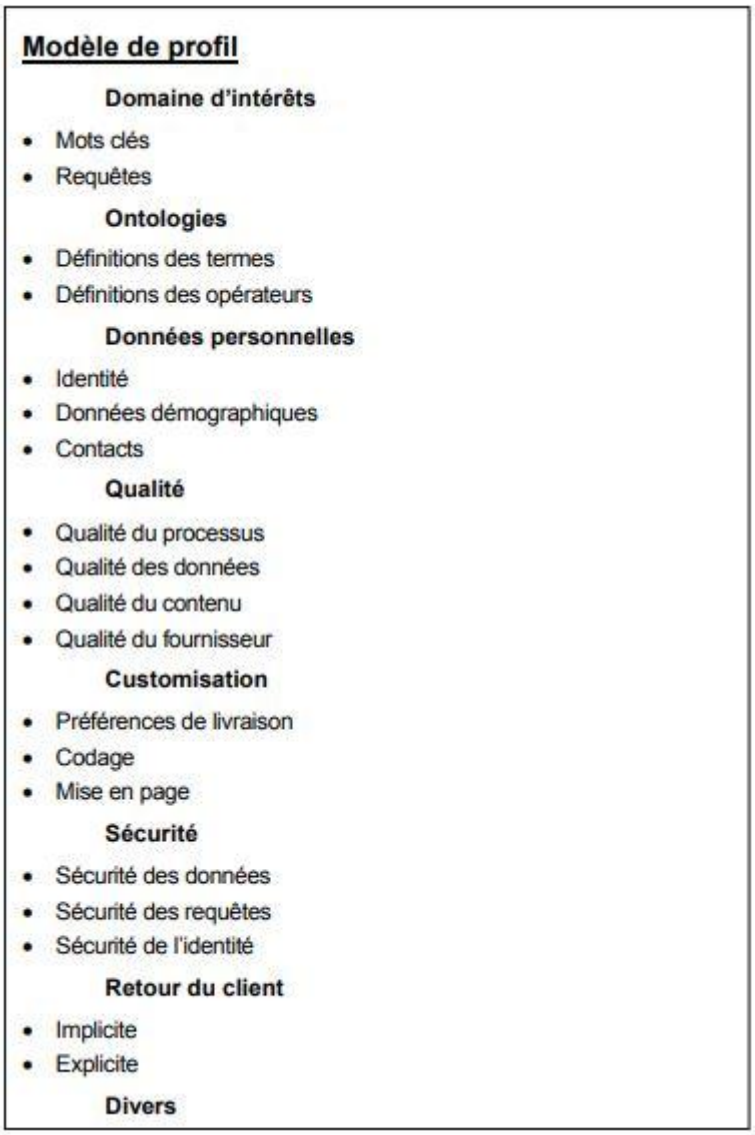


Figure 2.5: Structure d'un profil **[17].**

- **Données personnelles** : Ces données sont généralement stables, décrivent principalement les donnéesde l'identité de la personne, son domaine d'activité.
- **Domaine d'intérets :** C'est la partie dynamique d'un profil car ils sont divers et variés d'un profil et un autre et peut être décrit par des requêtes et mots clés.
- **Ontologie du domaine** : Elle explicite la sémantique de certains termes employéspar l'utilisateur dans son profil et donne en conséquence une meilleure interprétationet signification de ses centres d'intérêts selon le domaine et le contexte dans lequel iltravaille. Cette ontologie peut être spécifique à l'utilisateur et explicitement définie parlui ou générique relative à un domaine particulier et dont la terminologie est clairementdéfinie dans un thésaurus par exemple.

- **Qualité attendue des résultats délivrés:** Elle exprime les préférences de l'utilisateur, tels que l'origine de l'information, sa précision, sa fraîcheur, sa durée de validité, le temps nécessaire pour la produire ou la crédibilité de sa source qu'il l'a produit.
  - **Personnalisation (customisation) :** Elle concerne l'adaptation et lapersonnalisation de l'interface selon les préférences et les commodités de l'utilisateurtel que les modalités de présentation des résultats et les choix esthétiques ou visuelsde l'utilisateur, la quantité de résultats qu'il souhaite recevoir, etc.
  - **Sécurité et la confidentialité :** Pour la définition des droits d'accès au système etmême pour exprimer la volonté de l'utilisateur de cacher un traitement qu'il effectuepar la définition du degré de visibilité de certaines opérations.
  - **Retour de préférences :** Il désigne le « feedback » ou le retour de pertinence de l'utilisateur. Ce retour de pertinence peut être explicite et clairement fourni par lui ou implicite fourni par l'analyse de certaines informations récupérées ou dérivées à son insu.
  - **Informations diverses:** Il peut être parfois souhaitable de fournir certaines informations spécifiques selon l'exigence de l'application ou du contexte de travail.

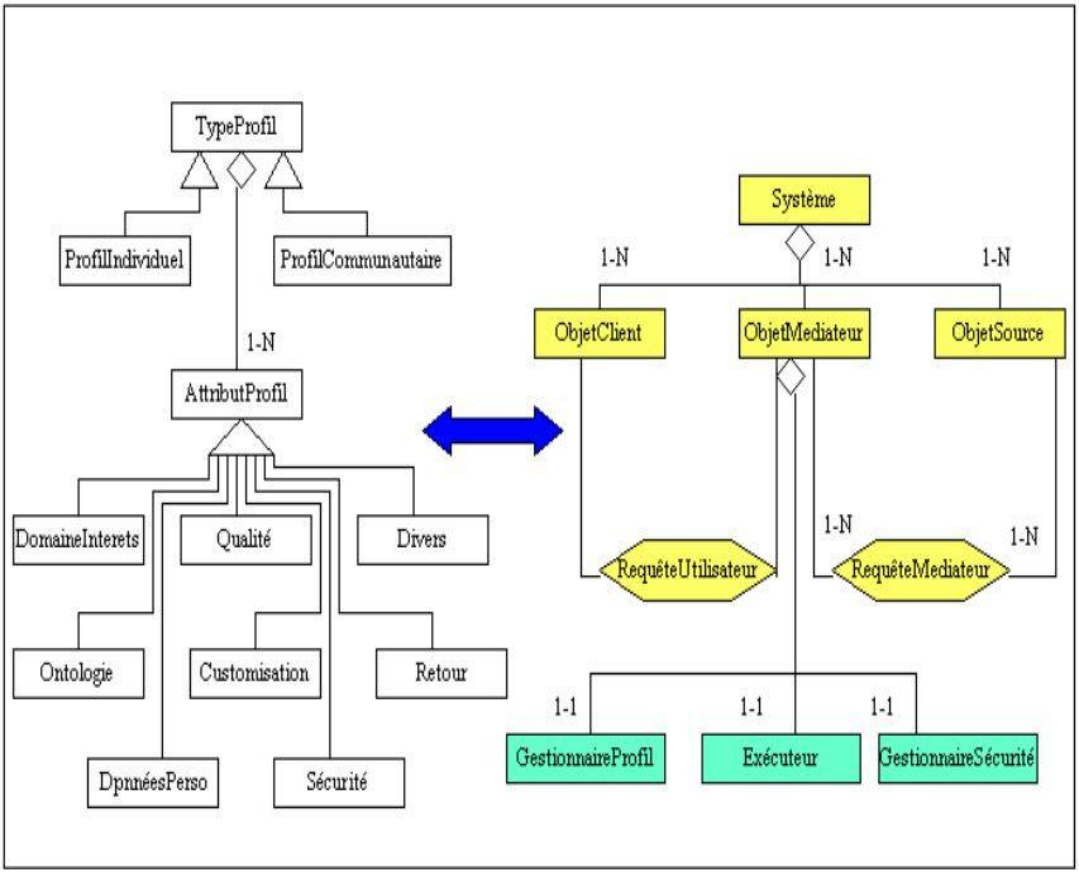


Figure 2.6: Modèle de départ d'un système de personnalisation **[17]**

6.4. Travaux de Boulkrinat

Boulkrinat**[49],** a adapté ces dernières dimensions pour le filtrage d'informations. Elle a proposé les dimensions suivantes : informations personnelles, contacts, préférences, qualité, centres d'intérêts, annotation, données statiques et divers. Pour la dimension centre d'intérêt, elle a associé une ontologie figure 2.4.

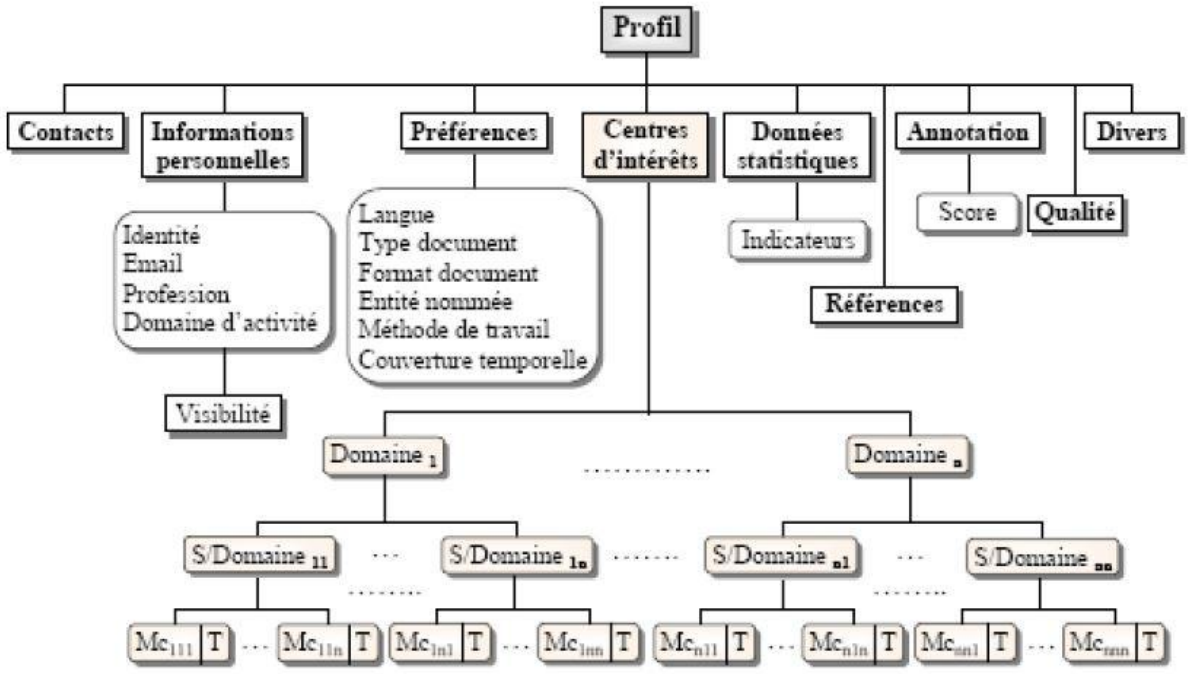


Figure 2.7: Description multidimensionnelle et hiérarchique du profil **[49].**

6.5. Travaux de Bouaissa

D.Bouaissa**[19]** a proposé un modèle du décideur figure 2.5 construit autour de quatre dimensions. Elles sont liées à l'objectif décisionnel, aux caractéristiques du décideur, aux facteurs contextuels et à la requête.

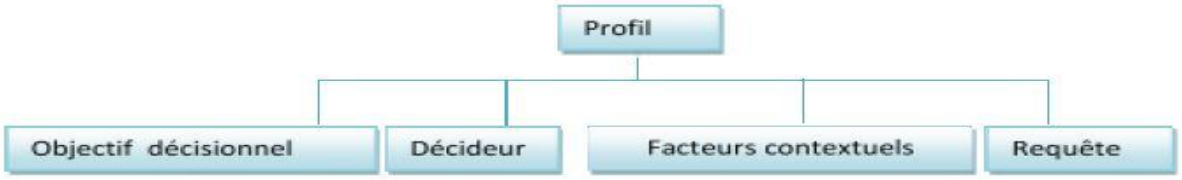


Figure 2.8: Dimensions du modèle de profil du décideur **[19]**

- **La modélisation de l'objectif décisionnel :** Cet élément va contenir le ou les objectifs du ou des projets décisionnels de l'entreprise auxquels le problème se rapporte.
- **Modélisation des caractéristiques du décideur :** L'objectif consiste à identifier les paramètres relatifs au décideur qui peuvent expliquer pourquoi il s'intéresse à tel ou tel événement.
- **Modélisation des facteurs contextuels :** Ces facteurs situent le décideur dans son environnement. C'est un utilisateur qui est confronté à une situation décisionnelle et qui se trouve dans un environnement informatique. Cette section porte sur la description des facteurs contextuels relatifs à la situation décisionnelle.
- **Modélisation de la requête :** Dans cette modélisation, elle a associé ce que privilégie le décideur dans son besoin informationnel et la façon d'interagir avec le système.

7. Conclusion

La personnalisation aide à faciliter l'expression du besoin de l'utilisateur et de lui permettre d'obtenir des informations pertinentes. Il faut considérer toutes les facettes du profil à savoir, l'utilisateur du système lui-même, le contexte dans lequel il émerge et les préférences exprimées par l'utilisateur. Les préférences exprimées par l'utilisateur traduisent ses propres besoins en informations que cela porte sur le contenu à travers l'expression du centre d'intérêt, le niveau de qualité qu'il désire ou des modalités de présentation de l'information résultat de la requête.

Ce chapitre a permis de cerner la thématique de la personnalisation de l'information dans sa globalité. Nous avons présenté en premier les éléments de base autour de la notion de la personnalisation ainsi que les profils utilisateurs. Ensuite nous avons défini les techniques de construction d'un profil utilisateur. Puis nous avons cité les domaines d'utilisation de la personnalisation. Enfin il était important de fouiller les travaux sur la modélisation des profils.

Algorithmes d'extraction des motifs fréquents

1. Introduction

L'extraction de connaissances dans les bases de données, également appelé data mining, désigne le processus permettant d'extraire des informations et des connaissances utiles qui sont enfouies dans les bases de données, les entrepôts de données (data warehouse) ou autres sources de donnée.

Depuis sa création, le Data Analytique joue un rôle important dans le processus de prise de décision, du coup plusieurs algorithmes FPM ont été développés pour accélérer les performances d'extraction.

Durant ce Chapitre, en intéresse à deux principaux problèmes liés à la prise de décision ou un décideur est face à une surcharge informationnelle, ou la plupart des algorithmes FPM ont ses problèmes des motifs cachés des ensembles d'éléments fréquents qui prennent plus de temps à exploiter lorsque la quantité de données augmente avec le temps ainsi que l'espace mémoire consommé lors de cette exécution.

Le but de cette étude est de relever les forces et les faiblesses des algorithmes FPM, afin de développer un algorithme efficace pour minimiser au maximum les problèmes et accéléré la performance d'extraction.

2. Définitions

Item:

Est tout objet, article, attribut, littéral, appartenant à un ensemble fini d'éléments distincts I = {x1, x2, ..., x m}. Dans les applications de type analyse du panier de la ménagère, les articles en vente dans un magasin sont des items. **[52]**

Itemset:

Est un ensemble de n Items. L'ensemble de tous les Itemsets possiblement formés par les éléments d'Items est 2^n. **[53]**

Itemset fréquent :

Un Itemset est fréquent si et seulement si son support est supérieur à un support minimum défini par l'utilisateur. [54]

Support minimal :

Notée minsup est le nombre minimum d'occurrence d'un Itemset pour être considéré comme fréquent. [55]

Transaction :

Est un ensemble d'items par exemple les items achetés par un client C à une date précise. Dans une base de données une transaction est représentée par trois attributs : idClient (identifiant d'un client), idDate (un identifiant pour une date), itemset (un ensemble d'items non vide). [55]

Basse de donnée transactionnelle :une base de données transactionnel le peut être représentée sous forme horizontale, verticale ou binaire.[56] le tableau 3.1 présente une matrice de transaction

Tableau 3.7: Matrice de transactions.

Transaction	Item 1	Item 2	Item 3
T1	0	1	1
T2	1	0	1

3. Classification des algorithmes d'extraction desmotifs fréquents

D'après[57] dans ses livre, Les algorithmes d'extractions de motifs fréquents peuvent être généralement classés en algorithmes générer-et-tester (candidate generate-and-test) et en croissance de modèles (pattern growth)comme montrer dans la figure 3.1.

Les premiers algorithmes d'extraction de modèles fréquents appartiennent à la première catégorie, tandis que les plus récents se situent dans la seconde. Les algorithmes examinés dans ce chapitre sont les algorithmes fondamentaux.

3.1 Générer-et-tester (Candidate-generate-and-test)

L'algorithme Apriori **[58]** a été le premier exemple de l'approche générer-et-tester en tant que moyen de découvrir des modèles fréquents à partir de données précises. 'algorithme Apriori extrait des ensembles d'éléments fréquents de longueur croissante à partir de données précises en effectuant plusieurs analyses d'une base de données d'entrée.

3.2 Croissance demodèles (patterngrowth)

L'une des principales lacunes des algorithmes générer-tester réside dans la nécessité d'analyser plusieurs fois la base de données en entrée complète chaque fois que des ensembles d'éléments de niveau supérieur candidats sont générés.

Une approche de croissance de modèle, supprime cette limitation en réduisant le nombre d'analyses de base de données à deux ou trois. Han et al**. [59]** ont introduit le premier algorithme basé sur des arbres permettant de lutter contre ce problème pour obtenir des données précises, appelé FP-Growth.

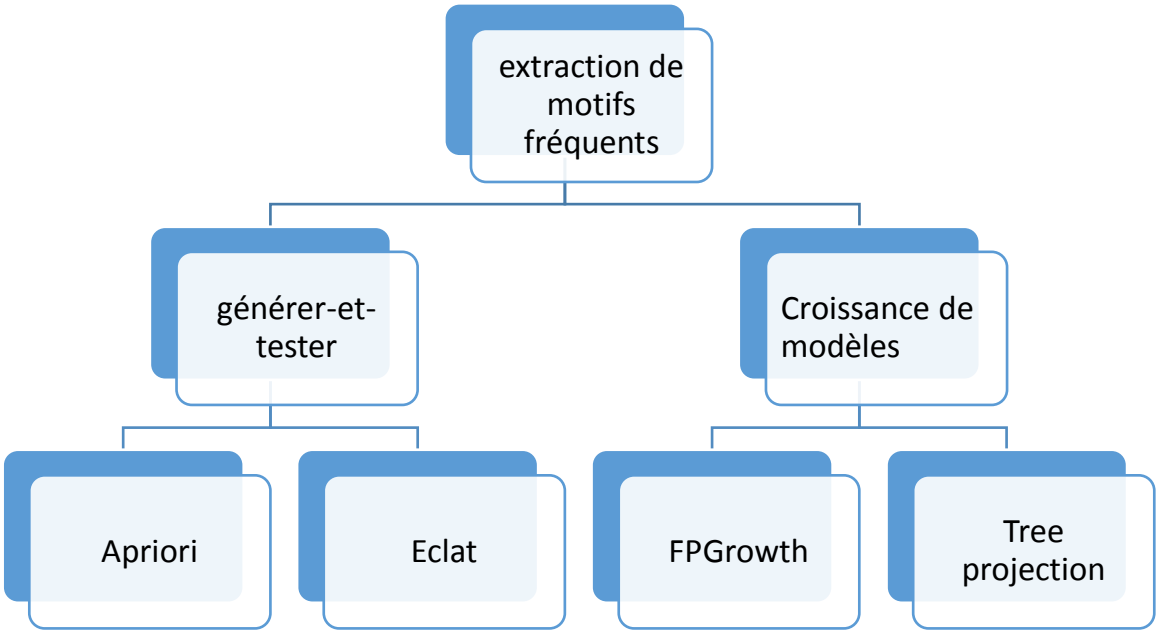


Figure 3.1: Classification des algorithmes d'extraction de motifs fréquents.

