CENTRE D'ETUDE DOCTORALES

Sciences de l'ingénieur

Thèse

présentée à la Faculté des Sciences et Techniques de Marrakech
pour obtenir le grade de :

Docteur

Formation Doctorale

Spécialité

Informatique

Fouille de données et analyse de qualité des règles d'association dans les bases de données massives : Application dans le domaine de la sécurité routière

---

par :

Addi Ait-Mlouk

(Master : Ingénierie des systèmes d'information)

Soutenue le 10 Mars 2018 devant la commission d'examen :

Abdellah AIT OUAHMAN	PES	ENSA-Marrakech	Président du jury
Abdelilah MAACH	PES	EMI-Rabat	Examinateur
Fatima GHARNATI	PES	FSSM	Examinatrice
Mohamed EL ADNANI	PES	FSSM	Examinateur
Rachid LATIF	PES	ENSA-Agadir	Examinateur

Dédicace

A la mémoire de mon père
Que Dieu ait son âme, implorant le Très-Haut de les accueillir
en son vaste paradis parmi ceux qu'il gratifie
de ses commensurables bienfaits.

A ma mère
qui m'a tout donné, m'a inculqué des valeurs telles que
l'éducation, la dignité, le respect, le travail...
Merci de m'avoir donné toutes les chances pour réussir.
Que ce travail soit pour vous le témoignage de mon infini amour.

A mes sœurs et mes frères
Merci pour votre amour et vos encouragements.

A tous ceux qui me sont chers (es).

Avant-propos

— **Nom et Prénom de l'auteur :** AIT-MLOUK Addi

— **Intitulé de travail :** Fouille de données et analyse de qualité des règles d'association dans les bases de données massives : Application dans le domaine de la sécurité routière.

— **Encadrant :**

1. Nom, Prénom et grade : GHARNATI Fatima, Enseignante chercheuse
2. Laboratoire et Institution : Gestion Intelligente des Énergies et Systèmes d'Information (GIESI), Département de physique, Faculté des Sciences Semlalia, Université Cadi Ayyad, Marrakech.

— **Co-encadrant :**

1. Nom, Prénom et grade : AGOUTI Tarik, Enseignant chercheur
2. Laboratoire et Institution : Laboratoire d'Ingénierie des Systèmes Informatiques, Département d'Informatique, Faculté des Sciences Semlalia, Université Cadi Ayyad, Marrakech.

— **Lieux (Laboratoire, Institution) :** Gestion Intelligente des Énergies et Systèmes d'Information (GIESI). Faculté des Sciences Semlalia, Université Cadi Ayyad, Marrakech.

— **Période de réalisation du travail de thèse :** Janvier 2014 - Décembre 2017

— **Rapporteurs autres que l'encadrant :**

1. — Nom et Prénom : EL ADNANI Mohamed
   — Grade : Professeur d'Enseignement Supérieur (PES)
   — Institution : Faculté des Sciences Semlalia, Université Cadi Ayyad, Marrakech.

2. — Nom et Prénom : LATIF Rachid
   — Grade : Professeur d'Enseignement Supérieur (PES)
   — Institution : Ecole Nationale des Sciences Appliquées, Université Ibn Zohr, Agadir.

3. — Nom et Prénom : MAACH Abdelilah
   — Grade : Professeur d'Enseignement Supérieur (PES)
   — Institution : Ecole Mohammadia d'Ingénieurs, Université Mohamed V, Rabat.

Ce travail a donné lieu aux résultats suivants(Publications, Communications,...) :

Articles dans des journaux internationaux :

1. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. (2017)An improved approach for association rules mining using multi-criteria decision support system : A case study in road safety. European Transport Research Review, Vol.9, No.3, pp.1–13, DOI : 10.1007/s12544-017-0257-5.
2. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. (2017) Mining and prioritization of association rules for Big Data : Multi-criteria decision analysis approach. Journal of Big Data, Vol.4, No.1, pp.1-21, DOI : 10.1186/s40537-017-0105-4.
3. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. (2017) Application of big data analysis with decision Tree for road accident. Indian Journal of Science and Technology, Vol. 10, No. 29, pp. 1-10, DOI : 10.17485/ijst/2017/v10i29/117325.
4. **Ait-Mlouk, A.**, Kamsa, I., Gharnati, F., Agouti, T. (2017) Intelligent transport system for road safety based data mining approach. International Journal of Control and Automation, Vol.10, No.8, pp.13-22.
5. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. (2016) Multi-criteria decisional approach for extracting relevant association rules. Int. J. of Computational Science and Engineering, Vol. 15, No.3/4, pp. 188–200.
6. **Ait-Mlouk, A.**, Gharnati, F. Agouti, T. (2016) Multi-Agent Based Modeling for Extracting Relevant Association Rules Using a Multi-Criteria Analysis Approach. Vietnam Journal of Computer Science, Vol.3, N.4, pp 235–245, doi :10.1007/s40595-016-0070-4.

Communications dans des conférences internationales :

1. **Ait-Mlouk, A.**, Gharnati, F., Agouti,T. Application of big data analysis with decision Tree for road accident. 3rd International Conference on Green Computing and Engineering Technologies, August 8-10, 2017. Killaloe, County Clare, Ireland.
2. **Ait-Mlouk, A.**,Kamsa, I., Gharnati, F., Agouti,T. Intelligent transport system for road safety based data mining approach. 3rd International Conference on Green Computing and Engineering Technologies, August 8-10, 2017. Killaloe, County Clare, Ireland.
3. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. An Approach Based on MapReduce and Decision Tree to Improve Road Safety in Morocco. International symposium on data engineering and information systems (DEIS'2017), Mai 19-05, 2017, Marrakech, Maroc.
4. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. Mining Traffic Accident from Big Data : The case study of Morocco. The First International Conference of High Innovation in Computer Science (ICHICS'2016), June 01-03, 2016, Kenitra, Maroc.
5. Formation Science des Données. La Conférence Internationale Francophone AAFD SFC, Mai 22-26, 2016, Marrakech, Maroc.

6. **Ait-Mlouk, A.**, Gharnati, F., Agouti,T. An approach based on association rules mining to improve road safety in Morocco. International Conference on Information Technology for Organizations Development (IT4OD), Fez, 2016, pp. 1-6.doi : 10.1109/IT4OD.2016.7479311
7. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. Association Rules Mining based On Electre Tri Method. The 4th International Conference on Software Engineering and New Technologies (ICSENT'2015), December, 20-24, 2015, Istanbul, Turkey.
8. **Ait-Mlouk, A.**, Gharnati, F., Agouti,T., and Derbali B. A choice of relevant association rules based on multi-criteria analysis approach. 5th International Conference on Information and Communication Technology and Accessibility (ICTA), Marrakech, 2015, pp. 1-6.
9. Aznaoui, H., **Ait-Mlouk, A.**, Classification of routing Protocol in WSNs. The 4th International Conference on Software Engineering and New Technologies (ICSENT-2015), December, 20-24, 2015, Istanbul, Turkey.
10. H. Aznaoui, S. Raghay, L. Aziz and **Ait-Mlouk, A.** A comparative study of routing protocols in WSN. 5th International Conference on Information and Communication Technology and Accessibility (ICTA), Marrakech, 2015, pp. 1-6.
11. **Ait-Mlouk, A.**, Gharnati, F., Agouti,T. Comparative survey of association rule mining algorithms based on multiple-criteria decision analysis approach. 3rd International Conference on Control, Engineering and Information Technology (CEIT), Tlemcen, 2015, pp. 1-6. doi : 10.1109/CEIT.2015.7233078
12. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. Evaluation of association rules extraction algorithms. International Conference on Networked Systems, May 13-15, 2015. Agadir, Maroc.

Communications dans des conférences nationales :

1. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. Un choix des règles d'association basé sur la méthode Electre Tri. Rencontre National en Informatique : Outils et Application (RNIOA'2015), Avril, 23-24. Er-Rachidia, Maroc.
2. **Ait-Mlouk, A.**, Gharnati, F., Agouti, T. On selecting interestingness measures for association rules extraction based on electre tri method. 1er forum de la recherche scientifique, Jun, 2015. Marrakech, Maroc.

Remerciements

Je remercie tout d'abord Dieu tout puissant de m'avoir donné le courage, la force et la patience d'achever ce modeste travail. Je remercie Mme Fatima Gharnati, d'avoir acceptée d'être la directrice de cette thèse, pour ses encouragements et pour son aide qu'elle m'a apporté pendant tout ce travail.

Tout particulièrement, je tiens à remercier M. Tarik Agouti, mon encadrant de thèse, qui m'a encadré durant ces longues années. Nous avons eu ensemble des discussions très enrichissantes qui m'ont orienté dans mes travaux de recherche. Je le remercie pour tous ses conseils avisés et pour sa disponibilité.