4 Algorithmes d'Extraction des motifs fréquent

4.1 Algorithme Apriori

Apriori est le tout premier algorithme pour l'extraction de motifs fréquents. Ila été donné par R agarwal et R srikant en 1994**[60].**d'apprêt Mishra,L et all en 2012**[61]** Apriori est un algorithme classique d'apprentissage des règles d'association. Il est conçu pour fonctionner sur des bases de données contenant des transactions, l'algorithme tente de trouver des sous-ensembles communs à au moins un nombre minimal K (seuil de confiance ou seuil de coupure) des ensembles d'éléments.

4.1.1 Procédure

- Le premier passage de l'algorithme consiste à compter les occurrences d'élément pour déterminer les grands ensembles d'items.

Ce processus est répété jusqu'à ce qu'aucun nouveau motif ne soit identifié. Items (k + 1) les ensembles d'éléments de longueur candidats sont générés à partir de grands ensembles d'éléments de longueur k.

- Les ensembles d'éléments candidats contenant des sous-ensembles de longueur k non importants sont élagués.

- La prise en charge de chaque ensemble d'éléments candidats est comptée en analysant la base de données.

- Éliminer les itemsets candidats qui sont petits (Seuil inferieur au Sup_min),la figure 3.2 montre tout cette procédure.

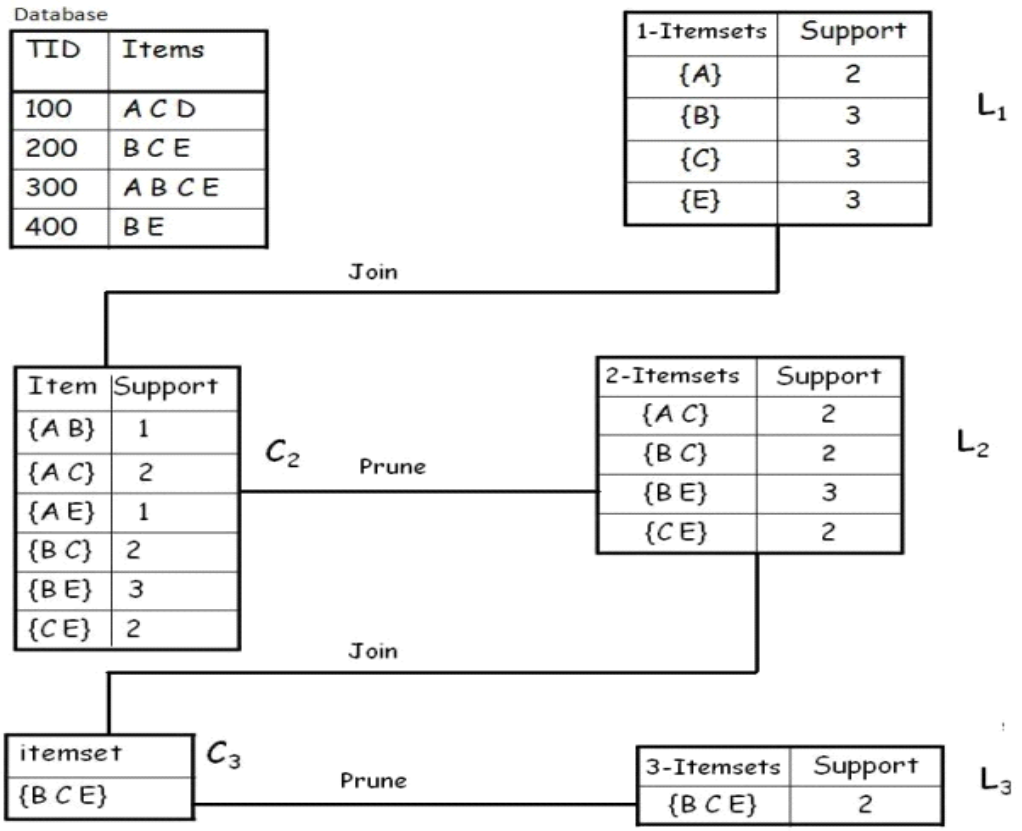


Figure 9.2: Algorithme Apriori, Génération d'ensembles d'éléments candidats et fréquents.[62]

4.1.2 Avantages et inconvénients

L'algorithme Apriori réduit considérablement la taille d'articles candidats de plus qu'il est facile à mettre en œuvre. Cependant, il souffre des limitations par rapport à la nécessité de nombreuses analyses de base de données ainsi que le grand nombre d'ensembles d'éléments candidats qui peuvent être encore généré si le nombre total d'ensembles des éléments fréquents augmente **[61].**

4.2 Algorithme FP-Growth

L'algorithme FP-Growth a été introduit par Han et al. **[59]** en 2000, ils ont dit qu'il est actuellement l'une des approches les plus rapides pour l'extraction fréquente d'ensembles d'articles.

C'est une méthode différente des approches par niveaux permettant d'extraire des itemsets fréquents sans génération de candidats, chose qui nous permet d'éviter les parcours et les visites répétés de la base de données.

4.2.1 Procédures

- On effectue un premier parcours de la base T pour déterminer les items fréquents en fonction du support minimum fourni. Ces items seront triés par la suite par ordre décroissant de support dans une liste (L). Les items ainsi triés seront traités dans cet ordre.

- Un second parcours de T est alors effectué. Chaque transaction est alors triée selon l'ordre des items dans L. Le nœud racine de l'arbre {null} est d'abord crée. Durant ce même parcours, une branche sera créé pour chaque transaction, mais des transactions ayant un même préfixe partageront le même début d'une branche de l'arbre, ainsi deux transactions identiques seront représentées par une seule et même branche.

La raison pour laquelle les items sont traités du plus fréquent au moins fréquent est que les items fréquents seront proches de la racine et seront mieux partagés par les transactions. Ceci fait du FP-tree une bonne structure compacte pour représenter les bases transactionnelles. **[59]**

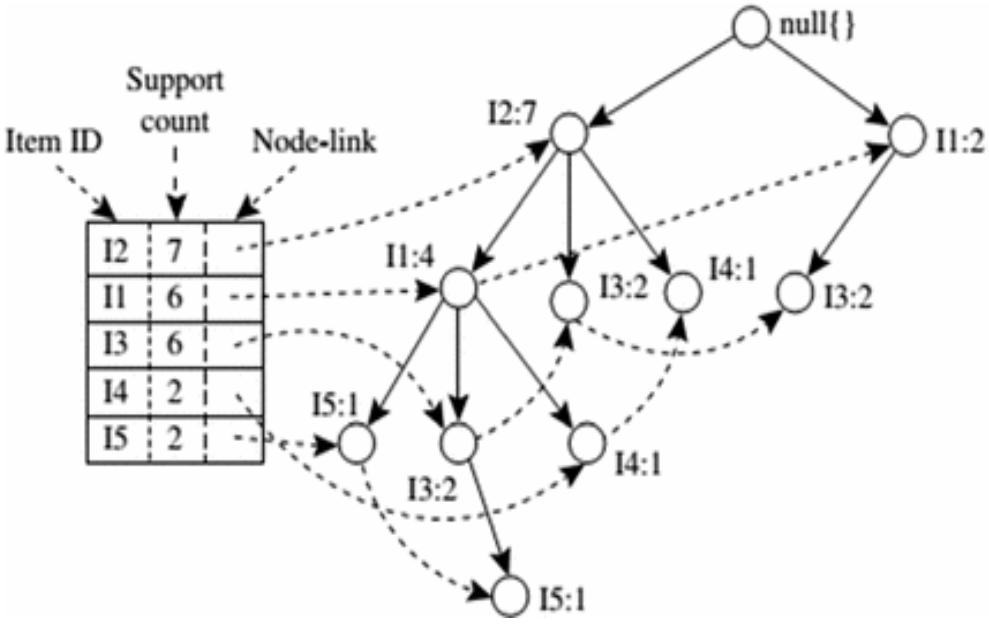


Figure 3.3: Exemple d'un FP-Tree construit **[63]**

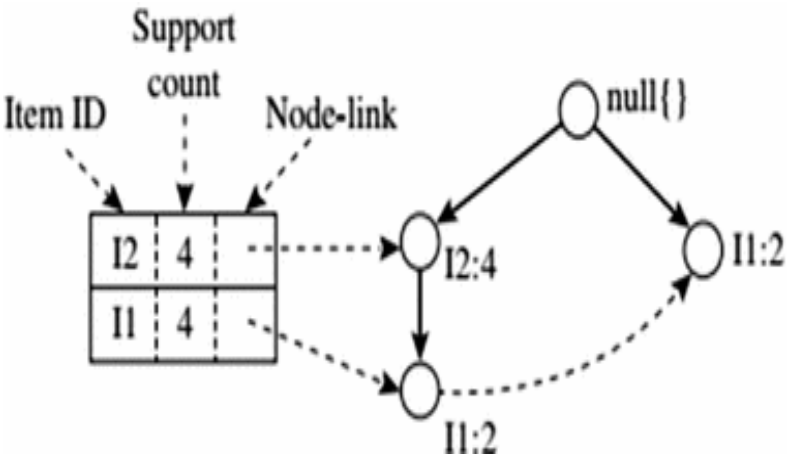


Figure 3.4: Exemple d'un FP-tree associé au nœud I3 **[63].**

4.2.2 Avantages et inconvénients

L'algorithme FP-Growth résout le problème de la nécessité de nombreuses analyses de base de données, vu qu'il ne fait que deux balayages de la base des transactions.

Néanmoins, cela ne garantit pas, dans le cas où la base de transactions est trop volumineuse, que toute la structure du FP-tree tiendra en mémoire centrale. De plus la construction de la structure FP-tree peut s'avérer longue et pourrait consommer beaucoup de ressources système. **[59]**

4.3 Algorithme ECLAT

ECLAT (Equivalence Class Transformation) a été introduit par Zaki, Parthasarathy, Ogihara et Li en 1997 **[64],** d'apret Javeed,MZ et all(1997) Eclat a été conçu pour surmonter les inconvénients de algorithme Apriori. Il utilise la mémoire agrégée du système en partitionnant les candidats en ensembles disjoints à l'aide du partitionnement par classe d'équivalence. Il dissocie la dépendance entre les transformateurs en droit en commençant de sorte que le coût de redistribution puisse être amorti par les itérations ultérieures. Eclat utilise la structure de base de données verticale qui regroupe toutes les informations pertinentes dans la liste des objets.

Il utilise algorithme de recherche en profondeur d'abord (Depth-First Search) et la base de données n'a pas besoin d'être scannée plusieurs fois pour que identifier les éléments (k + 1). La base de données est analysée une seule fois pour transformer les données du format horizontal dans le format vertical.

4.3.1 Procédure

Eclat est composé de trois phases principales [65](comme montrer dans les figures 3.5, 3.6 et3.7) :

- La phase d'initialisation : construction globale des 2-itemsets.
- La phase de transformation : partitionnement de l'ensemble des 2-itemsets fréquents et distribution de ces partitions aux autres processeurs. Transformation verticale de la base.
- La phase asynchrone : construction des k-itemsets fréquents.

itemset	TID_set
I1	{T100, T400, T500, T700, T800, T900}
I2	{T100, T200, T300, T400, T600, T800, T900}
I3	{T300, T500, T600, T700, T800, T900}
I4	{T200, T400}
I5	{T100, T800}

Figure 3.5: Le format de données vertical. [66]

itemset	TID_set
{I1, I2}	{T100, T400, T800, T900}
{I1, I3}	{T500, T700, T800, T900}
{I1, I4}	{T400}
{I1, I5}	{T100, T800}
{I2, I3}	{T300, T600, T800, T900}
{I2, I4}	{T200, T400}
{I2, I5}	{T100, T800}
{I3, I5}	{T800}

Figure 3.6:2-Itemsets au format de données vertical. [66]

itemset	TID_set
{I1, I2, I3}	{T800, T900}
{I1, I2, I5}	{T100, T800}

Figure 3.7: 3 itemsets au format de données vertical. **[66]**

4.3.2 Avantages et inconvénients

Analyser la base de données pour trouver le nombre de supports de (k + 1) éléments n'est pas requis, mais il a besoin plus d'espace mémoire et de temps de traitement sont nécessaires pour l'intersection de longs ensembles de TID **[63].**

4.4 Algorithme TreeProjection

TreeProjection algorithme a été introduit par Agarwal, Aggarwal, Prasad en 2001[1] est un algorithme qui extrait les itemsets fréquents à travers des techniques de recherche pour la construction d'un arbre lexicographique et les itemset sont projetés sur arbre lexicographique ou les noude sont les k-itemsets.**[63].**

TreeProjection peut être considéré comme un cadre générique qui préconise la notion de projection de base de données, contexte de plusieurs stratégies différentes pour la construction de l'arbre d'énumération, telles que en largeur d'abord, en profondeur d'abord, ou une combinaison des deux **[67]** La recherche est effectuée en parcourant l'arbre lexicographique avec une approche descendante.

L'arbre lexicographique figure3.8 est défini en la manière suivante **[68]** :

- Un sommet existe dans l'arbre correspondant à chaque motif fréquent. La racine de l'arbre est nul.
- On a $I=\{i_1,\ldots,i_k\}$ sont les motif fréquents, ou $i_1,i_2,\ldots i_k$ sont listés par ordre lexicographique, Le parent du nœud I est le itemsets $\{i_1,i_2,\ldots,i_{k-1}\}$

4.4.1 Avantages et inconvénients

Identifie rapidement les ensembles d'éléments fréquents, car seul le sous-ensemble de transactions pouvant probablement contenir les ensembles d'éléments fréquents est

recherché par l'algorithme. Par contre différentes représentations de l'arbre lexicographique présentent différentes limitations en termes d'efficacité pour la consommation de mémoire. **[68]**

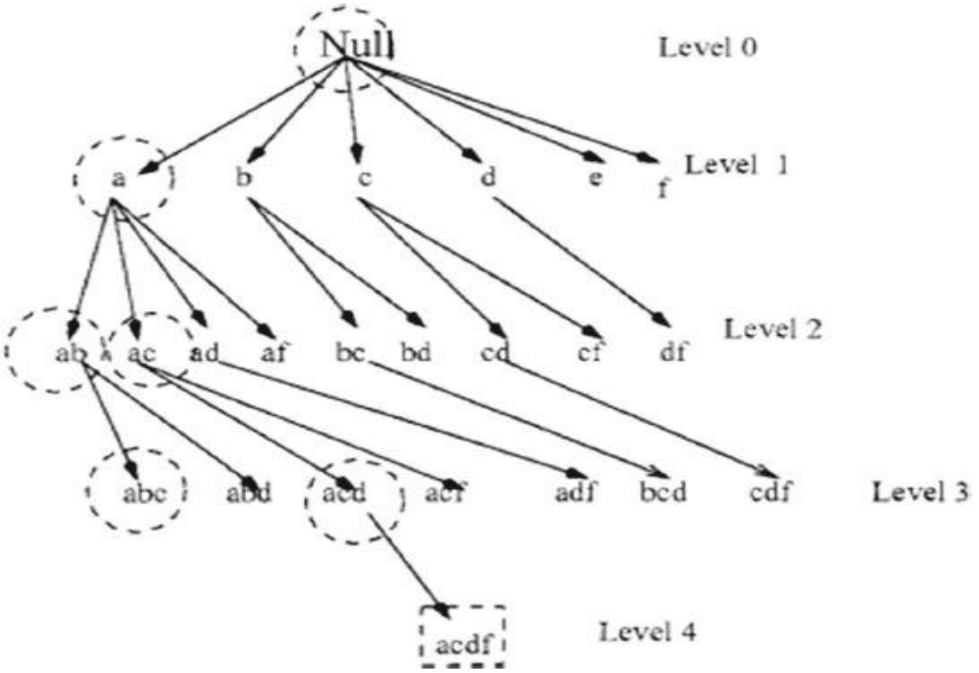


Figure 3.8: Arbre lexicographique. **[68]**

5 Comparaison des algorithmes d'extraction de modèlesfréquents

Heaton (2016)**[69],** a réalisé une étude de performance pour comparer les trois algorithmes (Apriori, Eclat, FPgrowth), en utilisant l'effet de la densité de données et l'augmentation de la taille de transaction.

5.1 Effets de la densité des données

L'algorithme Apriori, Eclat et FP-Growth fonctionne de manière similaire, jusqu'à ce que la densité dépasse 70%, Apriori a des besoins en mémoire considérablement plus importants que les autres algorithmes. À 70%,Eclat et FP-Growth affichent tous deux une croissance très similaire mais Apriori avait alloué la totalité de RAM de la machine de test. Cela a rendu nécessaire l'échange de stockage physique et a eu un impact considérable sur le temps d'exécution de l'algorithme. Il est également intéressant de noter qu'Eclat est légèrement en avance sur FP-Growth à faibles densités. Comme le montre la figure 3.9.

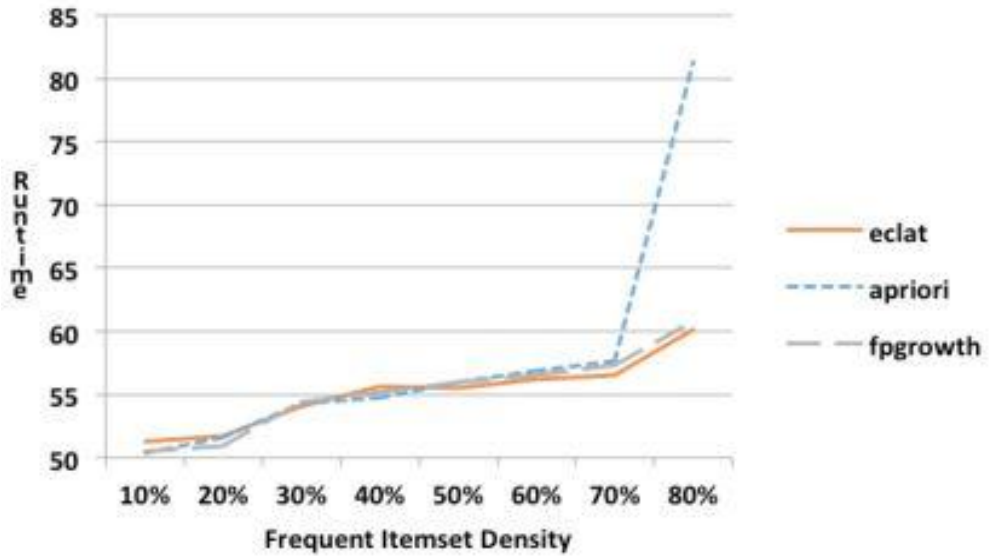


Figure 3.9: Effet de la densité d'ensemble d'articles sur le temps d'exécution (secondes).[69]

5.2 Effets l'augmentation de la taille des transactions

Les trois algorithmes montrent presque exactement les mêmes performances pour des tailles allant jusqu'à 60. Une fois supérieur à 60, Apriori semble croître beaucoup plus vite que les deux autres. Ceci est probablement dû à la mémoire accrue utilisée par Apriori. Fait intéressant, Apriori a réalisé le meilleur résultat entre 60 et 70 tailles de transaction maximales. Comme le montre la figure 3.10.

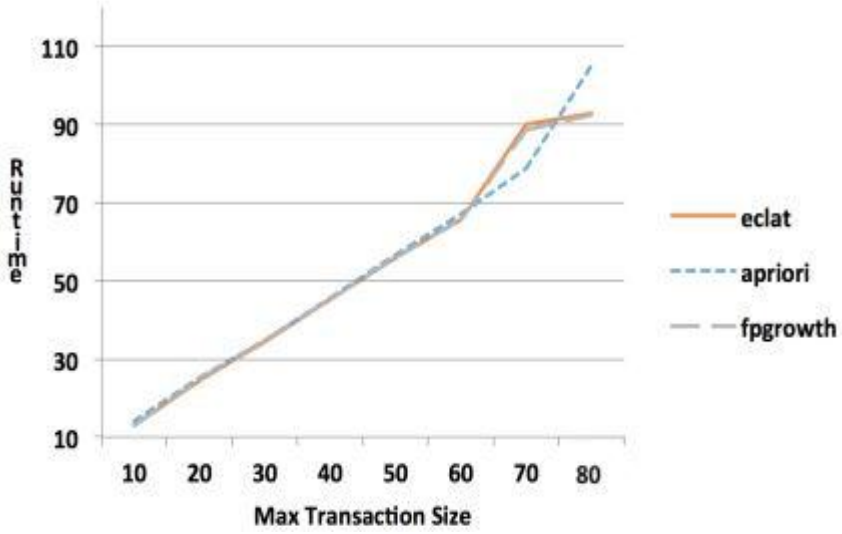


Figure 3.10: Effet de taille de transaction maximale sur le temps d'exécution (secondes).[69]

Les deux tableaux suivants mantre une compaaisant entre Apriorie et Eclat, Apriorie et FP-Growth.

Tableau 3.8: Différence entre l'algorithme Apriori et Eclat **[70].**

Paramètres	Algorithme Apriori	Algorithme Eclat
Technique	Il utilise une approche de recherche d'abord large et utilise la propriété apriori (tous les sous-ensembles non vides d'un ensemble d'éléments fréquent doivent être fréquents) et la méthode joy-prime	Il utilise une approche de recherche en profondeur d'abord et utilise l'intersection de la liste d'identifiants de transaction pour générer des ensembles d'éléments candidats
Utilisation de la mémoire	En raison de la grande quantité de candidats sont produites alors nécessitent un grand espace mémoire	Nécessite moins de mémoire que apriori si les itemsets sont peu nombreux
Bases de données	Convient aux jeux de données épars ainsi qu'aux jeux de données denses	Convient pour les jeux de données moyens et denses, mais pas pour les petits jeux de données
Temps	le temps d'exécution est plus que le temps perdu à produire des candidats à chaque fois.	Temps d'exécution est petit que l'algorithme apriori.

Tableau 3.9: Différence entre l'algorithme Apriori et FPGrowth **[71].**

Paramètres	Apriori	FPGrowth
Technique	Utiliser la propriété Apriori et Joindre et élaguer la propriété	Il construit une base d'arborescence FP et un modèle conditionnel à partir de la base de données
Utilisation de la mémoire	En raison du grand nombre de candidats générés, son utilisation nécessite un grand espace mémoire.	En raison de sa structure compacte et de l'absence de génération de candidats, il nécessite moins de mémoire
Nombre de scans	Analyses multiples pour générer des ensembles de candidats	Analyser la base de données seulement deux fois
Temps	La génération de jeux d'éléments candidats prend un temps exponentiel	Le temps d'exécution est plus petit que l'algorithme Apriori
La complexité du temps	Complexité temporelle exponentielle de Apriori est O (2^n)	La création d'arborescence est basée sur le nombre d'éléments dans la base de données. La complexité de l'arbre de création est O (| DB |)
Complexité de l'espace de recherche	N nombre de candidats sont générés, il faut donc beaucoup de mémoire	La complexité de la recherche par tous les

		chemins est alors délimitée par A (nombre d'en-têtes ^ 2 * profondeur de l'arbre)

6 CONCLUSION

Dans ce chapitre, nous donnons un bref aperçu des principaux algorithmes d'extraction des motifs fréquents tel qu'Apriori, FP-Grouth, Eclat et TreeProjection, L'objectif de cette étude est de passer en revue les forces et les faiblesses des algorithmes fondamentaux dans FPM (Frequent Pattern Mining).

Les grands problèmes dans la FPM, Premièrement, les motifs cachés qui existent fréquemment dans la base de données prend un grand temps quand la réalisation de l'extraction lorsque la quantité de données augmente. Cela provoque une consommation de mémoire importante suite à un calcul important effectué par l'algorithme extraction.

Apriori est algorithme facile à comprendre. Pour cette raison, Apriori est un point de départ populaire pour l'étude des motifs fréquents. Cependant, Apriori a de graves problèmes d'évolutivité et épuise la mémoire disponible beaucoup plus rapidement qu'Eclat et FP-Growth, et FP-Growth montrer des performances légèrement meilleures qu'Eclat.

Les chercheurs peuvent améliorer son efficacité en apportant d'autres nouvelles techniques avec les méthodes existantes. Ces algorithmes peuvent être modifiés efficacement pour réduire le temps d'exécution et utilisation de la mémoire.

Modélisation et conception

1. Introduction

Une des premières fonctionnalités que doit offrir un système de personnalisation est la modélisation de l'utilisateur au moyen d'un profil. Cette modélisation dépend de l'utilisateur lui-même, de l'application qu'il met en œuvre et de l'environnement de mise en œuvre. A ce niveau on définit les dimensions et leurs attributs qui peuvent caractérisés le mieux les besoins de l'utilisateur dans son contexte.

2. Approche proposée

Le rôle du profil du décideur ainsi que les caractéristiques psychologiques et son comportement ont une influence sur la prise de décision. Notre objectif consiste à identifier les paramètres relatifs au décideur qui peuvent expliquer pourquoi il s'intéresse à tel ou tel événement. Pour ce faire, nous avons identifié un certain nombre de paramètres, susceptible d'influencer sur la définition du problème et par la suite sur le processus de prise de décision. Dans notre approche, nous retenons les facettes suivantes :

- Modélisation du décideur ;
- Modélisation de l'objectif décisionnel ;
- Modélisation de l'organisation ;
- Modélisation de l'environnement.

2.1Modélisation du décideur

Pour être efficace, notre modèle doit tenir compte des particularités de chaque décideur en essayons de disposer des informations sur chaque décideur et sur son mode de raisonnement.

➢ Les données personnelles

Les données personnelles sont la partie statique du profil. Elles contiennent des informations qui décrivent le décideur et ne dépendent pas du système à interroger. Nous distinguons les caractéristiques individuelles suivantes:

- Son identité, composée d'un ensemble d'attributs d'identification du décideur. Des exemples de tels attributs sont le nom et le prénom du décideur, son adresse, son numéro de téléphone ou de fax, son adresse email etc.

- Ses formations initiales (universitaire ou autre) et poursuivies durant son parcours professionnel,
- Son expérience, exprimée au nombre d'années d'ancienneté. L'expérience peut influencer sur le style cognitif du décideur et donc sur sa façon de percevoir et de résoudre le problème ;
- Forme de représentation de ses résultats, ces résultats peuvent être sous forme de détails techniques, de données statistiques, de graphiques, de rapports etc.
- Ses préférences, une préférence est une expression permettant de hiérarchiser l'importance des informations dans un profil ou un contexte.

➢ Style cognitif

Le style cognitif peut être défini comme la façon propre à chacun de percevoir et de comprendre l'information perçue face à une nouvelle connaissance. D'après Bouaka **[51]**, le style cognitif a une influence sur la façon dont les individus examinent leur environnementpour recueillir de l'information et sur la façon dont ils intègrent leurs interprétations dans les modèles mentaux qui guident leurs actions (déduction, abduction, induction).

➢ Traits de personnalité

Selon MBTI (Myers-Briggs Type Indicator ou indicateur typologique de Myers-Briggs) **[72],** un test psychologique fondé sur les travaux du psychiatre C.G Jung (1875-1961) qui permet de classer la personnalité selon 16 grands types, avec les points forts et les points faibles, et de comprendre les sources de motivations. Son but est d'analyser le fonctionnement psychique selon 4 critères :

- **Orientation de l'énergie** : Elle se rapporte à la manière dont le décideur tire son énergie. Nous distinguons deux pôles :
  - Pole Extraversion (E): L'énergie est orientée et puisée dans le monde extérieur, elle est projetée, dirigée vers les gens et les choses extérieures.
  - Pole Introversion (I) : A l' opposé des Extravertis, l'énergie est orientée et puisée dans le monde intérieur (d'eux-mêmes).
- **Recueil de l'information** : Cette fonction est en rapport avec la manière dont le décideur perçoit le monde, ou plus précisément, les éléments sur lesquels il va porter son attention. Nous distinguons deux pôles :

- Pole Sensation (S) : Le décideur recueille l'information avec ses 5 sens, la sensation lui permet de voir, entendre, goûter, toucher et sentir le monde autour de lui.
- Pole iNtuition (I) : Le décideur recueille l'information avec le 6ème sens psychique (ce qui est imaginé/pensé). L'intuition lui permet de percevoir des renseignements abstraits, comme les symboles et les concepts.

- **Traitement de l'information** : Cette fonction est en rapport avec la manière dont le décideur prend ses décisions. Cela peut se faire de manière rationnelle (réflexion) ou émotionnelle (sentiment). Nous distinguons deux pôles :
  - Pole Thinking (T) : Le décideur prend ses décisions selon des critères objectifs (logique, raison).
  - Pole Feeling (F) : Le décideur prend ses décisions selon des critères subjectifs (émotions, valeurs).
- **Mode d'action :** Cette fonction se rapporte à la manière dont le décideur aime agir pour atteindre ses objectifs. Nous distinguons deux pôles
  - Pole Judgement (J) : Le décideur planifie ses actions à l'avance. Il permet de vivre dans un environnement structuré, ordonné et prévisible.
  - Pole Perception (P) : Le décideur adapte ses actions aux circonstances. Il permet l'expérimentation et l'ouverture aux changements. Il fait appel à des qualités de flexibilité, curiosité, réactivité et spontanéité.

2.2Modélisation de l'objectif décisionnel

Cet élément va contenir le ou les objectifs du ou des projets décisionnels de l'organisation auxquels le problème se rapporte. L'objectif décisionnel va correspondre à une triple finalité :

- il est une expression particulière des risques de l'organisation (positifs qui représentent les opportunités et négatifs qui représentent les dangers).
- il est une traduction partielle de l'enjeu du problème décisionnel ;
- et enfin, il est à l'origine la solution du décideur pour ce même problème.

La figure suivante illustre le modèle du décideur et son apport avec l'objectif décisionnel.

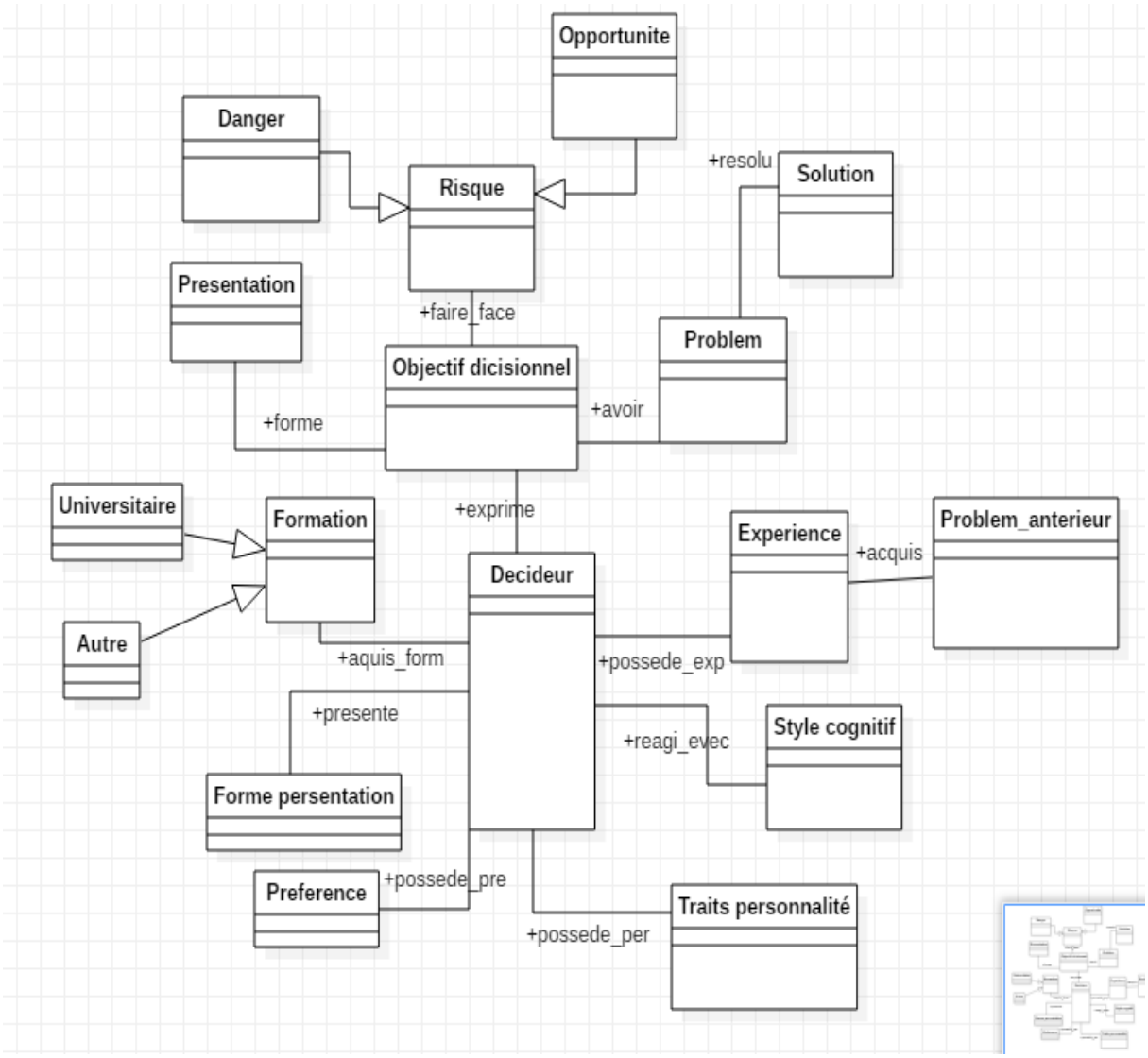


Figure 4.1: Modèle du décideur

2.3 Modélisation de l'organisation

Le modèle de l'organisation supporte l'analyse des facettes majeures de l'organisation afin de découvrir les problèmes et les possibilités de solutions par la spécification de sa stratégie, sa vision et ses missions. Il décrit aussi, l'organisation d'une façon structurée et modélise les informations sur l'organisation tels que sa structure et ses ressources.

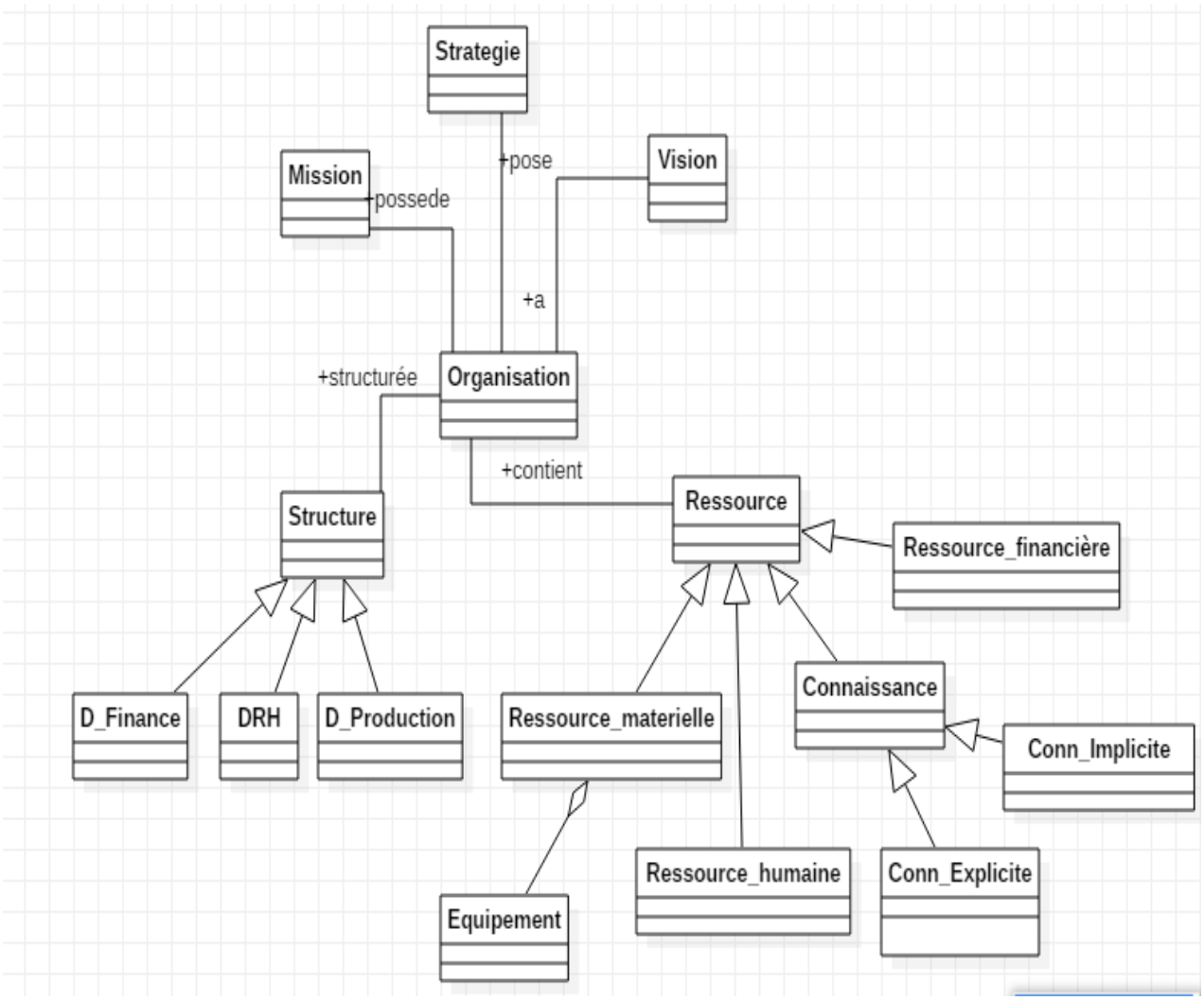


Figure 4.2: Modèle de l'organisation.

2.4 Modélisation de l'environnement

Toute organisation peut être vue comme un système à part entière. Ce système est soumis à des flux d'informations produites par le système lui-même et aussi à des flux d'informations reçus de l'extérieur. A ce jour l'environnement n'a pas encore connu de véritable modélisation,

nous cherchons donc à identifier dans l'environnement les facteurs que nous considérons sensibles, et qui peuvent intervenir dans la définition d'un problème décisionnel. Nous classons deux types :

- **L'environnement immédiat :** Dans notre modélisation nous avons choisissez le modèle des « **cinq forces de Porter** » qui a été élaboré en 1979 par le professeur deStratégie Michael Porter [73]. Selon Porter, cinq forces déterminent la structure concurrentielle d'une organisation qui influence la prise de décision:
  - **Les clients :** Si le pouvoir de négociation des clients est élevé, il influence la rentabilité du marché en imposant leurs exigences en matière de prix, de service, de qualité, etc.
  - **Les fournisseurs :** Le pouvoir de négociation des fournisseurs est très important sur la prise de décision. Des fournisseurs puissants peuvent imposer leurs conditions en termes de prix, de qualité et de quantité.
  - **Les produits de substitution :** Les produits de substitution peuvent être considérés comme une alternative par rapport à l'offre du marché. Les entreprises voient leurs produits être remplacés par des produits différents. Ces produits ont souvent un meilleur rapport prix/qualité et viennent d'un secteur où sont réalisés des profits élevés.
  - **Les entrants potentiels sur le marché :** Toute organisation a intérêt à créer autour d'elle des barrières d'entrée pour ne pas avoir une multitude de concurrents. Il s'agit de nouvelles entreprises ayant l'intention de se diversifier.
  - L**es concurrents :** Au sein d'un secteur, la concurrence entre organisations détermine l'attrait pour le secteur et un défi pour l'organisation.
- **L'environnement global :** Nous intégrons dans l'environnement global tous types d'environnement qui touche de près ou de loin à la compréhension du problème décisionnel, qui peut être regroupé en:
  - L'environnement social ;
  - L'environnement politique ;
  - L'environnement économique ;
  - L'environnement législatif ;
  - L'environnement scientifique.