Je tiens à remercier les membres du jury M. El Adnani Mohamed, M. Latif Rachid, et M. Maach Abdelilah d'avoir accepté d'être rapporteurs de cette thèse. Je remercie également M. Ait Ouahman Abdellah, pour l'honneur qu'il me fait d'avoir accepté d'être président du jury.
Je remercie également toute l'équipe du laboratoire GIESI de l'Université Cadi Ayyad, en particulier M. Lamchich Moulay Tahar le directeur de Laboratoire pour sa disponibilité et sa confiance.

Je ne saurais trouver les convenables mots pour remercier ma chère famille pour son soutien précieux moral et affectif durant mes longues années d'études, et sans elle, ce travail n'aurait pas pu avoir lieu.

Mes remerciements vont aussi à tous mes chers amis, je tiens à leurs exprimer mes amitiés et mes remerciements. Un grand merci à R. Es-sadki , Z. Erraji, A. Kadiri et Fatim Ezahra Elmazouari. Je ne peux finir sans avoir une pensée pour H. Aznaoui, et M. Ait Mlouk, I. Kamsa, B. Marzak et R. Mouachi en souvenir de notre collaboration et amitié.

Enfin, je tiens à remercier toutes les personnes qui ont croisé ma route pendant ces années, que ce soit dans le côté professionnel ou pédagogique.

Résumé

L'extraction de connaissances dans les bases de données (ECD), également appelée fouille de données, « désigne le processus non trivial d'extraction d'information implicite, précédemment inconnue et potentiellement utile ». La fouille de données est un domaine de recherche en plein essor visant à exploiter les grandes quantités de données collectées chaque jour dans divers domaines d'application de l'informatique. Ce domaine pluridisciplinaire est issu de l'intelligence artificielle, des statistiques et des bases de données.

Dans ce travail, nous nous intéressons au problème de l'extraction des règles d'association en introduisant de nouveaux algorithmes et approches d'aide à la décision multicritère. D'une manière générale, une règle d'association est une implication conditionnelle entre des ensembles d'attributs binaires appelés items. L'extraction de telles règles est décomposée en deux étapes principales, à savoir l'extraction des itemsets fréquents et la génération des règles d'association à partir de ceux-ci. Dans la majorité des approches existantes dans la littérature, l'extraction des règles d'association présente trois difficultés majeurs, à savoir ; la qualité des règles extraites, l'aspect spatiale de données et le temps de réponse des algorithmes d'extraction.

Pour surmonter ces difficultés, nous proposons dans cette thèse l'intégration de l'analyse multicritère au processus d'extraction des règles d'association pour l'analyse de la qualité. Ensuite, afin de prendre en considération l'aspect spatiale de données, et plus précisément l'estimation des distances métriques, nous avons proposé l'utilisation de la logique floue. Nous avons proposé également une intégration de l'algorithme FP-growth dans un environnement du Big Data pour l'extraction des règles d'association dans les bases de données massives.

En plus, en vue de tester concrètement l'apport des solutions proposées, nous avons conçu et développé un prototype logiciel constitué de trois interfaces interactives. La première intitulée *interface ARM*, est une interface web dédiée à l'extraction des règles d'association. La deuxième interface, intitulée *interface MCDA*, est une interface web dédiée à l'analyse de qualité des règles d'association extraites. Quant à la dernière, intitulée *Time Series Forcasting*, est une interface web dédiée à la prédiction des accidents routières en termes du nombre de blessures et décès. Ces interfaces interactives d'exploration de données ont été développées en utilisant le langage R et rshiny. En fin, les expérimentations menées sur quelques bases de données relatives aux accidents routières au Maroc montrent la faisabilité notable de nos contributions.

Mots clés : *Fouille de données, itemsets fréquents, règle d'association, règle d'association spatiale, analyse multicritères, mesures de qualité, la logique floue, Big Data.*

Abstract

Knowledge discovery in databases (KDD), often called Data Mining. « Is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data ». Data mining is an active field of research aiming to exploit the vast amounts of data collected every day in various fields of computer science applications. This multidisciplinary field comes from artificial intelligence, statistics, and databases.

In this thesis, we are interested in the problem of extracting association rules by introducing new algorithms and approaches. In general, an association rule is a conditional implication between sets of binary attributes called items. The extraction of such rules is composed of two main steps which are the extraction of frequent itemsets and the generation of association rules from them. The complexity of each of these steps is exponential : the number of frequent itemsets is exponential, and the number of association rules extracted can be very high, due to the quality measures used. In the literature, the extraction of the association rules is composed into two main difficulties, the response time and the memory space.

To overcome these difficulties, we propose in this thesis three main contributions respectively allowing the extraction of relevant association rules, the integration of the spatial component into the extraction process, and mining relevant association rules from big data. In the first contribution, we propose an extraction approach of the relevant association rules based on multicriteria decision analysis. Then, in the second contribution, we propose an efficient algorithm for extracting spatial predicates from which frequent sets of items and spatial association rules can be generated based on the preparation of the spatial context and the fuzzy set theory. We also proposed in the third contribution a distributed algorithm for the extraction of association rules from Big Data. Using these contributions, we were able to extract the relevant association rules and reduce the execution time and the memory space. In addition, in order to test concretely the contribution of the proposed solutions, we designed and developed a software prototype consisting of three interfaces. The first entitled *ARM interface*, is an interactive web interface dedicated to the extraction of association rules. The second interface, entitled *MCDA interface*, it is an interactive web interface dedicated to the evaluation and extraction of relevant association rules. For the last one, entitled *Time Series Forcasting*, it is an interactive web interface dedicated to the prediction of road accidents. Moreover, interactive and user-friendly interfaces have been developed by using R language and rshiny. Finally, the experiments conducted on some databases on road accidents in Morocco show the significant feasibility of our contributions.

Keywords : *Data mining, frequent itemsets, association rules, spatial association rules, multi-criteria analysis, quality measurements, fuzzy logic, Big Data.*

Notations

Description des symboles les plus utilisés dans cette thèse.

$\widehat{x}$	Estimateur de x
$P(A)$	Probabilité de l'ensemble A
$[a, b]$	Intervalle des valeurs comprises entre a et b
$A \backslash B$	Complémentaire de l'ensemble B dans l'ensemble A
$\preceq$	Inférieur ou égal
$\succeq$	Supérieur ou égal
$\forall$	Qu'il que soit
$\exists$	Il existe
∞	Infini
∇	Opérateur de la fermeture de la connexion de Galois
$\|x\|$	Valeur absolue du nombre x

Théorie des ensembles :

$\cup$	Union
$\bigcup$	Union généralisée
$\cap$	Intersection
$\bigcap$	Intersection généralisée
$\subseteq$	Inclusion
$\subset$	Inclusion sans égalité
$\in$	Appartenance
$\emptyset$	Ensemble vide
$\|E\|$	Cardinalité de l'ensemble E

Logique des prédicats :

$\vee$	Disjonction
$\wedge$	Conjonction
$\neq$	Négation
$\Rightarrow$	Implication
$\Leftrightarrow$	Equivalence

Abréviations et sigles

Description des abréviations et sigles utilisés dans cette thèse.

ECD	Extraction de Connaissances à partir de Données
ECDS	Extraction de Connaissances à partir de Données Spatiales
KDD	Knowledge Discovery from Databases
DM	Data Mining
FD	Fouille de données
FDS	Fouille de données Spatiales
DMS	Data Mining Spatial
GPS	Global Positioning System
RFID	Radio Frequency IDentification
ADMC	Aide à la Décision MultiCritère
FP-growth	Frequent Pattern growth
FP-tree	Frequent Pattern tree
ACF	Analyse des Concepts Formels
FM	Fréquents Maximaux
SIG	Système d'Information Géographique
MBR	Minimum Bounding Rectangle
GIS	Geographical Information System
ARGIS	Association Rules in Geographical Information System
AMC	Analyse Multicritère
ETL	Extraction Transformation Loading
METL	Ministère de l'Équipement, des Transports et de la Logistique
LHS	Left Hand Side
RHS	Right Hand Side
TEF	Théorie des Ensembles Flous
NFT	Nombres Flous Triangulaires
OMS	Organisation Mondiale de la Santé
CNPAC	Comité National de Prévention des Accidents de la Circulation
SOLAP	Spatial On-Line Analytical Processing
CSV	Comma Separated Values
HDFS	Hadoop Distributed File System
RDD	Resilient Distributed Dataset
PFP-growth	Parallel Frequents Patterns growth

Sommaire

Liste des figures

Liste des tableaux

Liste des figures

Liste des tableaux

Introduction générale

«Ne restez pas indéfénimment sur la route qui ne mène qu'à des endroits connus, abandonnez parfois les sentiers battus et entrez dans la forêt, vous découvrirez certainement quelque chose que vous n'avez jamais vu, bien sur ce ne sera qu'une petite chose, mais prêtez y attention, suivez la, explorez la, une découverte en amènera une autre, et avant même de vous rendre compte, vous aurez mis à jour une idée intéressante.»