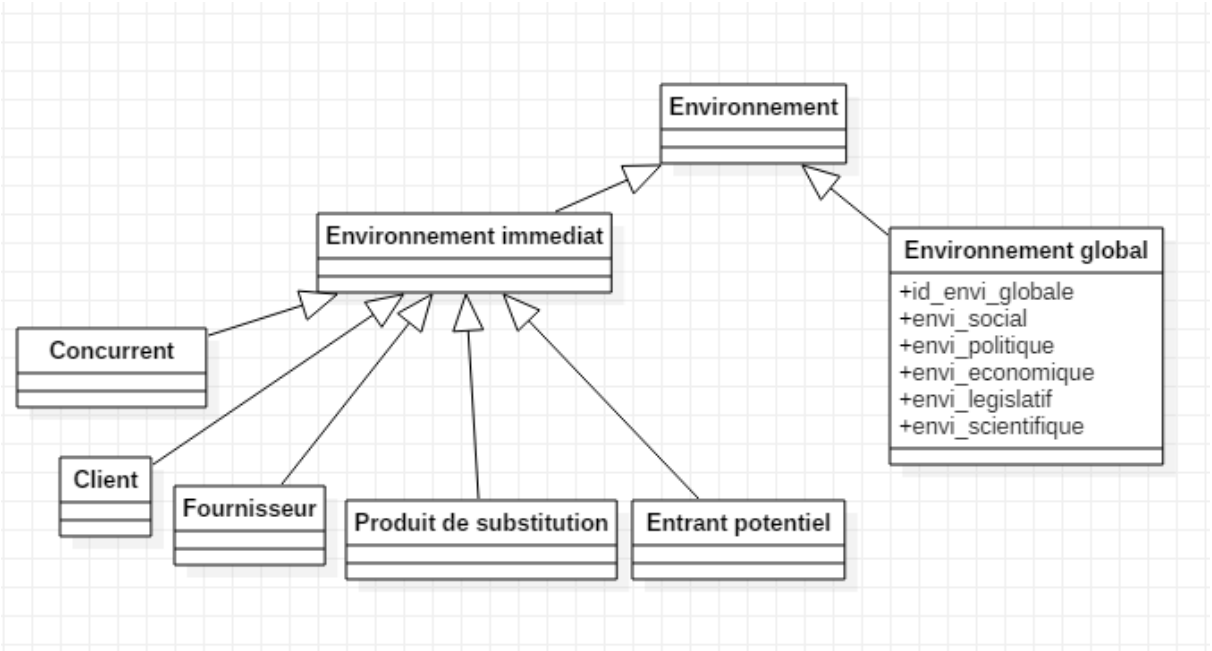


Figure 4.3 : Modèle de l'environnement

3. Cas d'étude : Extractions des motifs fréquents orienté besoin du décideur pour la gestion d'un hôpital

Dans notre cas d'étude, nous avons implémenté notre modèle dans le domaine médical, en utilisant la technique extraction des motifs fréquents, le processus est résumé dans le schéma de la figure 4.4.

- Premièrement le décideur saisît ses informations de profil pour créer un nouveau profil d'utilisateur. Ces données seront enregistrées dans la base de données.

- Après elles seront exploitées pour classer ce profil selon des clusters des décideurs créés par un clustering.
- D'autres données médicales seront appelées pour appliquer sur eux le processus d'extraction des motifs fréquents.
- Les résultats des deux opérations (le clustering et l'extraction des motifs fréquents) seront utilisés pour faire une classification. Cette classification appelle un modèle qu'on déjà crée avec un apprentissage supervisé des cas précédents.
- Les résultats de tout ce processus sont des motifs fréquents orientés besoin décideur.
- Toutes ces étapes seront détaillées par la suite.

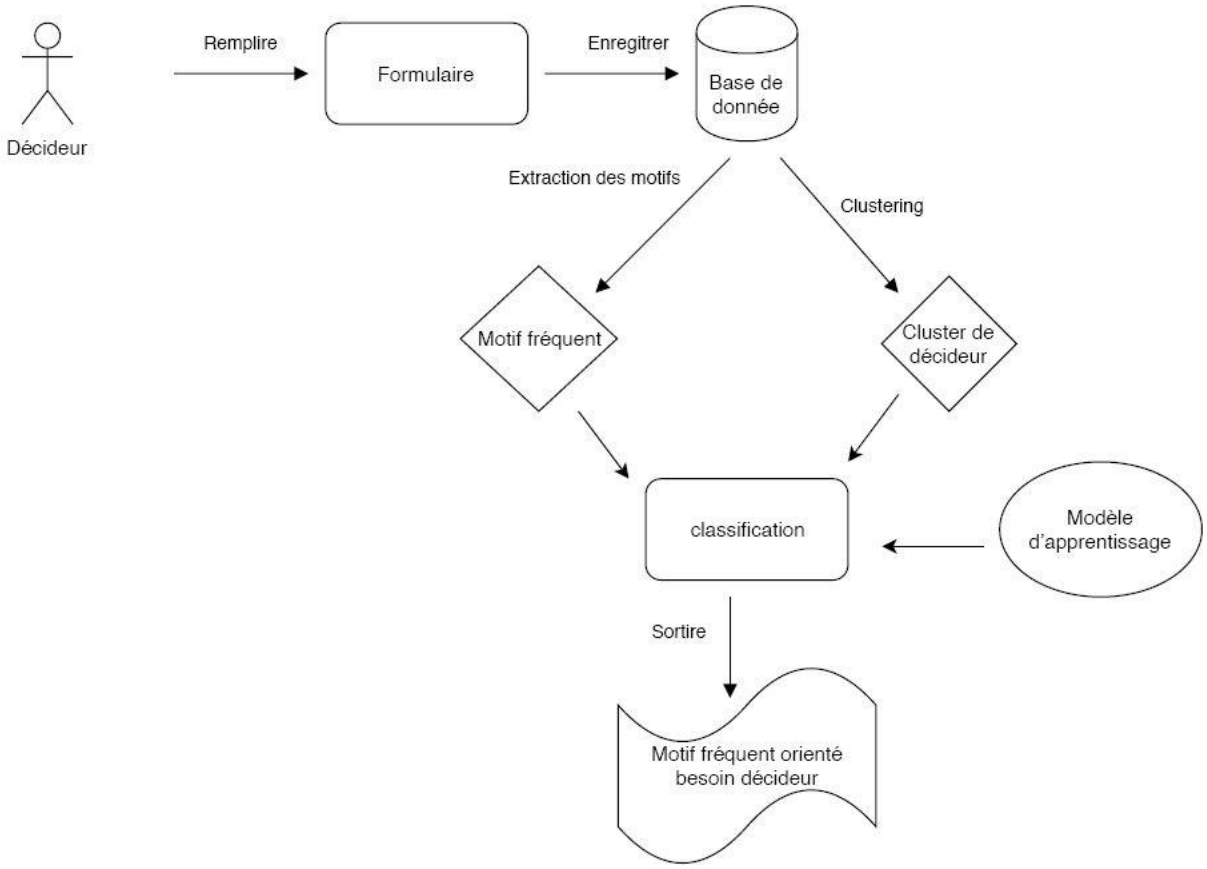


Figure 4.4: Schéma explicatif de processus d'extraction des motifs fréquents orienté besoin de décideur.

3.1 Utilité de l'extraction des motifs dans le domaine médical

Dans l'optique des méthodes d'exploration qui permettent d'aider le décideur dans sa prise de décision, il est souhaitable que fournir des connaissances claires, justifiées, expliquées et compréhensibles. Dans notre cas, il s'agit d'une extraction de motifs fréquents basé sur des données médicales.

Avant d'entamer notre travail d'extraction des motifs fréquents dans le domaine médical, nous citons quelques arguments qui servent comme l'atout de motivation pour l'utilisation de ses derniers :

- Les données médicales traitées concernent à la fois des attributs qualitatifs et quantitatifs, ce qui justifie les étapes de discrétisations pour obtenir des contextes booléens.
- Ces données sont volumineuses, ses caractéristiques posent de nombreux problèmes

- La fouille de ces données médicales poursuit un but d'exhaustivité des connaissances découvertes. À la différence des techniques statistiques, ce ne sont pas seulement les tendances globales des données qui sont recherchées mais également des propriétés locales qui concernent un petit nombre d'objets.

On prend en considération ces arguments et avec la numérisation des données des différents services hospitaliers, les praticiens ressentent le besoin de croiser les informations issues de ces différents services afin d'en déduire de nouvelles informations qui leurs permettront de diagnostiquer des cas qui présentent certaines complexités.

Le Data Mining est basée sur cette optique de croisement de l'information et d'extraction de nouvelles connaissances. Cela consiste à combiner plusieurs méthodes des différentes phases du processus de Data Mining.

3.2. Construction de modèle de personnalisation adapté au domaine médical

Notre problème décisionnel c'est la surcharge des patients dans un ou bien des services, ces services n'ont pas la capacité de prise en charge de tous les patients. Ainsi que les rendez-vous sont très loin. Cela oblige les clients (patients) à chercher d'autres concurrents (hôpitaux ou bien cliniques) pour prendre ses soins et l'organisation hospitalière va perdre ses clients.

Nous avons choisi dans notre travail les informations qui ont un impact sur l'objectif décisionnel et qui nous donnent des connaissances utiles. Ces connaissances peuvent influencer la décision :

3.2.1. Objectif décisionnel et Décideur

L'objectif décisionnel dans notre cas d'étude est de sélectionner :

- Les services qui souffrent de surcharge, donc la solution et d'agrandir ces services.
- Les services qui n'existent pas dans l'organisation hospitalière et qui ont beaucoup de demande (date de rendez-vous =loin), la solution est de construire un nouveau service.

✓ Pour répondre à notre objectif, nous devons mentionner les risques liés aux problèmes décisionnels, d'un côté positif (opportunité) ou bien négatif (danger)

- ✓ Le décideur est le directeur général de l'hôpital, c'est lui qui peut prendre la décision. Dans ce cas, il faut identifier :
- Son expérience : S'il a rencontré une situation pareille dans sa vie.
- Forme de présentation des résultats : comment le décideur veut voir ses résultats.
- Ses traits de personnalité : Quelle est la personnalité du décideur, en se basant sur le test MBTI.
- Ses préférences : Par exemple contrôler le monopole de la région.
- Ses formations : quelles sont les formations acquises par le décideur.

3.2.2. Organisation

La décision est prise pour faire une modification sur l'organisation (ajouter ou bien éclater un service) donc on est besoin de bien savoir :

- La structure : Les services qui existent.

Les ressources humaines, financières et matérielles pour notre objectif décisionnel.

3.2.3. Environnement

L'environnement a une influence directe ou indirecte sur la décision, donc il est utile de connaitre :

- Client avec toutes ces informations : dans notre cas d'étude le client est le patient.
- Fournisseur des produits médicaux ou bien matériels.
- Produits de substitutions : les services fournis par d'autres concurrents.
- Concurrents : les autres cliniques ou bien d'autres hôpitaux avec la distance de chaque concurrent.
- Entrant potentiel : c'est le risque d'avoir des nouveaux concurrents.

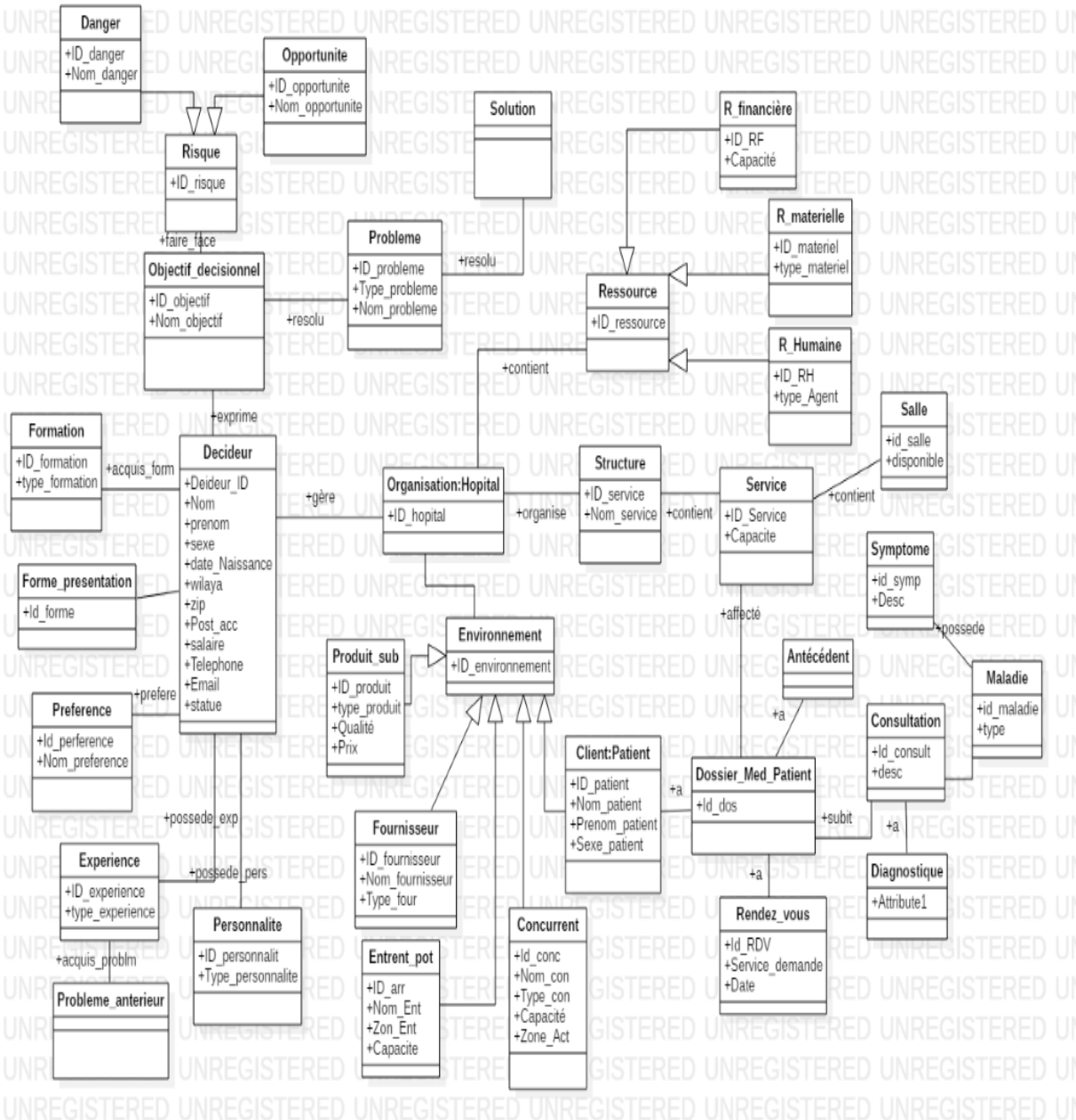


Figure 4.5: Modèle de personnalisation adapté au domaine médical.

3.3. Extraction des motifs

Nous avons utilisé la méthode d'extraction des motifs fréquents pour extraire les services qui rencontrent les problèmes cités, Cette extraction est faite par l'algorithme Apriori. Nous avons choisi comme ensemble des items {service, signe et date de rendez-vous} :

- Les services (item1): sont les services qui existent dans l'hôpital.
- Signe (item2) : sont les signes de la maladie traitée dans un service ou bien qui vont la traitée dans le rendez-vous.
- Date de rendez-vous (item3) : sont les rendez-vous fixés pour le patient est classés selon trois catégories (proche, loin, abandonner) :
  - Proche : les dates sont moins de deux mois.
  - Loin : les dates sont entre deux mois et six moins.
  - Abandonner : les dates sont plus de six mois.

Nous avons utilisé ces items pour modéliser notre problème et nous permettre d'extraire les services qui ont une surcharge, l'exemple est montré dans le tableau 4.1.

Tableau 4.1 : Exemples des transactions.

items	Item 1	Item 2	Item 3	Nombre d'apparition	Il y a un problème ?
T1	Gynécologie	Fibromes utérins	A	27	?
T2	Ophtalmologie	Uvéite	A	4	?
T3	Néphrologie	Insuffisance rénale	A	3	?

3.4. Création des Clusters des profils décideurs

Le nombre des profils des utilisateurs est très grand, donc nous créons des groupes de Clusters qui ont les mêmes caractéristiques, et pour se là nous avons utilisé la méthode de clustering de Data Mining, et nous avons choisi l'algorithme de k-means avec un choix de cinq clusters. Les résultats sont affichés dans la figure suivante.

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(422.0)	(110.0)	(70.0)	(96.0)	(96.0)	(50.0)
Nom_arr	Hopital	Clinique	Hopital	Hopital	Hopital	Hopital
Zone_arr	50KM	20KM	50KM	10KM	5KM	50KM
ID_preferences	1	10	12	1	1	10
ID_formation	2	2	2	1	1	2
ID_personnalite	5	10	13	2	5	5
ID_opp	4	4	2	1	3	4
ID_danger	3	4	4	3	3	4
Nom_Objectif	Ouvrir_Un_Nouveau_Service	Ouvrir_Un_Nouveau_Service	Agrandir_Un_Service	Ouvrir_Un_Nouveau_Service	Ouvrir_Un_Nouveau_Service	Ouvrir_Un_Nouveau_Service
ID_probleme	5	5	1	5	3	5
ID_service	10	10	1	19	11	10
ID_budget	3	3	2	3	3	3
ID_humain	1	69	1	1	12	1
suffisance_materiel	OUI	OUI	OUI	NON	NON	OUI

Figure 4.6:Résultats de clustering.

3.5. Utilisation de la personnalisation pour la classification

Après l'extraction des motifs fréquents et le clustering des profils des décideurs on veut faire la relation entre les deux résultats précédents pour répondre à la question «L'itemset est-il intéressant ou non».

Dans notre travail, nous avons proposé un formulaire qui contient les itmsets sur les lignes et les clusters sur les colonnes, et pour remplir la classe (Opinion), sa valeur prend deux cas (I : Intéressé, N : Non intéressé), l'exemple de formulaire est montré dans le tableau 4.2.

Tableau 4.10: Exemple du formulaire.

Itemset	C1	C2	C3	C4	C5
Itemset 1	I/N	I/N	I/N	I/N	I/N
..					
Itemset 168	I/N	I/N	I/N	I/N	I/N

Un sondage des gens du domaine médical a été fait pour remplir un formulaire. Ce formulaire représente la classe "Opinion" qui prend en considération les paramètres de chaque Cluster (Personnalité du décideur, Son objectif, les risques … etc.). La description de chaque Itemset associe à ces Clusters détermine le cas s'il est intéressant ou pas intéressant. Un schema explicatif de classification avec NaiveBayes est montré dans la figure suivante.

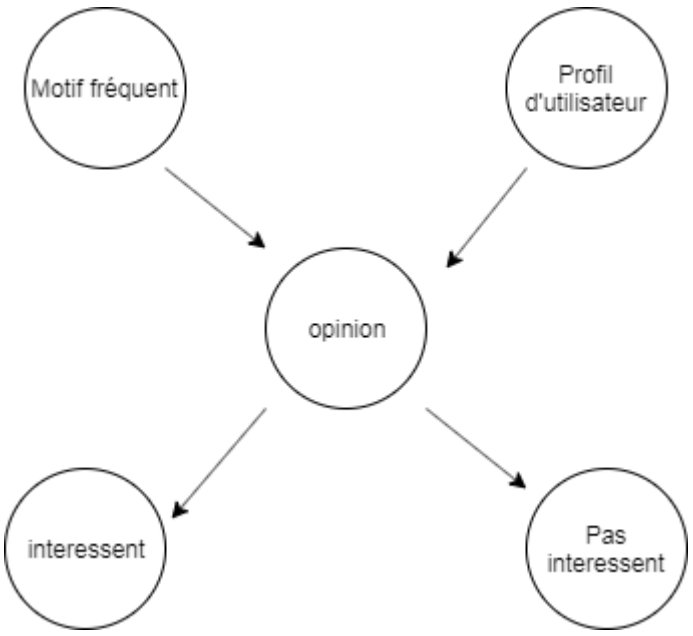


Figure 4.7 : Le réseau NaiveBayes qui représente le modèle de connaissances de décideur.

Exemple d'une ligne de remplissage :

- Itemset : INDICE_DATE=L.CODESERV=8 qui signifie que les dates des RDV sont loin dans le service de la gynécologie.

- Avec le Cluster (C1) :

  Le décideur à des préférences pour que tout le monde à une accessibilité aux soins (ID_preference =10), et une personnalité de type : (ISFJ- le Défenseur) qui est une personnalité des gens qui sont ouvert au changement, curieux et qui sont logique dans leurs prise de décisions.

Ce qui nous donne un Opinion = I (Intéressant) par son objectif qu'il est d'ouvrir un nouveau service.

Une fois les cases "Opinion" sont toutes remplies, l'algorithme d'apprentissage supervisé (NaiveBayes) est lancé pour créer notre modèle de connaissances de décideur. Le choix de l'algorithme NaiveBayes a été fait après une comparaison empirique entre plusieurs algorithmes de classification. Cette comparaison a monté que NaiveBayes a la meilleure précision. (Pour plus de détails, voir chapitre 5)

4. Conclusion

Dans ce chapitre, nous avons illustré la modélisation du profil d'un décideur et les approches utilisé pour l'extraction des itemsets fréquents, ainsi que l'utilisation de l'aspect de la personnalisation pour la classification. Le chapitre suivant est consacré à démontré l'efficacité de ces méthodes grâces aux résultats obtenus.

Tests et validation

1. Environnement de développement

1.1. L'environnement matériel

Nous avons utilisé :

- Un ordinateur portable Sony_Vaio avec les caractéristiques suivantes :
  - 4 GO RAM
  - 4096 MBytes DDR3
  - Nvidia GeForce 410M GPU
  - Intel (R) Core (TM) i5-2450M CPU @ 2.50 GHz
  - Système d'exploitation : Windows 10 de 64-bit

1.2. L'environnement logiciel

1.2.1. Java 8.0

Java 8 est la dernière version de Java et offre de nouvelles fonctionnalités, des performances accrues et des corrections de bug pour améliorer l'efficacité de développement et d'exécution des programmes Java. La nouvelle version de Java est d'abord mise à disposition des développeurs afin qu'ils disposent du temps adéquat pour effectuer les opérations de test et de certification. **[74]**

1.2.2. Netbeans IDE 8.2

L'EDI NetBeans est un environnement de développement - un outil pour les programmeurs pour écrire, compiler, déboguer et déployer des programmes. Il est écrit en Java - mais peut supporter n'importe quel langage de programmation. Il y a également un grand nombre de modules pour étendre l'EDI NetBeans. L'EDI NetBeans est un produit gratuit, sans aucune restriction quant à son usage. **[75]**

L'environnement de base comprend les fonctions générales suivantes :

- Configuration et gestion de l'interface graphique des utilisateurs,
- Support de différents langages de programmation,
- Traitement du code source (édition, navigation, formatage, inspection),
- Fonctions d'import/export depuis et vers d'autres IDE, tels qu'Eclipse ou JBuilder,
- Accès et gestion de bases de données, serveurs Web, ressources partagées,
- Gestion de tâches (à faire, suivi…),
- Documentation intégrée

1.2.3. SQL Server

Microsoft SQL Server**,** aussi appelé « SQL Server » ou parfois plus simplement « MSSQL », est un système de gestion de base de données relationnelle développé par Microsoft. **[76]**

Version : SQL Server 2008 R2 (nom de code Kilimanjaro)

SQL Server gère donc, comme la plupart des autres SGBDR du marché, différents types de fichiers, dans notre cas on a restauré une BD sauvegardé de type « .bak »

Ceci se fait à l'aide de la commande BACKUP :

```
RESTOREDATABASE Teste1 FROMDISK=N'D:\BACKUP\Test1.bak'
```

1.2.4. Weka 3.8

Weka est une collection d'algorithmes d'apprentissage automatique pour les tâches d'exploration de données. Il contient des outils pour la préparation des données, la classification, la régression, le regroupement, l'exploration de règles d'association et la visualisation. **[77]**. Weka est un logiciel open source distribué sous licence GNU General Public License.

2. Expérimentation et tests

2.1. Interface de saisie d'un nouveau profil

Chaque nouveau décideur devra entrer ses données personnelles ainsi que les données de son environnement, son objectif décisionnel, ses préférences et ses problèmes.

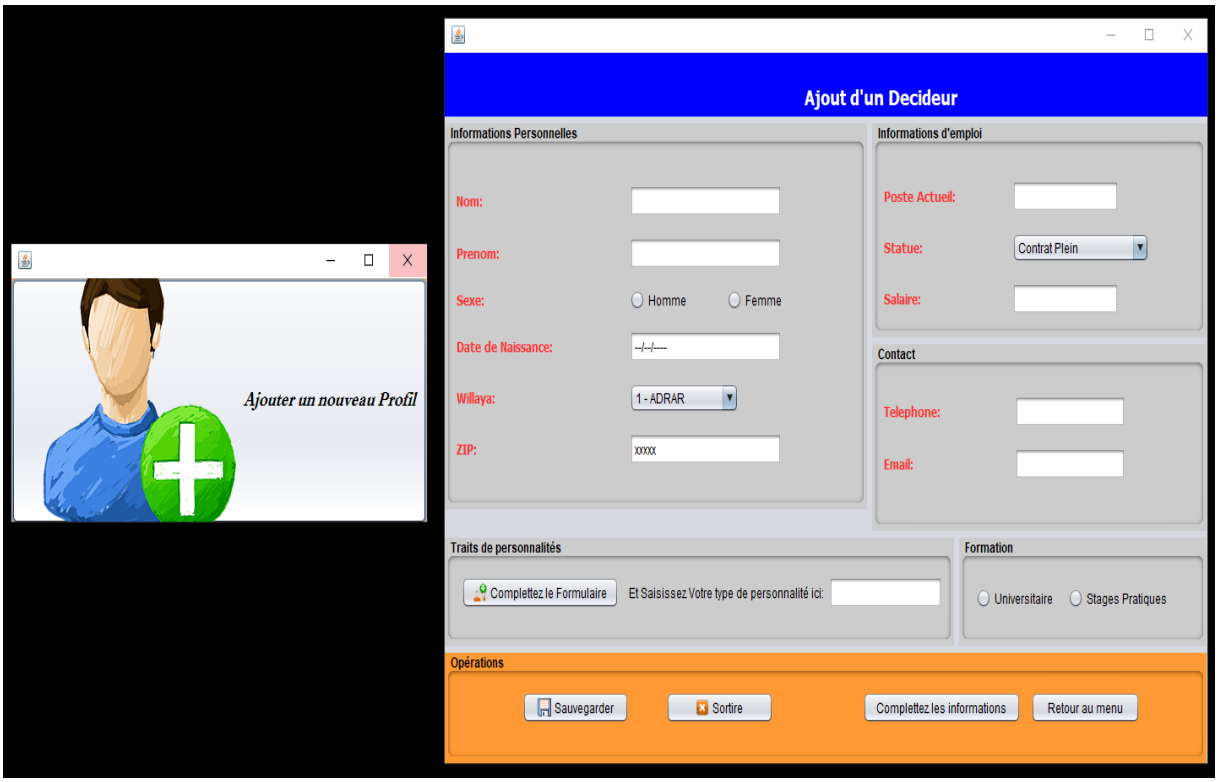


Figure 5.10: Fenêtre de saisie informations personnelles.

Pour les traits de personnalité, on a opté pour l'utilisation **MBTI** qu'il est l'outil d'évaluation psychologique déterminant le type psychologique.

Le test MBTI permet de décrire la personnalité de façon dynamique et détaillée, avec un langage simple permettant de donner un nom à des phénomènes insaisissables.

Pour ce cas on a utilisé l'outil en ligne « https://www.16personalities.com » Pour obtenir une description concrète et exacte de la personnalité du décideur basée sur des atouts scientifiques, un exemple des résultats est montrer dans la figure 5.2.

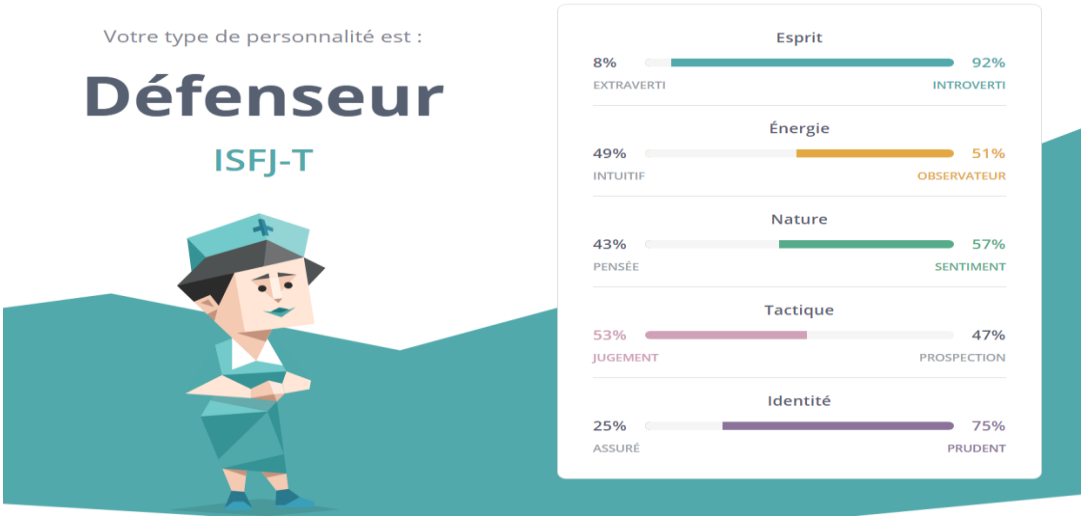


Figure 5.11: Résultat du teste MBTI.

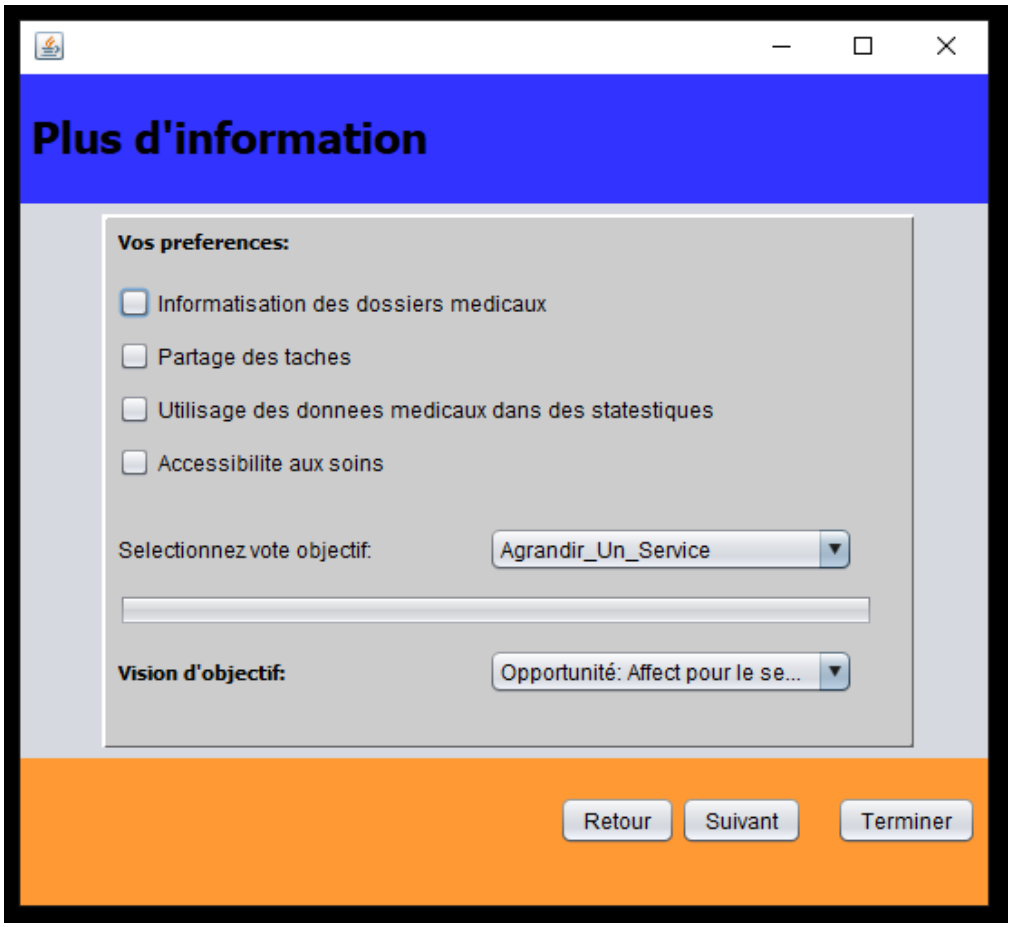


Figure 5.12: Fenêtre de saisie objectif et préférences.

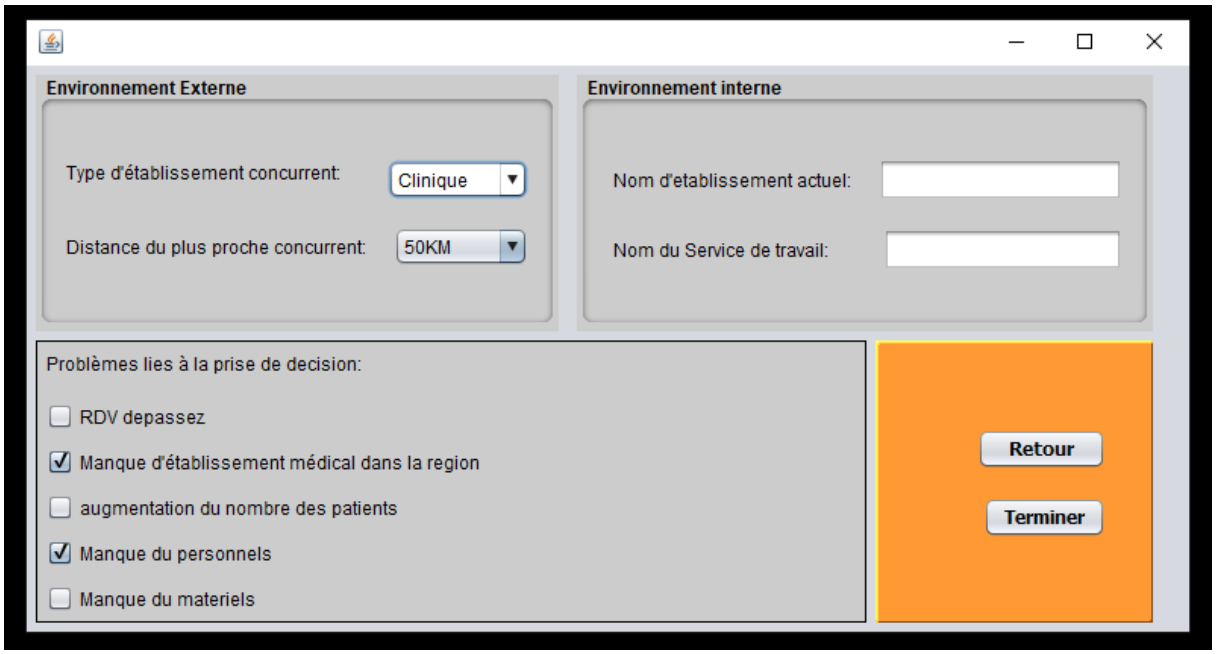
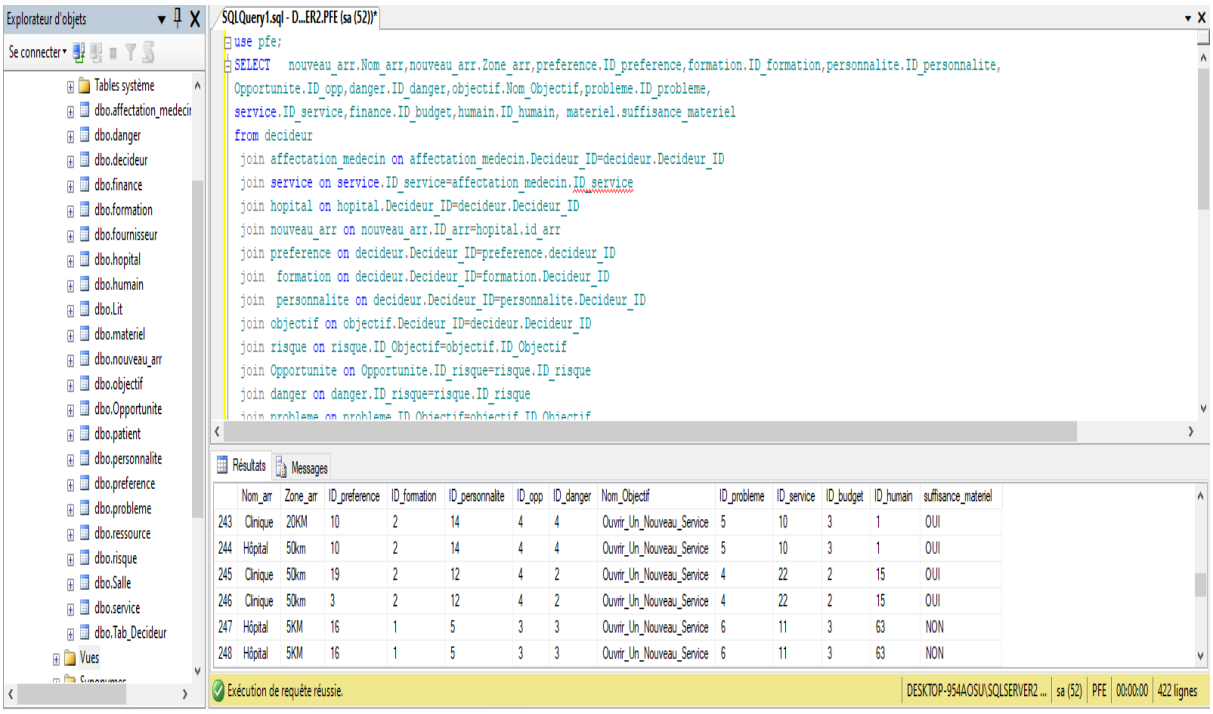


Figure 5.13: Fenêtre de saisie environnement et problèmes.

Après avoir saisie les informations de profil décideur, ses informations sont stockées dans la base de données. A l'aide d'une vue matérialisé figure 5.5 on récupère le fichier des profils décideurs entré, à fin de les classé dans des clusters.



	Nom_arr	Zone_arr	ID_preference	ID_formation	ID_personnalite	ID_opp	ID_danger	Nom_Objectif	ID_probleme	ID_service	ID_budget	ID_humain	suffisance_materiel
243	Clinique	20KM	10	2	14	4	4	Ouvrir_Un_Nouveau_Service	5	10	3	1	OUI
244	Hôpital	50km	10	2	14	4	4	Ouvrir_Un_Nouveau_Service	5	10	3	1	OUI
245	Clinique	50km	19	2	12	4	2	Ouvrir_Un_Nouveau_Service	4	22	2	15	OUI
246	Clinique	50km	3	2	12	4	2	Ouvrir_Un_Nouveau_Service	4	22	2	15	OUI
247	Hôpital	5KM	16	1	5	3	3	Ouvrir_Un_Nouveau_Service	6	11	3	63	NON
248	Hôpital	5KM	16	1	5	3	3	Ouvrir_Un_Nouveau_Service	6	11	3	63	NON

Figure 5.14: Vue matérialisé des profils décideurs

Une fois le fichier de la vue est fait, on utilise l'algorithme k-means de l'apprentissage non supervisé pour faire le clusturing. Ça nous permet de regrouper les profils dans cinq clusters C1, C2, C3, C4 et C5, les résultats sont montrer dans les figures 5.6 et 5.7.