Alexander Graham Bell

L'objectif de ce chapitre est de présenter le contexte général de cette thèse, les différentes problématiques d'extraction des règles d'association, et les objectifs de cette recherche.

Sommaire

0.1 Introduction

La fouille de données, dite Data Mining, « désigne le processus non trivial d'extraction de connaissances implicites, précédemment inconnues et potentiellement utiles à partir de données ». L'idée de base de la fouille de données est d'extraire les connaissances cachées à partir de données disponibles. Ces connaissances peuvent être sous forme de modèles, règles de décision, et concepts, qui sont utiles et compréhensibles. Les travaux de recherche liés à ce domaine sont motivés généralement par l'évolution très rapide des systèmes de collecte de données (GPS, RFID, Code à barres, satellite, etc.) et les technologies de stockage (diminution des coûts et augmentation de la capacité des disques durs).

L'extraction de connaissances à partir des bases de données (ECD) est un processus itératif constitué de plusieurs étapes allant de la sélection et le prétraitement de données jusqu'à la visualisation et l'interprétation des résultats, en passant par la phase d'apprentissage. L'intérêt d'appliquer les techniques d'extraction de connaissances est de découvrir les connaissances utiles, cachées à partir de données et les présenter comme élément valide pour l'aide à la décision. Les règles d'association sont l'une des formes les plus puissantes pour extraire les connaissances utiles à partir de données, elles ont été initiées par Agrawal [3, 4, 5] pour l'analyse des bases de données transactionnelles. Ces règles présentent des relations significatives entre objets selon leurs caractéristiques, et sont présentées sous la forme de motif : antécédent $\rightarrow$ conséquent. Dans cette forme, les deux parties d'une règle (c'est-à-dire l'antécédent et la conséquence) sont composées de nombreux éléments sous la forme $terme\ 1 \rightarrow terme\ 2 ... terme\ n\ (\alpha, \beta)$, soit par exemple la règle : Économie $\rightarrow$ Microéconomie (55%, 70%), une telle relation signifie que le terme Économie se trouve dans 55% des documents disponibles et que dans 70% des cas il s'y trouve en compagnie avec le terme Macroéconomie.

L'extraction des règles d'association se fait généralement en deux étapes ; l'extraction des itemsets fréquents et ensuite la génération des règles d'association à partir de ces itemsets. Dans la littérature, la majorité des algorithmes proposés pour l'extraction des règles d'association engendrent un temps de réponse très élevé et génèrent un très grand nombre de règles d'association. Pour réduire le nombre des règles d'association issues d'un contexte de la fouille de données, on utilise des critères communément appelés mesures de qualités. Plusieurs mesures de qualité [6] ont été proposées dans la littérature ce nombre important des mesures de qualité engendre de nouveaux problèmes, entre autres, le choix des mesures de qualité les plus convenables pour l'évaluation efficace des règles afin de retenir celles qui sont réellement pertinentes. De plus, la maturité des systèmes d'information géographique produit une énorme quantité de données spatiales collectées par les GPS, la télédétection et d'autres outils de collecte de données spatiales ce qui apporte de nouveaux défis pour développer des outils de la découverte des connaissances pertinentes à partir de grandes bases de données spatiales. En effet, certains algorithmes et méthodes d'exploration de données spatiales présentés dans la littérature portaient sur des données numériques, et ne peuvent pas être un moyen très efficace d'extraction des règles d'association spatiales [7, 8, 9]. Ainsi, il est indispensable de concevoir et développer des algorithmes permettant de tirer pleinement profit de la richesse informationnelle que recelé ces données, cela connue sous le nom de fouille de données spatiales [10].
De plus, l'analyse des bases de données massives impose un temps de réponse très élevé

pour les algorithmes itératifs, dans ce contexte, il est nécessaire de développer également des algorithmes distribués d'extraction des règles d'association pertinentes.

0.2 Problématique

Les méthodes proposées dans la littérature pour l'extraction des règles d'association classiques et spatiales basées sur l'algorithme de référence Apriori [4] pose quatre problèmes majeurs :

1. La tâche d'extraction des itemsets fréquents impose plusieurs parcours de base de données ce qui engendre un temps de réponse très élevé ;
2. L'utilité et la pertinence des règles d'association extraites par les algorithmes de génération constituent un problème primordial. En fait, les jeux de données réels conduisent à un très grand nombre de règles d'association, ce qui ne permet pas aux décideurs de faire eux-mêmes la sélection des règles d'association les plus pertinentes. La recherche des meilleures parmi le vaste ensemble de règles extraites impose la recherche et l'utilisation des bonnes mesures de qualités ;
3. L'algorithme ARGIS proposé par Salleb et Margoubi [7, 9] relatif à la découverte des règles d'association spatiales considère seulement deux couches thématiques à chaque comparaison. Autrement dit, il n'est pas possible d'extraire des règles regroupant plus de deux couches thématiques ;
4. Avec la maturité des systèmes de collecte de données (RFID, Réseaux sociaux, traitement des données en temps réel, et Internet des objets). Les entreprises ont investi en technologies de stockage afin de collecter et gérer leurs données. Pour tirer profit de ces données massives, il est indispensable de proposer des méthodes d'extraction des règles d'association en prenant en compte le temps de réponse et le nombre très élevé des règles d'association extraites.

Pour la formulation de ces problèmes, nous proposons d'étudier le domaine de la sécurité routière. Selon l'Organisation Mondiale de la Santé (OMS) [11], les accidents de la route sont la première cause de décès chez les jeunes âgés de 15 à 29 ans chaque année, près de 1.25 million de personnes décèdent dans un accident de la route et 20 à 50 millions d'autres sont blessés, parfois même handicapés. Le Ministère de l'Équipement, du Transport et de la Logistique du Maroc [12] a procédé aux statistiques provisoires nationales du mois de décembre 2016 comparées à ceux du mois de décembre 2015 se présente comme suit :

— 7117 accidents, soit : +3,46% ; ;
— 250 accidents mortels, soit : -7,06% ;
— 6867 accidents non mortels, soit : +3,89% ;
— 303 tués soient : +0,33% ;
— 682 blessés graves, soit : -15,17% ;

— 9181 blessés légers, soit : +1,15%.

Les accidents de la route entraînent des pertes sociaux-économiques considérables. En effet, ils nécessitent d'étudier profondément ce domaine. La fouille de données et l'extraction de la connaissance à partir de données répondent à ce besoin réel de l'analyse des accidents routiers pour tirer profit de la disponibilité croissante de données localisées et de la richesse potentielle en information de ces données. Dans notre domaine d'application, nous nous intéressons à l'analyse des accidents routiers tout en prenant en compte la phase d'apprentissage pour soulever les facteurs liés aux accidents, et ainsi la définition des critères d'interdépendances et de causalité entre les couches thématiques représentant les différentes infrastructures liées au domaine de la sécurité routière (Routes, Territoires, Institutions, etc).

0.3 Objectifs de la recherche

La fouille de données représente l'intégration de différents domaines, comprenant l'apprentissage automatique, l'intelligence artificielle, les bases de données, les statistiques et les théories d'information. L'extraction de la connaissance dans les bases des données est un processus complexe qui couvre de nombreuses étapes interdépendantes.
Certains algorithmes et méthodes d'extraction des règles d'association proposées dans la littérature retournent un très grand nombre de règles d'association dont la plupart sont des règles non utiles redondantes. Nous avons remarqué une perte très importante en termes de temps de réponse et d'espace de stockage au cours du processus de génération des règles d'association. Ainsi, et faisant référence à la problématique suscitée ci-dessus, nous pouvons citer les objectifs de cette thèse suivants :

1. L'amélioration des algorithmes utilisant les règles d'association pour l'extraction de la connaissance en tenant compte de la problématique relative à la pertinence et la qualité des règles d'association ;
2. Surmonter le problème majeur lié à l'intégration de la composante géographique pour l'extraction des règles d'association spatiales en se basant sur la théorie des ensembles flous, au niveau de calcul des distances entre les objets de différentes couches thématiques considérées ;
3. Intégration de l'algorithmes d'extraction des règles d'association FP-growth dans le contexte des bases de données massives (Big Data) pour résoudre le problème de stockage et améliorer le temps de réponse des algorithmes itératifs en se basant sur les calculs distribués ;
4. La formulation et la modélisation du système relatif à notre étude de cas des accidents routiers ;
5. La conception et le développement de prototype proposé pour la sécurité routière.