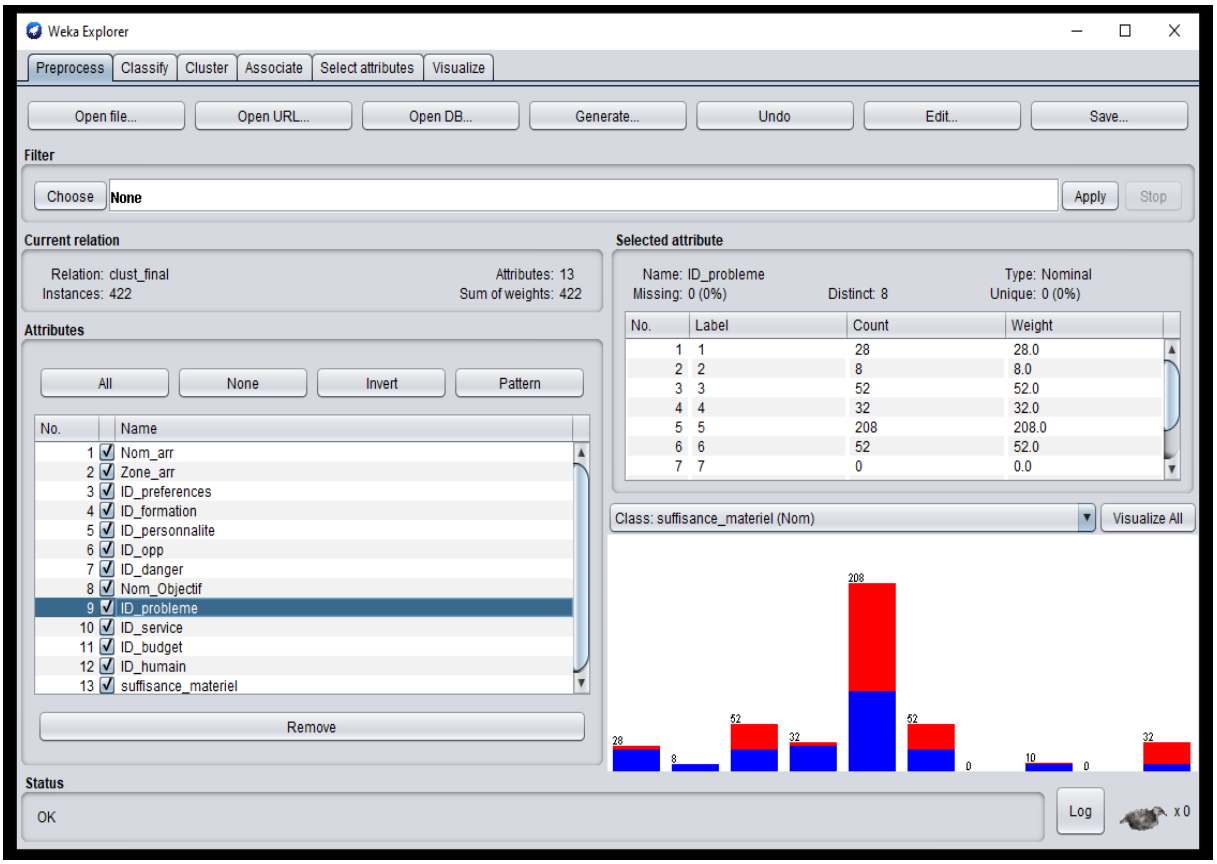


Figure 5.15: Profils des décideurs implémentés sur Weka.

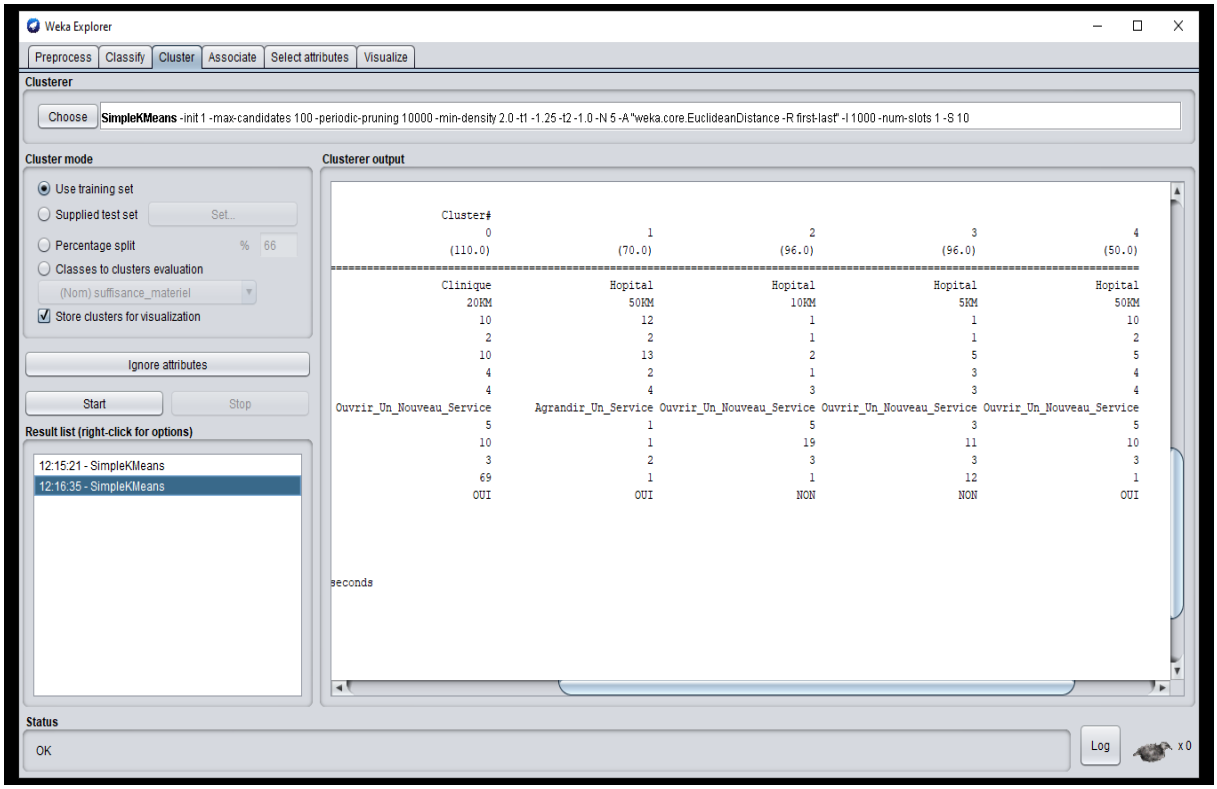


Figure 5.16: Résultat du Clusturing.

On a comme résultats cinq clusters avec des paramètres de décideurs propres à chaque catégorie de ses cinq clusters.

2.2 Phase d'extraction des itemsets fréquents

Pour cette étape, on a sélectionné les données à utiliser pour l'extraction des connaissances répondant à l'objectif décisionnel choisi qui est l'éclatement des services. Ces données sont les RDVs programmés pour chaque médecin qui appartient à un service, ainsi que les symptômes de la maladie enregistrées dans ses services, comme montrer la figure 5.8.

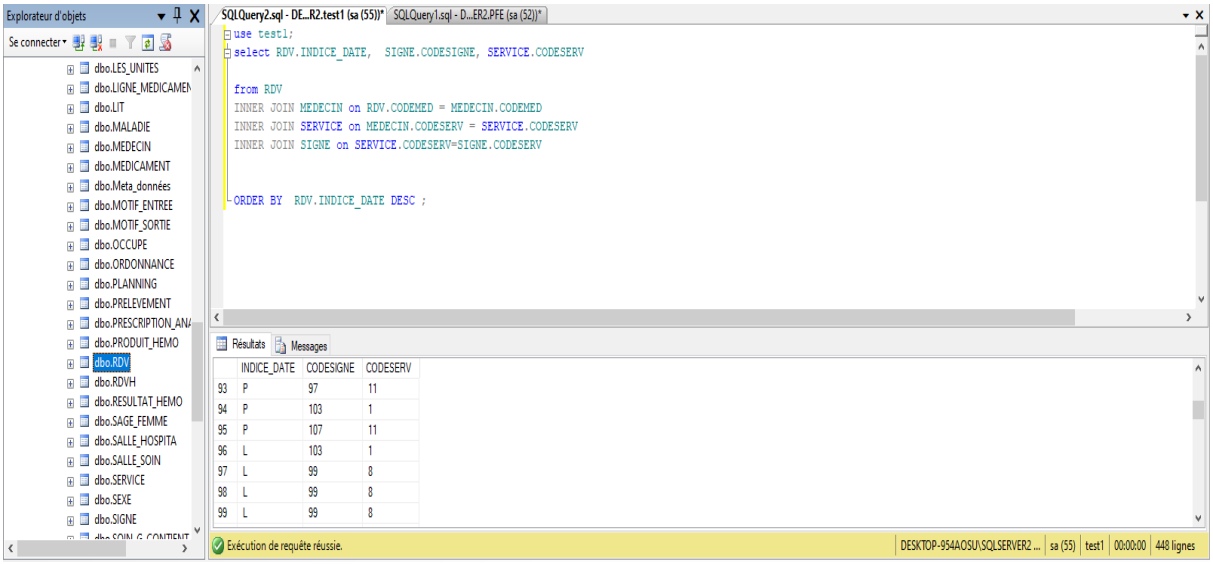


Figure 5.17: Vue matérialisé des données qu'on va les extraires.

On utilise « Apriori » qui est l'algorithme d'extraction des itemsets fréquents avec un MinSupp = 0.2. On obtient (168) Itemsets fréquents de taille 1 figure 5.9, taille 2 figure 5.10 et taille 3 figure 5.11 et la figure 5.12 montre la dispersion d'itemsets fréquents.

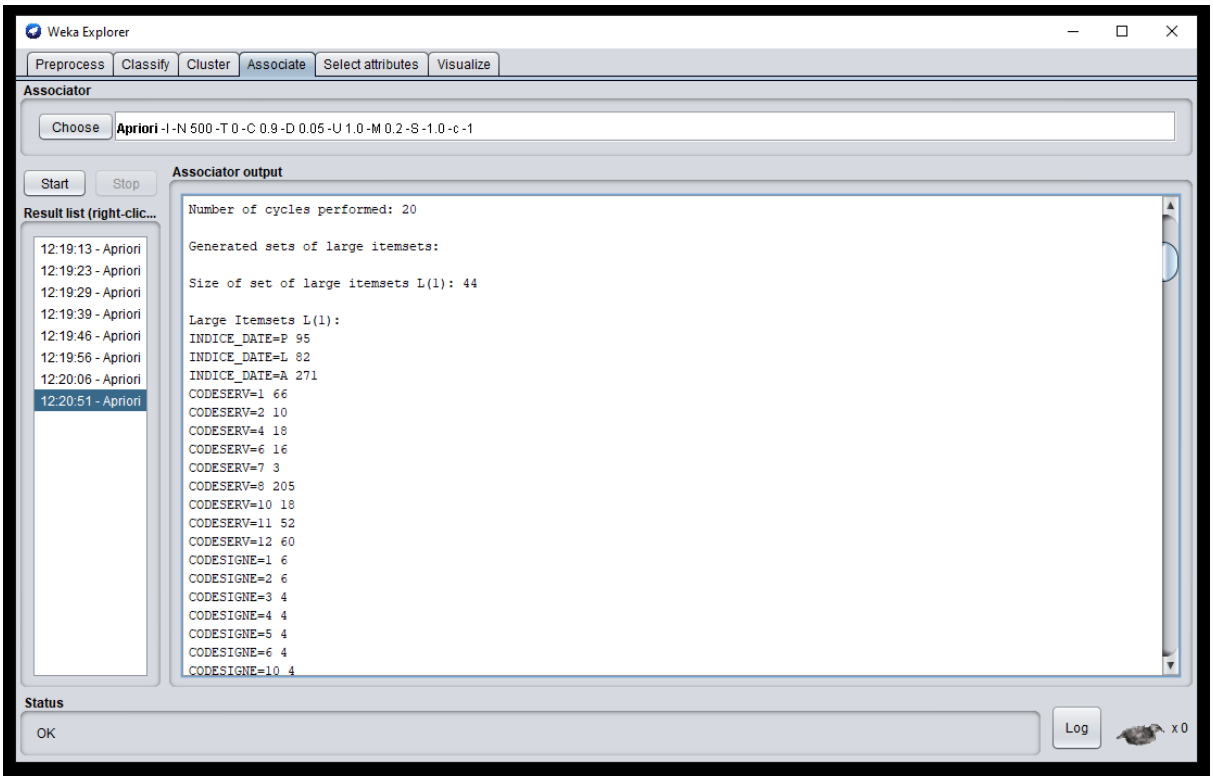


Figure 5.18: Itemsets de taille 1.

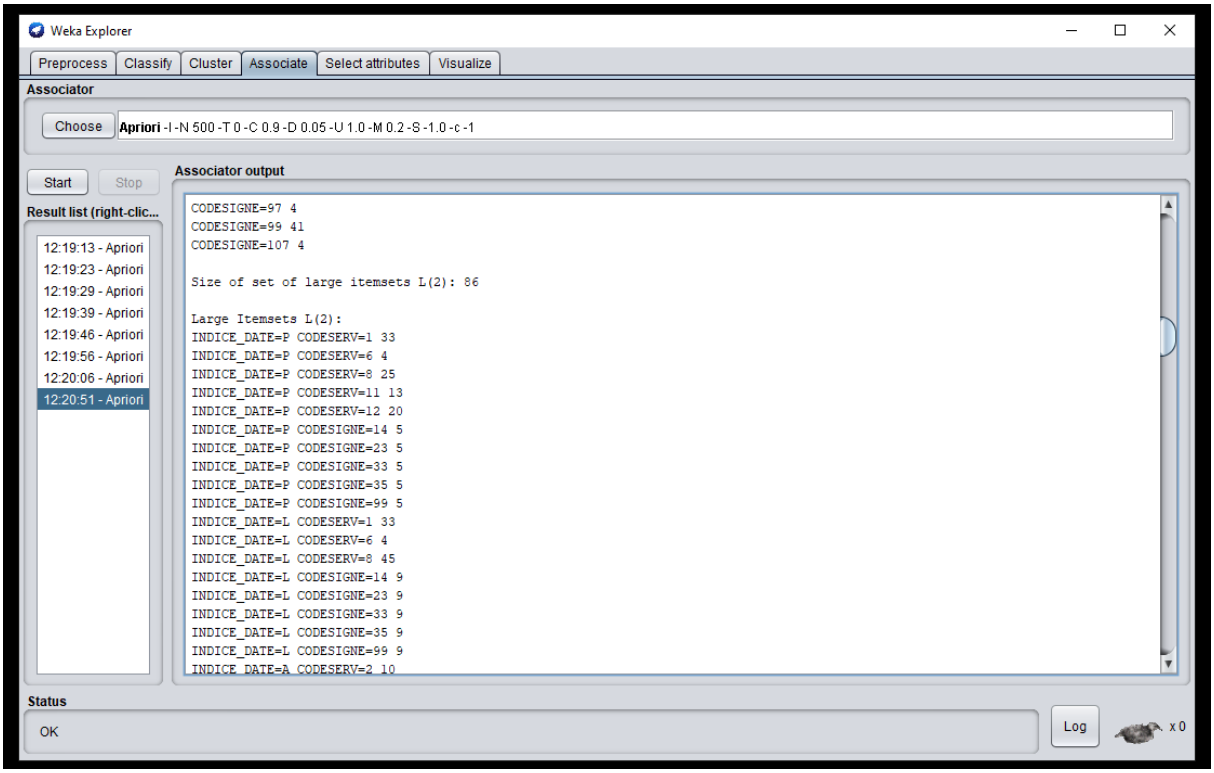


Figure 5.19: Itemsets de taille 2.

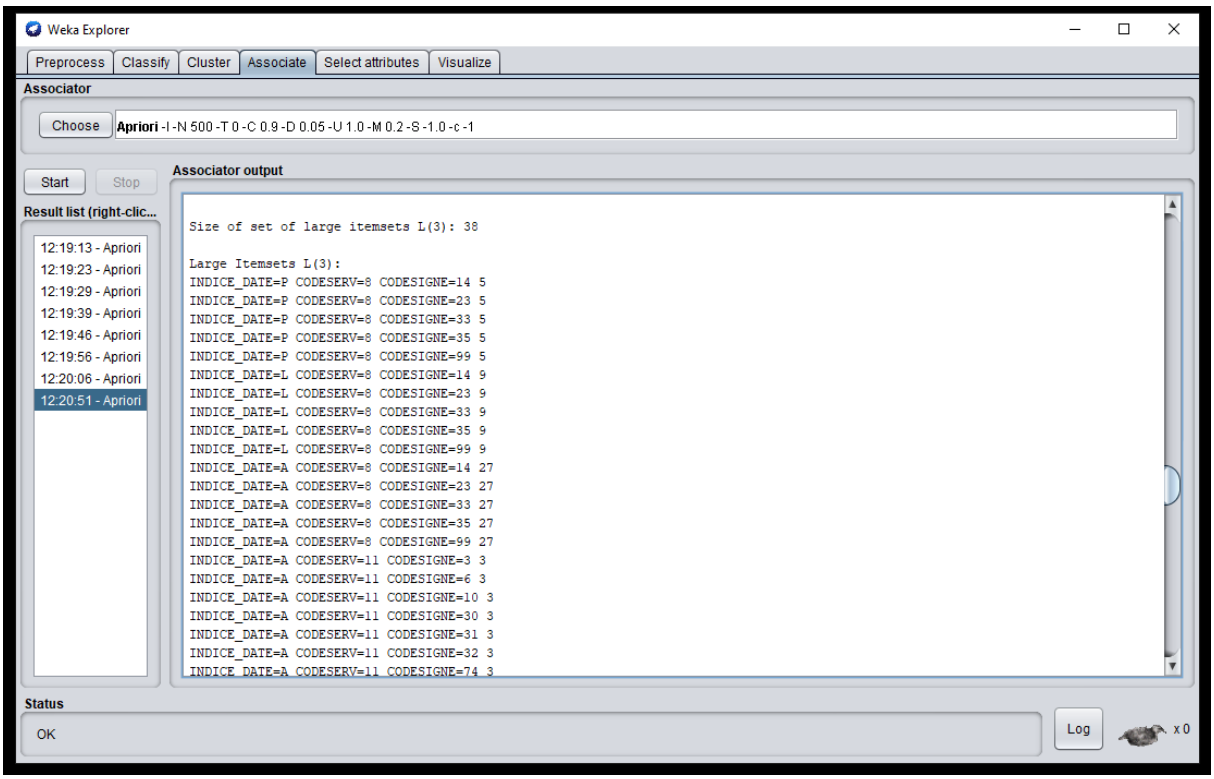


Figure 5.20: Itemsets de taille 3.

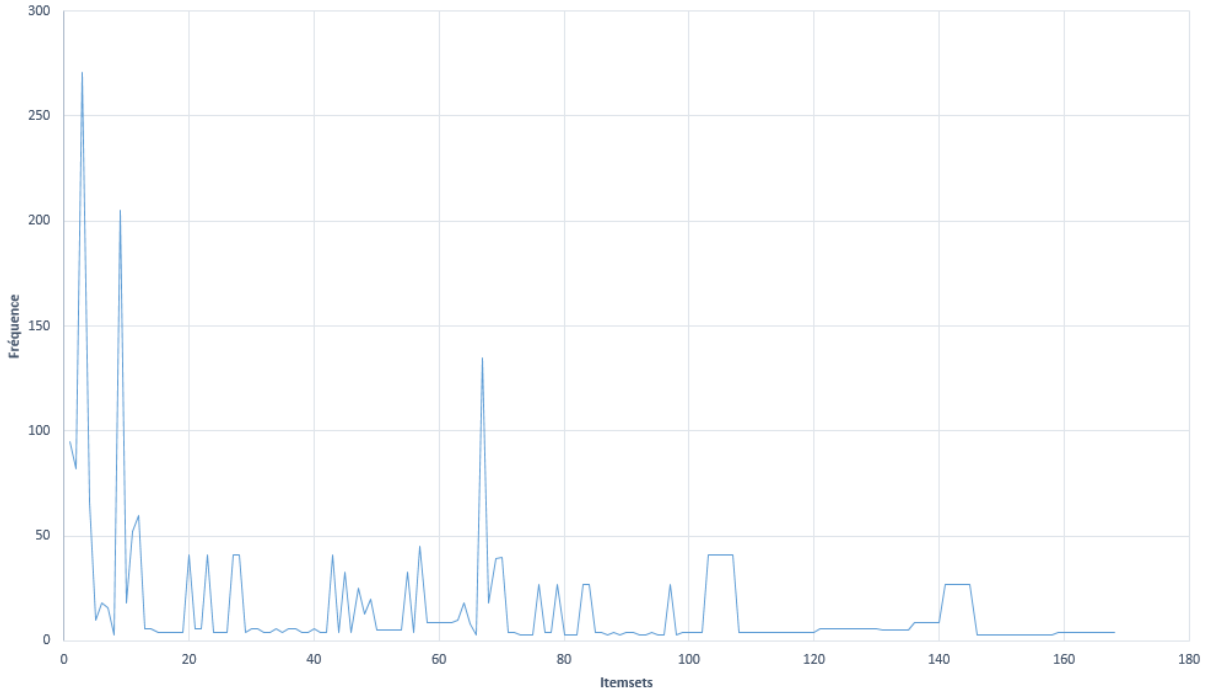


Figure 5.21: La dispersion d'itemsets fréquents.

a) Base de connaissances

Dans notre travail, On a proposé une valeur de classe (Opinion) pour Chaque Itemset résultants d'Apriori dont la classe prend deux valeurs (I : Intéressé, N : Non intéressé) en fonction des attributs de chaque Cluster du décideur. Dans notre cas, on a cinq clusters de décideurs, et (168) Itemset fréquents.

Tableau 5.11: Les données d'apprentissage de modèle de connaissance du décideur.

Itemset	C1	C2	C3	C4	C5
Itemset 1	I/N	I/N	I/N	I/N	I/N
..					
Itemset 168	I/N	I/N	I/N	I/N	I/N

Le remplissage de ce formulaire tableau 5.1 à étais fait par un sondage des gens des domaines médical, la classe opinion à étais remplis en prenant en considération les paramètres de chaque Cluster (Personnalité de décideur, Son objectif, les risque … etc). La description de chaque Itemset associe à ces Clusters détermine le cas s'il est intéressant ou pas.

b) Création du Modèle d'apprentissage

On a utilisé l'algorithme d'apprentissage supervisé (NaiveBayes) pour crée notre modèle d'apprentissage supervisé qui représente le modèle de connaissances du décideur, les resutats sont dans la figure 5.13.

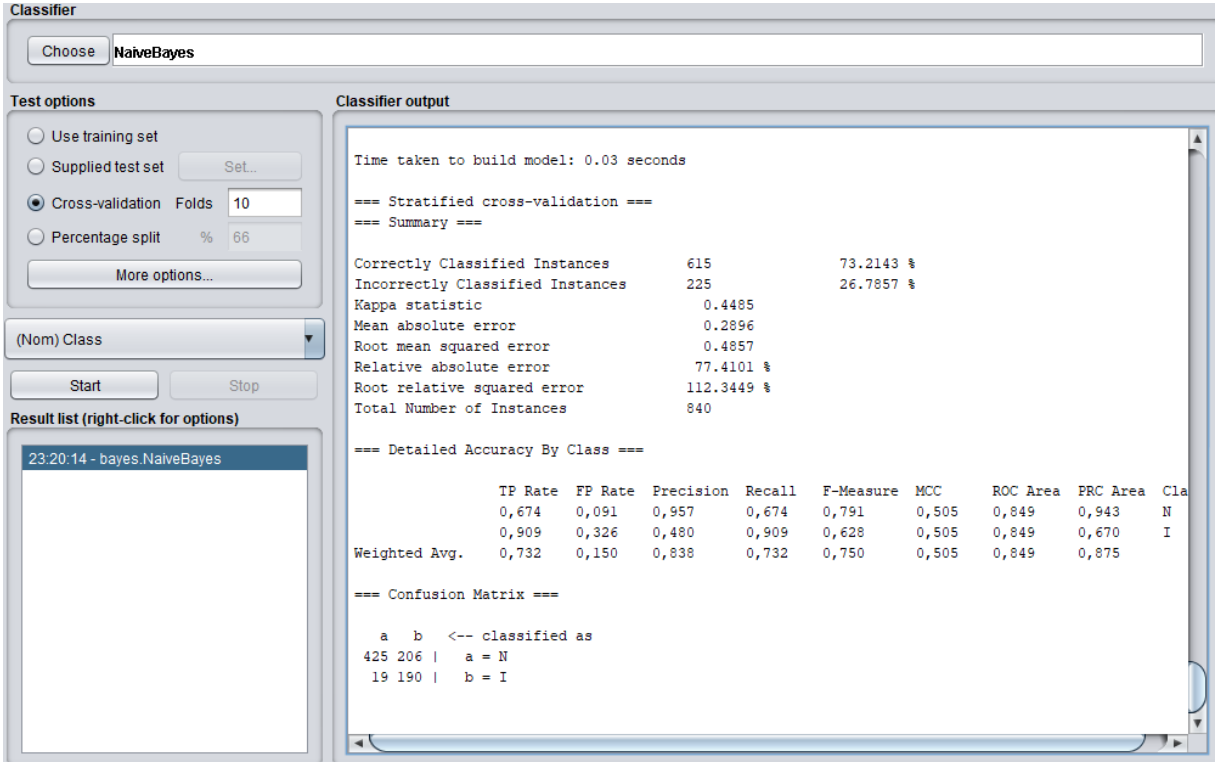


Figure 5.22: La Classification de la classe « Opinion » avec NaiveBayes.

On a opté pour cet algorithme après une comparaison qui a été faite sur la base de la précision du modèle généré. Comme c'est montré sur les figures5.14 ,5.15 et la figure 5.16, le bayesNaive est plus précis que les algorithmes : Arbre de décision J48 et Table de décision jRip.

```
=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,986    0,469    0,864      0,986    0,921      0,639    0,930     0,975     N
                0,531    0,014    0,925      0,531    0,675      0,639    0,930     0,839     I
Weighted Avg.   0,873    0,356    0,879      0,873    0,860      0,639    0,930     0,941

=== Confusion Matrix ===
```

Figure 5.23: Précision du NaiveBayes.

```
=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,960    0,359    0,890      0,960    0,924      0,664    0,930     0,974     N
                0,641    0,040    0,843      0,641    0,728      0,664    0,930     0,833     I
Weighted Avg.   0,881    0,279    0,878      0,881    0,875      0,664    0,930     0,939
```

Figure 5.15: Précision du jRip.

```
=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                1,000    1,000    0,751      1,000    0,858      ?        0,497     0,750     N
                0,000    0,000    ?          0,000    ?          ?        0,497     0,248     I
Weighted Avg.   0,751    0,751    ?          0,751    ?          ?        0,497     0,625
```

Figure 5.16: Précision du J48.

Voilà, maintenant que le modèle est créé, en peux entamer les tests.

3. Tests et validation

Dans cette phase de notre travail, nous avons introduit des nouveaux critères de sélection, pour cerner mieux notre objectif et ensuite orienté la prise de décision. Nous avons inclus une nouvelle table “hotelerie” Qui contient les informations sur la gestion d’hôtellerie de l’hopital, les services de l’hôpital avec l’état de disponibilité des salles pour chaque service. Ce type de connaissance n’existait pas dans le modèle d’apprentissage créer dans l’étape précédente de notre travail. Ce modèle va générer automatiquement de nouvelles connaissances qui seront exploitées par la suite dans l’extraction des motifs frequents.

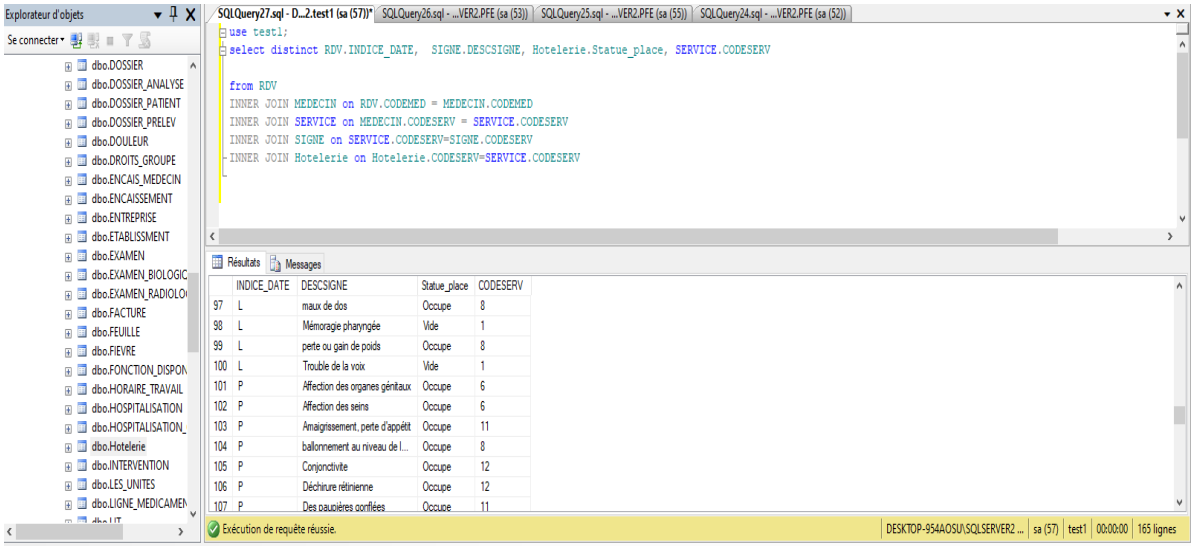


Figure 5.24: Vue matérialisé des données à utiliser lors de cette extraction.

On exécute toujours l'algorithme « Apriori » avec un MinSupp de 0.1, on obtient les (20) Itemsets fréquents.

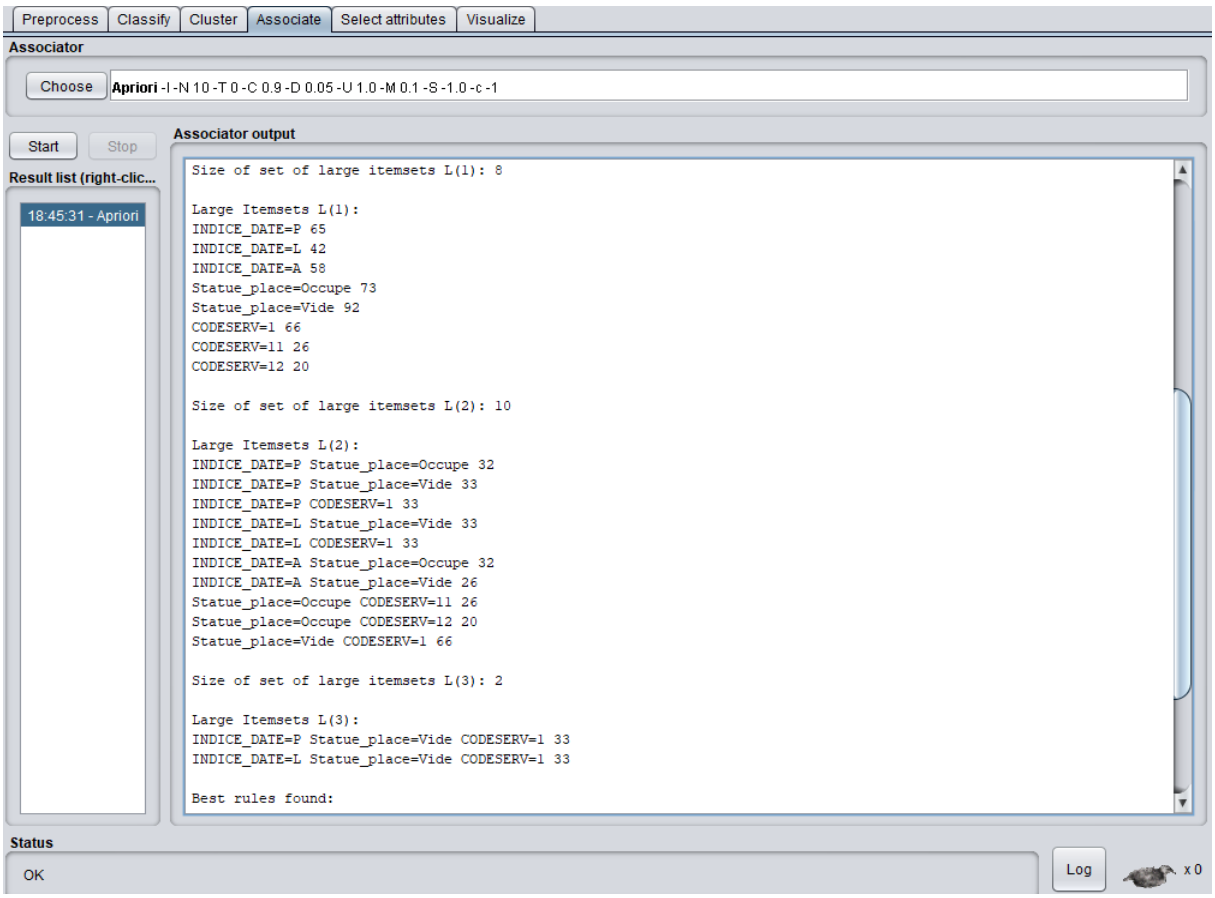


Figure 5.18 : Itemsets pour le test.

On utilise notre modèle d'apprentissage déjà généré pour voir parmi les quelles d'entre ses itemsets sont Intéressants pour un profil de décideur choisis.

Pour ce test, on tire deux profils parmi les cinq (C1 et C2):

Pour le Profil C1 :

```
Time taken to build model: 0 seconds

=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1        1:?        1:N       0.83
        2        1:?        2:I       0.881
        3        1:?        2:I       0.881
        4        1:?        1:N       0.632
        5        1:?        1:N       0.632
        6        1:?        1:N       0.67
        7        1:?        1:N       0.67
        8        1:?        1:N       0.67
        9        1:?        1:N       0.632
       10        1:?        1:N       0.632
       11        1:?        1:N       0.83
       12        1:?        1:N       0.632
       13        1:?        1:N       0.632
       14        1:?        1:N       0.632
       15        1:?        1:N       0.632
       16        1:?        1:N       0.632
       17        1:?        1:N       0.632
       18        1:?        1:N       0.632
       19        1:?        1:N       0.632
       20        1:?        1:N       0.632

=== Evaluation on test set ===
```

Figure 5.19: Prédiction classe pour C1

On remarque que parmi les (20) itemsets, il y a que deux itemsets (2 et 3) qui sont intéressant pour ce type de profils du décideur qui appartient C1.
Les instants qui sont intéressant, sont ceux avec des Rendez-vous loin ou abandonné.

Pour le Profil C2

```
Time taken to build model: 0.01 seconds

=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1        1:?        1:N       0.699
        2        1:?        2:I       0.939
        3        1:?        2:I       0.939
        4        1:?        2:I       0.55
        5        1:?        2:I       0.55
        6        1:?        2:I       0.508
        7        1:?        2:I       0.508
        8        1:?        2:I       0.508
        9        1:?        2:I       0.55
       10        1:?        2:I       0.55
       11        1:?        1:N       0.699
       12        1:?        2:I       0.55
       13        1:?        2:I       0.55
       14        1:?        2:I       0.55
       15        1:?        2:I       0.55
       16        1:?        2:I       0.55
       17        1:?        2:I       0.55
       18        1:?        2:I       0.55
       19        1:?        2:I       0.55
       20        1:?        2:I       0.55

=== Evaluation on test set ===
```

Figure 5.20: Prédiction classe pour C2

Pour les décideurs qui ont un Profils C2, les Itemsets qui ne les intéressent pas sont ceux où il y'a une Date de rendez-vous Proche, le reste de ses Itemsets sont de type Intéressant.

3.1 D'autre Exemple de test

Voici d'autre résultats de test pour d'autre Instances, pour voir si les résultats son correcte ou pas.

On a opté pour (4) Itemsets dont chaqu'un est associé à un profil d'un décideur pour voir le résultat de la prédiction.

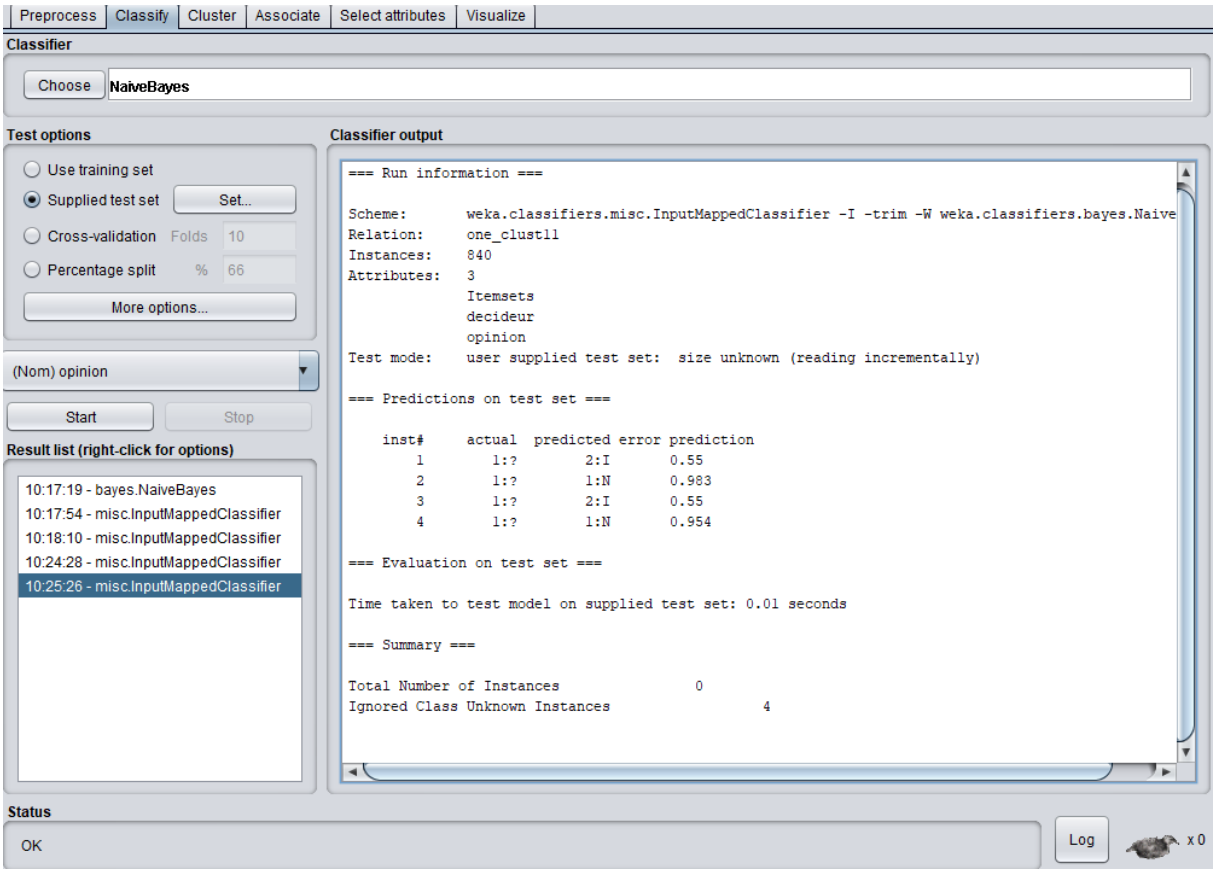


Figure 5.21: résultats du test sur Weka.