0.4 Organisation de thèse

Le rapport ci-présent est organisé selon deux grandes parties. La première présente un état de l'art et concepts de base de la fouille de données, l'analyse multicritère et la logique floue. Elle est constituée de deux chapitres nécessaires pour comprendre ultérieurement l'intérêt des nouvelles approches.
En plus de l'introduction générale, le premier chapitre présente les notions utiles considérées comme les fondements mathématiques de l'extraction des règles association, règles d'association spatiale et notion de base des systèmes d'information géographique, y compris les états de l'art sur les notions des itemsets fréquents et les méthodes d'extraction des règles d'association spatiales. Nous présentons aussi les différentes mesures de qualité des règles d'association. Notons que les mesures de qualité sont naturellement utilisées pour capturer les règles utiles et pertinentes à partir d'un contexte de fouille de données. Ce chapitre décrit également les différents algorithmes d'extraction des règles d'association.

Le deuxième chapitre s'occupe de définir les fondements mathématiques de l'analyse multicritères, y compris l'état de l'art des méthodes d'analyse multicritère (AMC). Nous y parlerons par ailleurs de l'intérêt que porte l'analyse multicritère à l'extraction des règles d'association pertinentes. Nous y présentons également les concepts de la logique floue et son apport au processus d'extraction des règles d'association spatiales.

La deuxième partie nommée contributions est constituée de quatre chapitres. En effet, le premier chapitre représente la description du premier objectif que nous nous sommes fixés à savoir l'extraction des règles d'association pertinentes à l'aide de l'analyse multicritère. Le deuxième chapitre représente le deuxième objectif à savoir la proposition d'un algorithme d'extraction des règles d'association spatiales basée sur la logique floue. Le troisième chapitre, représente l'extension de l'approche proposée dans le premier chapitre au contexte du Big Data pour l'extraction des règles d'association dans les bases de données massives. Quant à le dernier, représente l'implémentation d'un prototype logiciel en vue de tester concrètement l'apport des approches proposées en présentant les tests de performance et les résultats obtenus afin de répondre à l'objectif stratégique de la sécurité routière.

Ce rapport se clôture par une conclusion générale dans laquelle, nous établissons le bilan des travaux réalisés et proposons des perspectives possibles pour ces nouvelles approches.

Première partie

État de l'art et concepts de base

Chapitre 1

Extraction des itemsets fréquents et règles d'association

«Those who do not remember the past are condemned to repeat it.»

George Santayana

L'objectif de cette partie est de donner un bref aperçu sur la fouille de données et la fouille de données spatiales. Nous présentons les différents algorithmes d'extraction des règles d'association, ensuite, nous décrivons les concepts de base de la fouille de données.

Sommaire

1.1 Introduction

Au cours des dernières années, les outils d'analyse statistiques traditionnels présentent des difficultés pour la gestion de l'énorme volume de données collectées. En plus, les méthodes statistiques nécessitent une connaissance plus large de données afin de définir des hypothèses principales pour l'analyse. En conséquence, l'analyse devient plus coûteuse en termes de temps et de stockage. Les méthodes classiques deviennent un outil inapproprié et inadéquat pour l'analyse des grandes quantités de données. Cette accumulation d'informations dans les bases de données a motivé le développement d'un nouveau champ de recherche appelé fouille de données ou l'Extraction de Connaissances dans les bases de données (ECD). Ce champ est issu des bases de données, des statistiques, et de l'intelligence artificielle. L'idée de base de la fouille de données est d'extraire des informations implicites, précédemment inconnues et potentiellement utiles à partir d'ensemble de données, de point de vue utilisateur ces informations peuvent être des règles d'association, des concepts, des modèles, etc. Fayyad [1] décrit le processus de l'ECD comme un processus itératif semi-automatique constitué de plusieurs étapes allant de l'objectif d'extraction et la sélection des données jusqu'à la visualisation et l'interprétation des résultats, en passant par la phase de recherche de connaissances. Les différentes étapes de ce processus sont présentées dans la Figure 1.1.

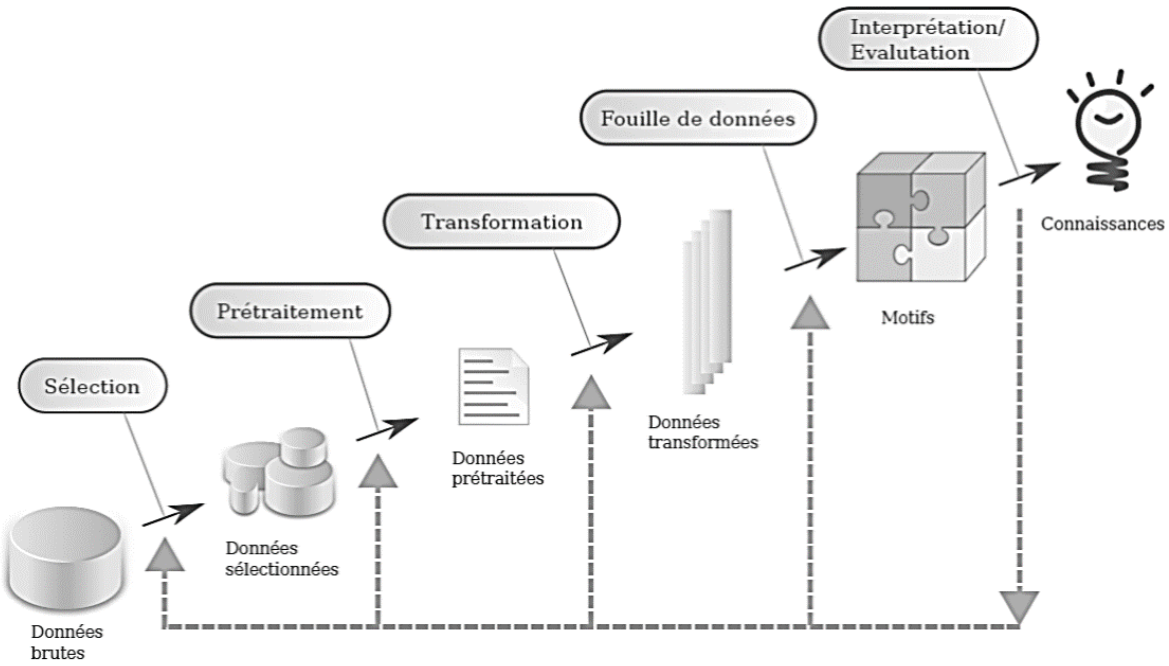

Figure 1.1 – Processus de l'extraction de connaissance dans les bases de données [1]

Les étapes principales du processus de l'ECD sont comme suit :

1. **La sélection** : Une fois l'objectif de l'extraction est fixé, l'étape de base est la sélection des échantillons significatifs de données. Toutes les données brutes ne sont

pas nécessairement pertinentes pour une application des algorithmes de la fouille de données, il est nécessaire de sélectionner un sous-ensemble adapté à l'étude à mener et déterminer la structure générale des données ainsi que les règles utilisées en identifiant les informations exploitables et vérifier leur qualité et leur facilité d'accès.

2. **Le Prétraitement** : Les erreurs de saisie, les champs nuls, et les valeurs manquantes, impose généralement une phase de nettoyage de données, celle-ci a pour objectif de corriger ou de contourner l'inexactitude et les erreurs de données comme suit :
   - Exclure les enregistrements incomplets ;
   - Identifier et traiter les valeurs manquantes, les valeurs erronées ou incertaines et les inconsistances ;
   - Remplacer les données manquantes ;
   - Prédire les valeurs (valeur moyenne des objets similaires, la régression) ;
   - Utiliser l'absence de valeur comme une information ;
3. **Transformation** : En vue d'appliquer un traitement spécifique aux données précédemment sélectionnées, il est nécessaire d'adapter leur structure dans un format approprié à la tâche de la fouille de données choisies dans la troisième étape.
4. **Fouille de données** : Dans cette phase, des méthodes intelligentes sont utilisées afin d'extraire les connaissances utiles à partir de données et les présenter sous une forme synthétique. Lors de cette étape plusieurs techniques peuvent être utilisées à savoir, le clustring, la classification, la régression, les règles d'association, etc.
5. **Interprétation** : Cette étape identifie les modèles intéressants représentant les connaissances, en se basant non seulement sur des mesures d'intérêt, mais aussi sur l'avis de l'expert. Cette évaluation prend généralement une forme graphique ou textuelle et contribue fortement à améliorer la lisibilité et la compréhension des résultats et facilite le partage de la connaissance. Les résultats produits par les algorithmes de fouille de données ne sont pas toujours exploitables directement. En effet, il est utile de définir des nouvelles mesures de qualité afin d'assister le décideur a utiliser les règles d'association les plus pertinentes.

Il existe une distinction précise entre le concept de l'ECD et celui de la fouille de données. En effet, ce dernier n'est qu'une des étapes de découverte de connaissances correspondant à l'extraction de connaissances à partir de données. Cette étape consiste à recouvre uniquement l'extraction de connaissances à partir de données en appliquant les méthodes de découverte. Les méthodes de la fouille de données peuvent être divisées en deux grandes familles : les méthodes descriptives et les méthodes prédictives. En ce qui concerne la première famille (Figure 1.2) , elle est caractérisée par son mode de découverte et d'analyse descriptive de données. La deuxième famille (Figure 1.3) cherche à prédire une donnée particulière.

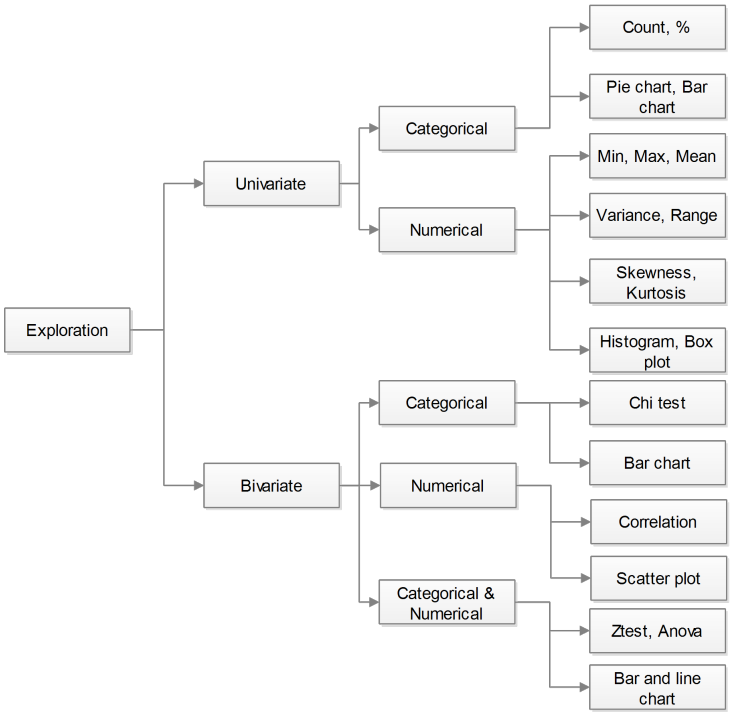

Figure 1.2 – Les techniques descriptives de fouille de données [2]

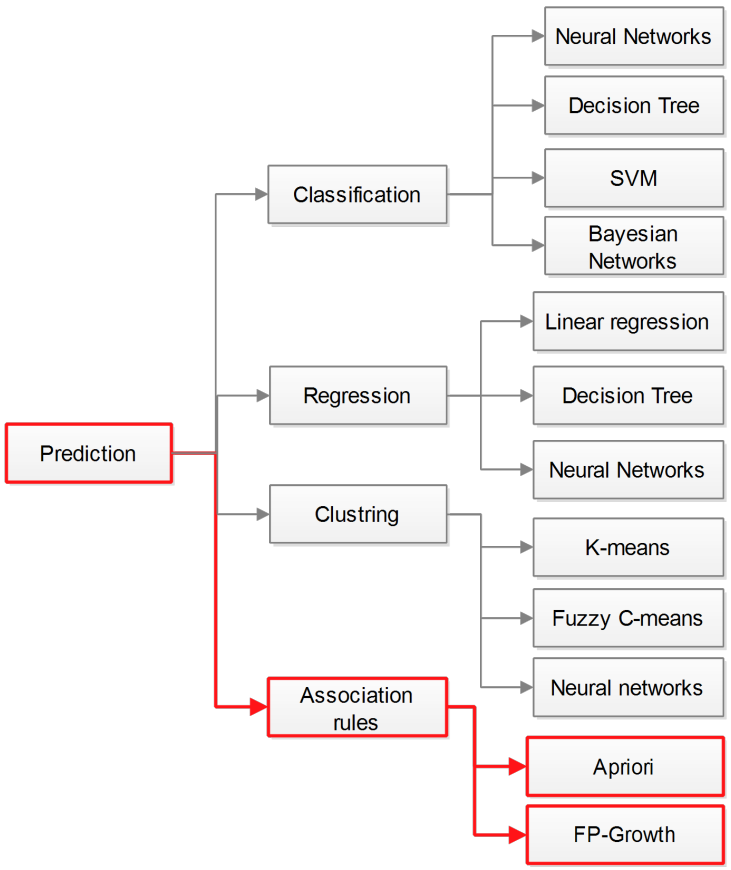

Figure 1.3 – Les techniques prédictives de fouille de données [2]

Parmi les techniques prédictives utilisées en fouille de données, nous pouvons signaler les règles d'association qui est sans doute la tache phare qui a attiré l'attention des chercheurs et pour laquelle beaucoup des travaux ont été effectués. Cette technique a pour vocation d'extraire les règles intelligible et exploitable à partir des grands volumes de données. Ces règles d'association sont des implications de la forme : $a \rightarrow b$ ou a et b sont des items (variables booléennes de la forme attribut = valeur). Une telle règle décrit une corrélation entre l'ensemble d'items dans une base de transactions. Autrement dit, étant donné un ensemble d'items, l'objectif est de découvrir si l'occurrence de cet ensemble est associée à une autre occurrence d'un autre ensemble d'items. Par exemple, « 90% des clients qui achètent un smartphone achètent aussi une pochette et un abonnement à Internet » est une règle d'association associant l'item smartphone aux items pochette et abonnement à Internet. Dans cette thèse, nous nous intéressons plus particulièrement aux techniques des règles d'association qui présentent des avantages, notamment leurs domaines d'applications tels que les télécommunications, aide au diagnostic médicale, la maintenance préventive, le marketing, l'analyse de données spatiales, les réseaux sociaux, la fouille de texte, analyse des accidents routiers, etc.

1.2 Les types de données utilisées en fouille de données

L'ensemble de données non structuré n'est pas analysable par les méthodes de fouille de données. Ainsi il faudra extraire d'une situation complexe de données, une situation analysable, exprimable sous forme de tableaux de données (Figure 1.4). Un tableau de données présente deux ensembles d'objets : les lignes correspondent aux échantillons (Enregistrements) et les colonnes correspondent aux attributs ou champs.

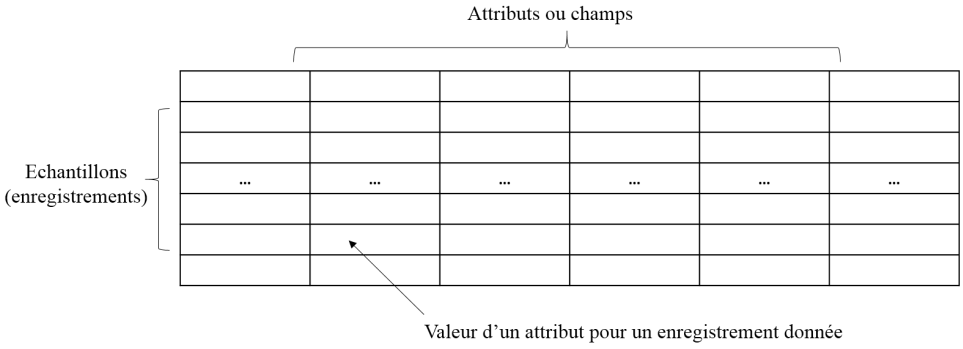

Figure 1.4 – Structure d'un tableau de données

Une donnée est un enregistrement au sens des bases de données, elle est caractérisée par un ensemble de «champs», ces champs peuvent être de deux types numériques ou symboliques. Les attributs numériques comportent les variables réelles ou entières telles que la longueur, le poids, l'âge, etc. Les attributs symboliques (appelés aussi catégoriels) tels que la couleur, etc.

1.3 Définitions

Ce paragraphe définit les principaux concepts liés à la tâche d'extraction des règles d'association. Nous illustrons ces concepts par la base de transactions données dans la Figure 1.5.

Exemple 1.1 :
Pour instance, nous prenons l'exemple de contexte de fouille cité dans [7] : un complexe cinématographique à décidé de fidéliser son public en lançant la carte d'abonnement au cinéma dite « illimité ». Les films vus par chaque cinéphile sont enregistrés dans une base de données à chaque fois que le client se présente au guichet. Elle est exploitée par la suite pour comprendre les attitudes de « consommation » du cinéma, les types de films les plus prisés par le public, les heures auxquelles les gens préfèrent venir voir un film, etc. le tableau $\mathcal{D}$ ci-dessous est un extrait (fictif) et donne pour chaque cinéphile identifié par un numéro Tid, l'ensemble des films qu'il a vu durant le mois courant. Les films concernés sont donnés dans le tableau I. Par exemple la ligne d'identificateur $Tid = 1$ de $\mathcal{D}$ concerne un client ayant vu dans le mois les deux films suivants : « Harry Potter » et « Star Wars II ».