Tableau 5.12: Exemple de test du modèle.

inst#	actuel	prédiction	taux de prédiction
INDICE_DATE=L.Statue_place=Occupe,C2,?	?	I	0.55
INDICE_DATE=P.Statue_place=Vide,C4,?	?	N	0.983
INDICE_DATE=A.CODESERV=9.Statue_place=Occupe,C2,?	?	I	0.55
INDICE_DATE=A.CODESERV=12.CODESIGNR=40.Statue_place=Vide,C3,?	?	N	0.954

Le tableau en dessus, montre la prédiction de la classe (Opinion : Intéressant, Non intéressant) pour chaque instance Itemset, profil d'un décideur.

L'INDICE_DATE représente les RDV fixé pour une consultation donné, son définie par :

-L'indice P : pour les rendez-vous proche (moins de 2mois)
-L'indice L : pour les rendez-vous loin (entre 2 et 6 mois)
-L'indice A : pour les rendez-vous abandonné (plus de 6mois)

La Statue_place représente l'information qui concerne la disponibilité de place dans un service, elle prend deux valeurs : Occupe ou Vide.

Pour le CODESERV c'est l'indicateur du service, par exemple :

CODESERV=9 indique le service d'Hématologie.

La même pour CODESIGNE c'est l'indicateur du symptôme d'une maladie, par exemple : CODESIGNR=65 indique le symptôme de Myopie.

3.1.1 Explication des résultats

INDICE_DATE=A.CODESERV=9.Statue_place=Occupe,C2,?
On veut savoir si l'itemset (INDICE_DATE=A.CODESERV=9.Statue_place=Occupe) avec le profil de décideur de catégorie C2, est intéressant ou pas.
On a des RDV abandonné dans le service d'Hématologie (9) avec des places non disponibles dans le même service.

Et on voit les paramètres du profil C2 qu'il s'agit de son objectif qu'il est d'Agrandir le service et en se basant sur sa personnalité (Professeur) qu'elle se base sur la recherche d'explications, bâtir des théories, à trouver des solutions. Il est organiser il préfère un environnement structuré, ordonné et prévisible, qu'il peut le contrôler.

En même temps imaginatif dans la recueille des informations abstraites et intangibles. Avec une vision évolutive, que les Rdv dépassé l'empêche de réaliser, On voie bien que la prédiction de la classe opinion qu'elle est de type (I : Intéressant) est une prédiction qui tient la route dans le cas de cet profil C2.

-Dans l'instance 2 du teste on a l'instance suivants :

INDICE_DATE=P.Statue_place=Vide,C4,?

Dans ce cas-là, tous les RDV ont des dates proches, ainsi qu'une disponibilité des places dans le service.

La prédiction de 0.983 pour la classe opinion (N : Non intéressant) est totalement justifier, on voit l'objectif du profil décideur C4 qu'il est d'ouvrir un nouveau service.

-Dans l'instance 4 du teste on a l'instance suivants :

INDICE_DATE=A.CODESERV=12.CODESIGNR=40.Statue_place=Vide,C3,?

Des dates Abandonné pour les Rdv du service Ophtalmologie=12, dont la maladie est dus au symptôme (40) : Rétinopathie diabétique.
Mais avec une Statue_place=Vide, ce qui signifie une disponibilité de places dans ce service.
La prédiction est (N : Non intéressant) Par L'ouverture d'un nouveau service qu'il est l'objectif du profil C3.

Malgré les RDV dépassé, le modèle de classification a prédit de 95.4% la valeur (N) pour la classe opinion.

Il s'agit du paramètres Nouveau_arr ainsi que Zone_Act qui signifiant les concurrents et leurs placements par rapport au service étudie.

Dans ce cas en a un Hôpital à 10KM de ce service, donc les résultats et si logique en voyons le contexte d'un autre établissement proche ainsi que la disponibilité déjà des places dans ce service malgré les RDV dépassé.

A la fin, Ces résultats de tests sont logiques, mais malgré ça il pourrait y'en avoir d'autre Itemset fréquents en fonction d'autre Profils de décideurs, qui vont apparaitre fausses.

Cela est dû au modèle d'apprentissage, qui ne contient pas beaucoup d'information (dataset d'apprentissage contient 168 enregistrements), du coup plus le modèle englobe plus d'information plus les résultats sont plus fiables.

4. Conclusion

Dans ce chapitre nous avons présenté l'environnement matériel et l'environnement logiciel, les diffèrent méthodes et algorithmes utilisé pour implémenter notre approche. On a effectué des tests pour vérifier la qualité des résultats de notre approche.

Conclusion Générale

1. Conclusions

L'extraction de connaissances à partir de données (ECD) est définie comme un processus de découverte d'informations implicites, inconnues auparavant et potentiellement utiles à partir de données. Néanmoins, ces connaissances peuvent être utiles pour un utilisateur et non pour un autre. C'est dans ce contexte que s'intègre notre travail. Plus précisément, on s'est intéressé dans notre travail à l'extraction des motifs fréquents orientée besoin du décideur. La génération d'un nombre important des motifs fréquents est parmi les caractéristiques de cette méthode, ce qui il nous a poussé à la choisir comme étant une technique de Data Mining dans notre cas d'étude.

Nous avons consacré le premier chapitre pour l'extraction des connaissances à partir de données (ECD) et nous avons évoqué le processus d'ECD et ses étapes. Nous avons donné une brève description de technique de base de data mining qui sont relatives à l'analyse des données et définie les deux type d'apprentissage (supervisé é et non supervisé) qu'on a utilisée, a la fin on a parlés sur le data mining dans le domaine médical et ses problèmes.

Dans le deuxième chapitre, nous avons présenté l'aspect de la personnalisation basé sur les profils, nous avons aussi explicité quelques notions de la personnalisation et définit les profils et leur techniques de construction.

Dans le troisième chapitre, nous somme focalisé sur la technique d'extraction des motifs fréquents on détaler les algorithmes fondamentaux de la technique et faire des comparaisons pour objectif est de passer en revue les forces et les faiblesses de ses algorithmes.

A la fin, nous avons proposé un modèle des besoins de décideur et nous l'avons fusionné avec l'extraction des motifs en utilisant des technique de data mining (apprentissage superviser et non superviser) pour pouvoir d'avoir des motifs fréquents orienter besoin de décideur. Nous avons testé la solution pour voir sa fiabilité, mais il n y a pas un système entièrement parfait.

2. Perspectives

- Enrichir le modèle du décideur afin de couvrir plus d'informations qui ont une influence sur la décision.

- Donnée la capacité de voter aux les utilisateurs pour permettre à notre système d'appondre plus.
- Intégrer les données du modèle de décideur à l'intérieur d'un des algorithmes d'extraction des motifs fréquents afin de préfiltrer les motifs fréquents suivant les contraintes de décideurs et générer que les motifs intéressants.

Références

Références

[1] R. Chalal. " »Une approche pour la capitalisation coopérative des connaissances sur les risques produit en phase initiale d'un projet industriel ». Thèse doctorat, INI, Décembre 2007.

[2]U.M. Fayyad, G. Piatesky-Shapiro, P. Smyth, R. Uthurusamy, From Data Mining to KnowledgeDiscovery : An OverviewAdvances in KnowledgeDiscovery and Data Mining, janvier 34 AAAI Press MIT Press, 1996.

[3] D. Salerno, A.Martin. L'extraction des connaissances au Knowledge Management, Dominique Crié. Dans Revue française de gestion 2003/5 (no 146), pages 59 à 79

[4] Margaret H. Dunham «Data Mining Introductory and Advanced Topics», Prentice Hall, 2003.

[5] ZERF Nadjet « *Gestion des connaissances dans le domaine médical* », Mémoire de magister Université Saad Dahlab Blida, 2010.

[6] Zighed&Rakotomalala, « Extraction des Connaissances à partir des Données (ECD) », in Techniques de l'Ingénieur, 2002.

[7] Alsagheer R.H.A., Alharan A. F. H., Al-Haboob A.S.A. Popular Decision Tree Algorithms of Data Mining Techniques: A Review. International Journal of Computer Science and Mobile Computing,6(6) : 133-142; 2017.

[8]MacQueen, J. B. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability (pp. 281–297). Berkeley, CA: University of California Press,1967.

[9]Maitrise, P. De. *Data Mining avec Weka*, 2014.

[10]Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 657–668. https://doi.org/10.1109/TPAMI.2005.95, 2005.

[11] Morgan J., Sonquist J.A. Problems in the Analysis of Survey Data, and a Proposal, Journal of the American Statistical Association, 58:415-435, 1963.

[12]Caron, S. Une introduction aux arbres de décision. *Tokyo University*, 9, 2011.

[13] John F. Roddick, Peter Fule ,Warwick J. Graco. Exploratory Medical Knowledge Discovery : Experiences and Issues School of Informatics and Engineering Flinders University of South Australia GPO Box 2100, Adelaide 5001 South Australia, 2003.

[14] John F. Roddick, Peter Fule ,Warwick J. Graco. Exploratory Medical Knowledge Discovery : Experiences and Issues School of Informatics and Engineering Flinders University of South Australia GPO Box 2100, Adelaide 5001 South Australia, 2003.

[15] M.Bouzeghoub., D.Kostadinov. « Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils ». Laboratoire PRiSM, Université de Versailles 45, avenue des Etats-Unis, 78035 Versailles, 2005.

[16] Hamdi Y., Khrouf K., Feki D., Personnalisation dans les entrepôts de documents. Atelier des Systèmes Décisionnels, ASD Avril 2012, Blida, Algérie, 2012.

[17] Kostadinov D., Personnalisation de l'information et gestion des profils utilisateurs, Rapport de DEA,Université de Versailles, France, 2003.

[18] W. Janowski, A. Sarner. Five Opportunities for Personalization. Gartner Group, pp. 1, 05/2001

[19] Bouaissa D. Un modèle du décideur pour l'aide a la spécification de ses besoins.USD BLIDA, Algerie, 2012.

[21] Mobasher B., Data mining for personalization, In The Adaptive Web: Methods And Strategies Of Web Personalization. Brusilovsky P. and Kobsa A. and Nejdl W. (Ed.), p. 90-105, 2007

[21] Mobasher B., Web Usage mining and personalization, In Practical Handbook Of Internet Computing. Singh Munindar P. (Ed.), 2005.

[22] Han J., Kamber M., Data mining: concepts and techniques, Jim Gray (Ed.). Morgan Kaufmann, 2006.

[23] Chan. P., Constructing Web User Profiles: A Non-invasive Learning Approach, In Web Usage Analysis and User Profiling, LNAI 1836, Springer-Verlag, p. 39-55, 2000.

[24] Eirinaki M., Vazirgiannis M., Varlamis I ., Sewep: using site semantics and a taxonomy to enhance the web personalization process, In Proceedings Of The Ninth AcmSigkdd International Conference On Knowledge Discovery And Data Mining Washington Dc USA, p. 99-108, 2003.

[25] Ferguson S., BEAGLE: A genetic algorithm for Information Filter Profile Creation, Technical Report CS-692, University of Alabama, 1995.

[26] Song R.,Chen E., Zhao M., SVM Based Automatic User Profile Construction for Personalized Search, In Proceedings of the International Conference on Intelligent Computing ICIC , China, p. 475-484, 2005.

[27] Jin X., Zhou Y., Mobasher B., A unified approach to personalization based on probabilistic latent semantic models of web usage and content, In Proceedings Of The Aaai 2004 Workshop On Semantic Web Personalization (Swp'04). 2004.

[28] Michlmayr E., Cayzer S., Learning User Profiles from Tagging Data and Leveraging them for Personal(ized) Information Access,WWW2007, Banff, Canada, p. ,2007

[29] Dorigo M., Di Caro G., Sampls M., Ant algorithms, 3rd International Workshop on Ants Algorithms (ANTS), Brussels, Belgium, 2000.

[30] Mushtaq N., Werner P., Tolle K., Zicari R., Building and Evaluating Non-Obvious User Profiles for Visitors of Web Site, In Proceedings of the IEEE International Conference on E-Commerce Technology (CEC'04), p. 9-15, 2004.

[31] Lieberman H., Letizia: An Agent That Assists Web Browsing, In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, p. 924-929, 1995.

[32] Sugiyama K., Hatano K., Yoshikawa M., Adaptive web search based on user profile constructed without any effort from users, In Proceeding of the 13th International World Wide Web Conferences (WWW), New York, USA, p. 675-684, 2004.

[33] Eisenstein, J., Puerta A., *Adaptation in Automated User-Interface Design*, In Proceedingsof the International Conference on Intelligent User Interfaces, LA, USA, p. 74-81, 2000

[34] Shearin S., LiebermanH, *Intelligent Profiling by Example*, In Proceedings of the 2001International Conference on Intelligent User Interfaces, Santa Fe, USA, p. 145 151, 2001

[35] Boulkrinat, S. : Modélisation hybride du profil utilisateur pour un système de filtrage d'informations sur le web. Thèse de Magister. ESI. 2007.

[36] Ciro S., Newton V., *Use Reformulated Profile in Information Filtering*, In Proceedings ofthe AAAI Workshop on Semantic Web Personalization, San Jose, California, 2004.

[37] Bradley K., Rafter R., Smyth B., *Case-Based User Profiling for Content Personalisation*,In Proceedings of the International Conference on Adaptive Hypermedia and AdaptiveWeb-based Systems, Trento, Italy, p. 62-72, 2000.

[38] Shearin S., Lieberman H., *Intelligent Profiling by Example*, In Proceedings of theInternational Conference on Intelligent User Interfaces, USA, p. 145-151, 2001.

[39] Ferreira J., Silva A., *MySDI: A Generic Architecture to Develop SDI PersonalisedServices*, In Proceedings of the 3rd International Conference on Enterprise InformationSystems, Setubal, Portugal, p. 262-270, 2001.

[40] Zemirli N., Tamine-Lechani L., Boughanem M., *Présentation et évaluation d'un modèle d'accès personnalisé à l'information basé sur les diagrammes d'influence*, Dans les Actes du XXVème congrès INFORSID, Perros-Guirec, France, p. 89-104, 2007.

[41] Kießling W., *Foundations of Preferences in Database Systems*, In Proceedings of the28th Conference on Very Large Data Bases, Hong Kong, China, p. 311-322, 2002.

[42] Kießling W., *Preference Queries with SV-Semantics*, In Proc. of the 11th InternationalConference on Management of Data (COMMAD 2005), Goa, India, p. 15-20, 2005.

[43] Koutrika G., Ioannidis Y. E., *Personalization of Queries in Database Systems*, InProceedings of the 20th International Conference on Data Engineering, Boston,Massachusetts, USA, p. 597-608, 2004.

[44] Koutrika G., Ioannidis Y. E., *Constrained Optimalities in Query Personalization*, InProceedings of the ACM SIGMOD, Baltimore, Maryland, USA, p. 73-84, 2005.

[45] Chomicki J., *Querying with Intrinsic Preferences*, In Proceeding of the 8th InternationalConference on Extending Database Technology (EDBT), Prague, Czech Republic, p. 34-51, 2002.

[46] Borzsonyi S., Kossmann D., Stocker K., *The Skyline Operator*, In Proceedings of theIEEE Conference on Data Engineering (ICDE), Heidelberg, Germany, p. 421-430, 2001.

[47] Lacroix M., Lavency P., *Preference: Putting More Knowledge into Queries*, In Proceeding of the 13th Conference on Very Large Data Bases, Brighton, England, p. 217-225, 1987.

[48] Rocacher D., Liétard L., *Préférences et quantités dans le cadre de l'interrogationflexible: sur la prise en compte d'expressions quantifiées*, Dans les actes des 22e JournéesBases de Données Avancées (BDA), Lille, France, 2006.

[49] Shearin S., LiebermanH, Intelligent Profiling by Example, In: Proceedings of the 2001 International Conference on Intelligent User Interfaces, Santa Fe, USA, January 2001

[50] Platform for Privacy Preferences Project, at http://www.w3.org/P3P/

[51] Bouaka,N : Développement d'un modèle pour l'explicitation d'un problème décisionnel: un outil d'aide à la décision dans le contexte d'intelligence économique. Thèse de doctorat. Labo Loria .2003.

[52] Ansaf Salleb, Recherche de motifs fréquents pour l'extraction de règles d'association et de caractérisation, doctorat, Orléans, 2003.

[53] Abdelhak Mansoul, fouille de données biologiques : étude comparative et expérimentation, magister, oran, 2010.

[54] Hassane Hilali, application de la classification textuelle pour l'extraction des règles d'association maximales, université du québec, 2009.

[55] Allia Mohamed Rachid, BOUADI Tassadit, El MOUTAOUKIL Sami, et KEIRA Mamadou, Fouille de données : Règles séquentielles, master.

[56] Ansaf Salleb. Recherche de motifs fréquents pour l'extraction de règles d'association et de caractérisation, doctorat, Orléans, 2003.

[57] Aggarwal CC, Prasad VVV A tree projection algorithm for generation of frequent JParallel Distrib Comput 61(3):350–371, 2001.

[58] Borgelt, C. An implementation of the FP-Growth algorithm. Chicago, Illinois. August 21, 2005.

[59] Han, J., Pei, J., Yin, Y. MiningFrequent Patterns without Candidate Generation Proc. 2000 ACMSIGMOD Int. Conf. on Management of Data(SIGMOD'00),Dallas,TX.

[60] AgrawalR, SrikantR. Fast algorithms for mining association rules. In: Paper presented at the proceedings of the 20th international conference on very large data bases, Santiago, 1994.

[61] Mishra,L., Choubey,A. Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data.Shri Shankaracharya College of Engineering and Technology, Bhilai C.G. India, 2012.

[62] Agarwal, R., Srikant, R. Fast Algorithms for Mining Association Rules. http://www.columbia.edu/~rd2537/docu/apriori(abstract).pdf, 2018.

[63] Chee, C.-H., Jaafar, J., Aziz, IA, Hasan, MH et Yeoh, W. Algorithmes pour l'extraction d'éléments fréquents: une revue de la littérature. Examen de l'intelligence artificielle. doi:10.1007 / s10462-018-9629-z, 2018.

[64] Javeed,MZ. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12(3):372–390, 2000.

[65] Javeed,MZ., Parthasarathy,S. &Li,W. A Localized Algorithm for Parallel Association Mining Department of Computer Science, University of Rochester, Rochester, NY 14627, 1997.

[66] Han, J., Kamber, M., Pei J. Data mining concepts and techniques. Elsevier, Atlanta, 2012.

[67] Aggarwal CC, Bhuiyan MA, Hasan MA. Frequent Pattern Mining algorithms: a survey. In: Aggarwal CC, Han J (eds) Frequent Pattern Mining. Springer, Basel, pp 1964, 2014.

[68] Aggarwal CC, Prasad VVV . A tree projection algorithm for generation offrequent itemsets. J Parallel Distrib Comput 61(3):350–371, 2001.

[69]Heaton, J. Comparing dataset characteristics that favor the Apriori, Eclat or FP Growth frequent itemset mining algorithms. SoutheastCon , 2016.

[70] S. S. Kadam and S. S.Deshmukh. "Eclat Algorithm for FIM on CPU-GPU coperative &parallel environment," *IOSR J. Comput. Eng.*, vol. 16, no. 2, pp. 88–96, 2014

[71] Alagukumar, S., & Lawrance, R. *A Selective Analysis of Microarray Data Using Association Rule Mining. Procedia Computer Science, 47, 3–12*, 2015.

[72]http://www.16-types.fr/modele-MBTI-0-introduction.html

[73] Fuller, Mark B., and Michael E. Porter. "Coalitions and global strategy from." *Competition in global industries* 315, 1986

[74]https://www.java.com/fr/download/faq/java8.xml?printFriendly=true.

[75]https://netbeans.org/index_fr.html.

[76]https://sql.sh/sgbd/sql-server.

[77]https://www.cs.waikato.ac.nz/ml/weka/.