I		
Item	Titre	Réalisateur
a	Harry Potter	Chris Colombus
b	Star Wars II	George Lucas
c	Attrape moi si tu peux	Steven Spielberg
d	Un homme d'exception	Ron Hsoward

D	
Tid	Transaction
1	ab
2	ac
3	ad
4	bcd
5	abcd

Figure 1.5 – Exemple de base de transaction

Définition 1.1 (Item). *Un item est tout objet, article, attribut, appartenant à un ensemble fini d'éléments distincts* $\mathcal{I} = \{x_1, x_2, .., x_m\}$.

Exemple 1.2 :
Les articles en vente dans un magasin sont des items, dans la base $\mathcal{D}$, a, b et c sont des items correspondant aux filmes.

Définition 1.2 (Itemset). *On appelle itemset tout sous-ensemble d'items de I. Un itemset constitué de k items est appelé un k-itemset.*

Exemple 1.3 :
Dans la base $\mathcal{D}$ la transaction 5 est un 4-itemsets constitué de trois items a, b, c et d qui correspondants à tous les films.

Définition 1.3 (Transaction). *Une transaction est un itemset identifié par un identificateur unique* Tid. *L'ensemble de tous les identificateurs des transactions* $Tids$ *sera désigné par l'ensemble* T.

Exemple 1.4 :
Si nous supposons que les items présentés dans le contexte $\mathcal{D}$ (Figure 1.5) sont les articles achetés dans un magasin, nous pouvons dire que le contexte $\mathcal{D}$ est composé des transactions suivantes : $\{1, 2, 3, 4, 5\}$.

Définition 1.4 (Tidset). *On appelle Tidset tout sous-ensemble d'identificateurs de transactions (tids) de T.*

Exemple 1.5 :
Les Tidset $\{1, 2\}$, représentent l'ensemble des identificateurs des transactions 1 et 2 du contexte $\mathcal{D}$.

Définition 1.5 (Taxonomie). *Une taxonomie est un graphe orienté acyclique donnant une classification des items selon un critère de généralité. Les nœuds non-feuilles sont des items généralisés. Une arête (n_i, n_j) signifie que l'item du nœud n_i est plus général que l'item du nœud n_j. Ancêtre(x) désignera l'ensemble de tous les items x qui sont ancêtres de l'item x.*

Exemple 1.6 :
La Figure 1.6 donne un exemple de taxonomie sur les films. Nous avons par exemple *ancêtre(comédies musicales) = {films,comédies}*.

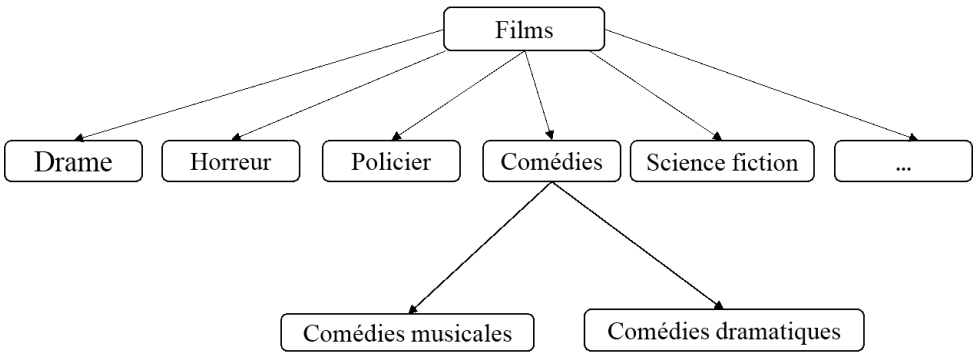

Figure 1.6 – Exemple de taxinomie

Définition 1.6 (Base transactionnelle). *Une base transactionnelle B est un ensemble de couples formés d'un identificateur de transaction Tid et de la transaction proprement dite.*

$$\mathcal{D} = \{(y; X_y)/y \in T; X_y \subset I\}$$

Exemple 1.7 :
Le contexte $\mathcal{D}$ (Figure 1.5) représente une base transactionnelle de 5 transactions 1, 2, 3, 4, 5 et de 4 items $\{a, b, c, d\}$.
I= {a,b,c,d}
T= {1,2,3,4,5}

$\mathcal{D}$= *{(1,ab),(2,ac),(3,cd),(4,bcd),(5,abcd)}*

Comme présenté dans la Figure 1.7, une base de données transactionnelle peut être représentée sous forme horizontale, verticale ou booléenne.

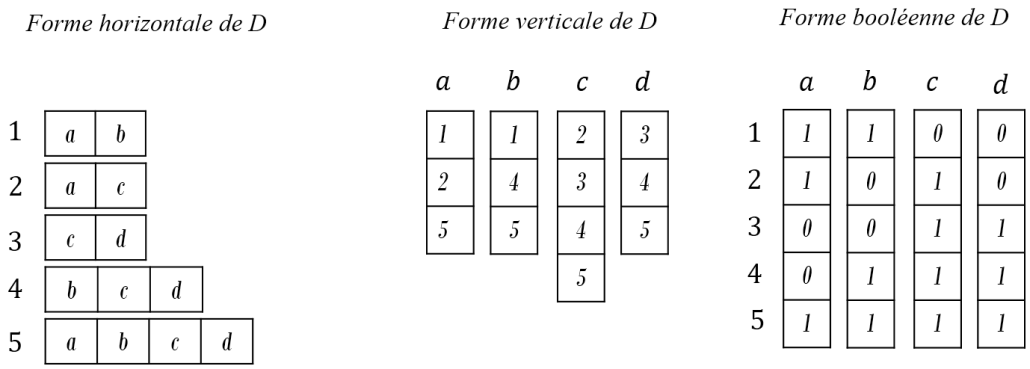

Figure 1.7 – Formes de présentation d'une base transactionnelle

Définition 1.7 (Fréquence). *La fréquence d'un itemset X, noté freq(X), est le nombre de transactions de $\mathcal{D}$ contenant X :*

$freq(X) = |\{(y, X_y) \in D / X \subseteq X_y\}| = |t(X)|$, tel que $t(X)$ le nombre de transaction contenant X

Exemple 1.8 :
Dans l'exemple 1.1, on a $freq(ab) = 2$, vu que l'itemset ab apparait dans les transactions 1 et 5 de D.

Définition 1.8 (Support). *Le support d'un itemset X, noté Supp(X) est la proportion de transactions de $\mathcal{D}$ contenant X, le support prend sa valeur dans l'intervalle [0,1].*

$$Supp(X) = \frac{|\{(y,X_y) \in D / X \subseteq X_y\}|}{|D|} = \frac{|t(X)|}{|D|}$$

Exemple 1.9 :
Dans l'exemple 1.1, on a $Supp(ab) = 2/5 = 0,4(40\%)$ vu que l'itemset ab apparait dans deux transactions parmi 5 de $\mathcal{D}$.

Définition 1.9 (Itemset fréquent). *Étant donné un seuil y, appelé support minimum donné par l'utilisateur, un itemset X est dit fréquent, si son support dépasse le seuil y fixé a priori.*

$$X \text{ est fréquent ssi } Supp(X) \succeq y$$

Exemple 1.10 :
Dans l'exemple 1.1, pour un support de 40%, l'itemset cd de support égal à 60% est fréquent.

Définition 1.10 (Règle d'association). *Une règle d'association est un couple(A,B), où A et B sont des itemsets non vides disjoints, i.e.*$A \neq \emptyset$ *,*$B \neq \emptyset$ *et* $A \cap B = \emptyset$*. On note classiquement une telle règle sous la forme* $A \rightarrow B$*. A s'appelle l'antécédent de la règle et B le conséquent de la règle.*

Un exemple de règle d'association extraite d'une base de données de ventes de supermarché est : $cereales \wedge sucre \rightarrow lait(Support\ 70\%, Confiance\ 50\%)$. Cette règle indique que les clients qui achètent des céréales et du sucre ont également tendance à acheter du lait. La mesure de support définit la portée de la règle, c'est-à-dire la proportion de clients ayant acheté les trois articles, et la mesure de confiance définit la précision de la règle, c'est-à-dire la proportion de clients ayant acheté du lait parmi ceux ayant acheté des céréales et du sucre.

Définition 1.11 (Règle redondante). *Une règle d'association est dite redondante si elle n'est pas porteuse de connaissances supplémentaires par rapport à l'ensemble des règles résultantes.*

Ainsi, il est nécessaire de prendre en considération les mesures de qualité des règles d'association, les mesures les plus utilisées sont le support, la confiance et le lift. Cependant il existe d'autres mesures de qualité que nous verrons un peu plus loin dans la fin de ce chapitre.

Définition 1.12 (Support d'une règle). *On définit le support d'une règle d'association comme étant le support de l'itemset* $A \cup B$ *(i.e. la proportion de transactions contenant à la fois A et B) .*

$$Supp(A \rightarrow B) = Supp(A \cup B) = \frac{|t(A \cup B)|}{t|A|}$$

Définition 1.13 (Confiance d'une règle). *La confiance d'une règle* $A \rightarrow B$*, notée* $Conf(A \rightarrow B)$*, représente la probabilité conditionnelle qu'une transaction contenant B sachant qu'elle contient A*

$$Conf(A \rightarrow B) = \frac{Supp(A \cup B)}{Supp(A)} = \frac{|t(A \cup B)|}{|t(A)|}$$

Remarque : $0 \preceq Conf(A \rightarrow B) \preceq 1$

Pour illustrer la notion de confiance, on considère deux ensembles de transactions $t(A)$ et $t(B)$, (Figure 1.8). La confiance est une mesure permettant d'évaluer la solidité de la règle d'association, elle mesure le degré d'inclusion de A dans B, où $t(A)$ représente le nombre de transaction contenant A et $t(B)$ représente le nombre de transaction contenant B.

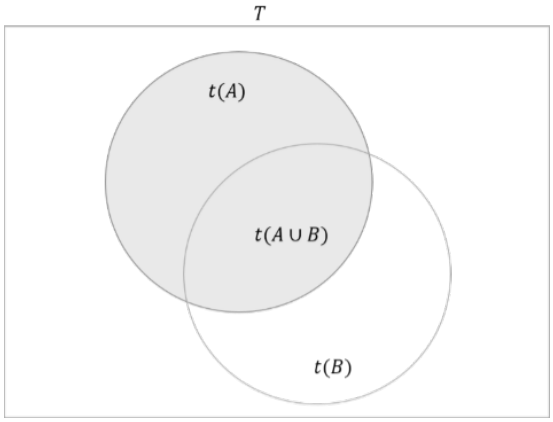

Figure 1.8 – Inclusion des transactions de A dans celles de B

L'extraction des règles d'association consiste à extraire les règles dont le support et la confiance sont au moins égaux respectivement, à des seuils minimaux de support et de confiance définis par l'utilisateur. La plupart des approches proposées pour l'extraction des règles d'association reposent sur les quatre phases suivantes :

Préparation des données : Cette phase consiste à sélectionner les données (attributs et objets), de la base de données, utiles à l'extraction des règles d'association et transformer ces données en un contexte d'extraction.

Extraction des ensembles fréquents d'attributs : Cette phase consiste à extraire du contexte tous les itemsets fréquents, un itemset est fréquent si son support est supérieur ou égal au seuil minimal de support défini par l'utilisateur.

Génération des règles d'association : Cette phase consiste à utiliser les itemset fréquents extraits de la phase précédente pour générer les règles d'association ayant la confiance supérieure ou égale à la confiance minimale.

Visualisation et interprétation : Cette phase consiste à la visualisation des règles d'association extraites afin d'en déduire des connaissances utiles pour l'amélioration de l'activité concernée.

1.4 Algorithmes d'extraction des règles d'association

De nombreux algorithmes ont été développés pour résoudre le problème de recherche des motifs fréquents [3, 4, 13]. Ces algorithmes peuvent être classés en trois grandes catégories : les algorithmes d'extraction des motifs fréquents, des motifs maximaux [14, 15, 16] et des motifs fermés [17, 18, 19, 20]. Dans ce qui suit, nous allons présenter l'état de l'art des algorithmes d'extraction des motifs fréquents.

1.4.1 Algorithme de génération des règles d'association

Le concept de règles d'association a été initié en 1993 par Agrawal et al. [4, 21]. L'une des applications typiques de l'extraction des règles d'association est l'analyse du panier de la ménagère. Elle consiste à rechercher des corrélations entre les produits à travers les tickets de caisse des clients afin de comprendre leurs habitudes de consommation, organiser les promotions, gérer les stocks, agencer les rayons du magasin, etc. dernièrement, cette technique est utilisée dans les domaines cherchant à regrouper des objets entre eux par le degré de similarité, tels que le domaine médical, CRM, accidentologie, industrie, les réseaux sociaux, texte mining etc.
Une règle d'association décrit une association entre des ensembles d'items dans une base de données. En d'autres termes, étant donné un ensemble d'attributs, le but de l'extraction des règles d'association est de découvrir si l'occurrence d'un tel ensemble d'items (article) dans une transaction est associée à l'occurrence d'un autre ensemble d'attributs. Par exemple, « 70% des clients qui achètent un ordinateur achètent aussi une imprimante et un abonnement à Internet » est une règle d'association associant l'item ordinateur aux items imprimante et abonnement à Internet.

Le cadre classique de problème de la fouille des règles d'association peut être décrit de la façon suivante : Soit $I = \{x_1, x_2, \dots x_n\}$ un ensemble d'items. Un sous-ensemble de I est aussi appelé itemset. Soit $\mathcal{D}$ un ensemble de transactions où chaque transaction est un ensemble d'items, identifié par un identifiant unique (Tid). Soit X_1 et X_2 deux itemsets disjoints. Une règle d'association est une implication de la forme $X_1 \rightarrow X_2(S\%, C\%)$. La valeur de S représente la proportion des transactions dans $\mathcal{D}$ contenant X_1 et X_2 cette mesure est appelée le support, tandis que la valeur de C représente la proportion de transactions parmi celle contentent X_1 qui contiennent aussi X_2, cette mesure est appelée la confiance. L'extraction des règles d'association passe par deux étapes importantes, la première étape consiste à extraire l'ensemble de tous les itemsets fréquents, toute on respectant la valeur du support minimum proposé par le décideur. La deuxième consiste à générer, à partir de l'ensemble des itemsets fréquents extraits, les règles d'association en tenant compte du seuil de confiance proposé.

Une règle d'association R, est une relation entre deux itemsets fréquents L_1 et L_2 tel que $L_1 \prec L_2$, notée $L_2 \rightarrow L_1 - L_2$, L_2 et $L_1 - L_2$ sont appelés, respectivement, antécédents et conséquents de la règle R. pour générer les règles d'association, on considère l'ensemble F des itemsetrs fréquents trouvés dans la phase de l'extraction des itemsets fréquents d'un algorithme d'extraction par exemple Apriori [4]. Les règles d'association valides sont celles dont la confiance est supérieure ou égale au seuil minimal fixé par l'utilisateur. La procédure de génération des règles d'association notée $Gen - Rules(l_k, H_{m+1})$ (Figure 1.9) est basé sur la propriété qui permettant de ne pas considérer tous les sous-ensembles des itemsets fréquents.

Proprieté 1.1.*Soit l un itemset fréquent,nous avons :*

$$\forall C \subset l, C \neq \emptyset, [(l - C) \rightarrow C] \text{ est solide} \Longrightarrow \widetilde{C} \subset C, \widetilde{C} \neq \emptyset, [(l - \widetilde{C}) \rightarrow \widetilde{C}] \text{ est solide}$$

Cette propriété signifie que si une règle d'association avec une conséquence C est solide, toutes les règles ayant pour conséquences des sous-ensembles de C sont aussi solides. Dans L'algorithme présenté dans la Figure 1.9, F représente l'ensemble des itemsets fréquents et H_m l'ensemble des m-itemsets conséquents de règles.

Notations :
F : itemsets fréquents.
H_m : m-itemsets qui sont les conséquences des règles valides générées à partir de k itemset fréquents l_k.

Entrées : k-itemsets fréquents l_k, H_m de m-itemsets conséquences de règles valides générées à partir de l_k, Support minimum, $minconf$
Sorties : Ensemble AR de règles d'association valides générées à partir de l_k

```
1  si (k ≻ m + 1 alors
2  |  H_{m+1} ← Apriori − Gen(H_m);
3  |  pour chaque h_{m+1} ∈ H_{m+1} faire
4  |  |  confiance(r) ← support(l_k)/(l_k − h_{m+1});
5  |  |  si (confiance(r) ⪰ minconf alors
6  |  |  |  AR ← AR ∪ {r : (l_k − h_{m+1}) → h_{m+1}};
7  |  |  |  sinon
8  |  |  |  |  Supprimer h_{m+1} de H_{m+1};
9  |  |  |  fin
10 |  |  fin
11 |  fin
12 |  Gen − Rules(l_k, H_{m+1});
13 fin
```

Figure 1.9 – Procédure de génération des règles d'association

L'algorithme considère chaque itemset fréquent de F de taille supérieure à 1. Pour chacun de ces itemsets l_k l'ensemble H_1 des itemsets de taille 1 sont générés et pour chaque itemset h_1 la règle est générée en tenant compte du seuil de minconfiance. La procédure $Gen - Rules(l_k, H_m)$ reçoit en entrée un k-itemsets fréquent, un ensemble h_m des itemsets conséquents de règles valides générées et un seuil minimal de confiance. La ligne 1 correspond au test d'arrêt des appels récursifs de la procédure. Ensuite, l'ensemble h_{m+1} des (m+1)-itemsets des conséquents de règles valides générées à partir de l_k est créé en appliquant la procédure $Apriori - Gen(H_m)$ à l'ensemble H_m des m-itemsets (ligne 2). Chaque règle dont la conséquence est un (m+1)-itemsets de H_{m+1} est alors testées (ligne 3 à 8) si la Règle est valide, elle est insérée dans AR (ligne 6) sinon l'itemset (m+1) est supprimé de H_{m+1} (ligne 8).

Exemple 1.11 :
La génération des règles d'association à partir de l'ensemble F des itemsets fréquents de

l'étape précédente (Figure 1.9) pour un seuil minimal de $minconf = 2$ sont présentés dans le Tableau 1.1.

Tableau 1.1 – Les règles d'association générées

Itemset	Règle	Support	Confiance	Solide ?
ab	$a \longrightarrow b$	2/5	2/3	oui
	$b \longrightarrow a$	2/5	2/3	oui
ac	$a \longrightarrow c$	2/5	2/3	oui
	$c \longrightarrow a$	2/5	2/4	non
bc	$b \longrightarrow c$	2/5	2/3	oui
	$c \longrightarrow b$	2/5	2/4	non
bd	$b \longrightarrow d$	2/5	2/3	oui
	$d \longrightarrow b$	2/5	2/3	oui
cd	$c \longrightarrow d$	3/5	3/4	oui
	$d \longrightarrow c$	2/5	3/3	oui
bcd	$bc \longrightarrow d$	2/5	2/3	oui
	$bd \longrightarrow c$	2/5	2/2	oui
	$cd \longrightarrow b$	2/5	2/2	oui

1.4.2 Algorithmes d'extraction des itemsets fréquents

Pour cette catégorie, nous trouvons les algorithmes de base ayant résolu le problème de la détermination des itemsets fréquents. Ces algorithmes sont basés sur la propriété d'antimonotonicite [4] : « Tout sous-ensemble d'un ensemble d'articles fréquent est fréquent, et tout sur ensemble d'un ensemble non fréquent est non fréquent ». Parmi les algorithmes de cette catégorie, nous pouvons citer l'algorithme Apriori dans lequel, deux paramètres, support minimum, et confiance minimale sont introduits.

L'algorithme de base Apriori

Apriori est l'algorithme clé pour l'extraction des règles d'association dans les bases de données transactionnelles, proposé par Agrawal et al. [4]. Il constitue la base de la majorité des algorithmes ayant pour objet l'extraction des règles d'association. C'est un algorithme itératif de recherche des itemsets fréquents par niveau. Il permet de chercher les règles d'association en deux étapes dont la première a pour objet de rechercher tous les sous-ensembles d'items fréquents, tels que $support(X) \succ minsup$. Une fois ces derniers sont générés ils constituent les éléments d'entrées pour la recherche de toutes les règles d'associations vérifiant la condition $confiance(X) \succ minconf$.

Principe de base :
Étant donnée une base de données transactionnelle, un support minimum S et une confiance minimale C. le processus d'extraction des règles d'association peut être effectué en deux étapes principales :

Étape 1 : Extraction des itemsets fréquents

L'algorithme Apriori (Figure 1.10) utilise une approche itérative par niveaux pour générer les itemsets fréquents. En effet, l'algorithme calcule les supports des 1-itemsets, en effectuant un premier parcours de la base de données. Les itemsets, dont le support n'a pas atteint le seuil minsupp fixé, sont considérés comme non fréquents. Par la suite, un nouvel ensemble des 2-itemsets, dits ensemble des itemsets candidats, est généré. Après un autre parcours de la base de données, les supports des 2-itemsets candidats sont calculés. Les 2-itemsets non fréquents sont écartés et le processus décrit précédemment est relancé, jusqu'à ce que l'on ne puisse plus générer les itemsets fréquents. La description de la k_{ieme} candidat sera dénotée par C_k et l'ensemble des k-itemsets fréquents de taille k par F_k.

$$C_k = \{(C_k, Supp(C_k))|\forall x \subseteq C_k, x \neq \emptyset, Supp(x) \geq \delta\}, \text{ où } \delta \text{ le support minimum}$$

$$F_k = \{(l_k, Supp(l_k))|l_k \; est \; un \; k-itemset \; et \; Supp(l_k) \geq \delta\}$$

La procédure de génération des itemsets candidats fréquents $Apriori-Gen(F_{k-1})$ (Figure 1.10) est appelée en ligne 3 et prend comme paramètres F_{k-1}, et retourne comme résultat l'ensemble de tous les k-ensembles candidats. Cette fonction s'exécute en deux étapes : l'étape de la jointure qui réalise les jointures possibles des (K-1) et l'étape d'élagage, cet élagage se fait en se basant sur la propriété d'antrimonotonicité des itemsets fréquents [4]. Une fois l'ensemble C_k Des itemsets candidats sont calculés, la base de transactions est parcourue afin de trouver le support de chaque candidat. Si c'est le cas, alors le support de ces candidats est augmenté (ligne 7). Parmi les candidats, seuls ceux qui ont le support supérieur à minsupp sont retenus et qui sont des itemsets fréquents.

Entrées : Base de transaction $\mathcal{D}$, Support minimum δ
Sorties : Ensemble F_k des itemsets fréquents

$F_1 \leftarrow \{1 - itemsets\ fréquents\}$;
pour $(k \leftarrow 2, F_{k-1} \neq \emptyset; k++)$ **faire**
  $C_k \leftarrow Apriori - Gen(F_{k-1})$;
  pour chaque $objeto \in \mathcal{D}$ **faire**
    $C_o \leftarrow Subset(C_k, o)$;
    pour chaque $condidatc \in C_o$ **faire**
      $c.support++$;
    fin
  fin
  $F_k \leftarrow \{c \in C_k | c.support \geqslant \delta\}$;
fin
retourner $\bigcup_k F_k$

Entrées : Ensemble F_{k-1} de (k-1) itemsets fréquents
Sorties : Ensemble C_k k-itemsets condidats

insérer dans C_k;
sélectionner $p[1], p[2], ..., p[k-1], q[k-1]$;
de $F_{k-1}p, F_{k-1}q$;
où $p[1] = q[1], ..., p[k-2] = q[k-2], p[k-1] \prec q[k-1]$;
pour chaque *itemset condidat* $c \in C_k$ **faire**
  pour chaque $sous - ensemble\ s\ de\ c\ de\ taille\ k-1$ **faire**
    si $s \notin F_{k-1}$ **alors**
      supprimer c de C_k;
    fin
  fin
fin
retourner C_k

Figure 1.10 – Extraction des itemsets fréquents à l'aide de l'algorithme Apriori

Exemple 1.12 :
Pour illustrer les différentes étapes de l'algorithme Apriori on l'applique pour la base de données transactionnelle de l'exemple 1.1 pour extraire les itemsets fréquents en utilisant un support minimum $S = 0.4$ (Figure 1.11). Dans la première étape, chaque item I de transaction T est un 1-itemset de C_1. Un premier parcours de $\mathcal{D}$ permet de trouver le support de chaque item (1-itemset), si le support est supérieur ou égal au support minimum, les 1-itemset sont fréquents et sont gardés dans F_1. Ensuite, pour découvrir les 2-itemsets, l'algorithme effectue une jointure de $F_1 \times F_1$ pour trouver l'ensemble C_2. Un second parcours de $\mathcal{D}$ est effectué pour déterminer le support de 2-itemsets candidats. Seules les 2-itemset fréquents sont gardés dans C_2. Les itemsets n'ayant pas le support supérieur ou

égal au support minimum sont supprimés pour construire les itemset candidats C_3 par la jointure de $F_2 \times F_2$. Un troisième parcours de $\mathcal{D}$ est effectué pour déterminer les 3-itemsets fréquents. Les itemsets n'ayant pas le support supérieur ou égal ou support minimum sont supprimés pour construire F_3. De nouveau, on effectue la jointure de $F_3 \times F_3$ pour trouver l'ensemble des itemsets candidats C_4, cet ensemble est vide, car on n'a plus qu'un seul élément de taille 3.

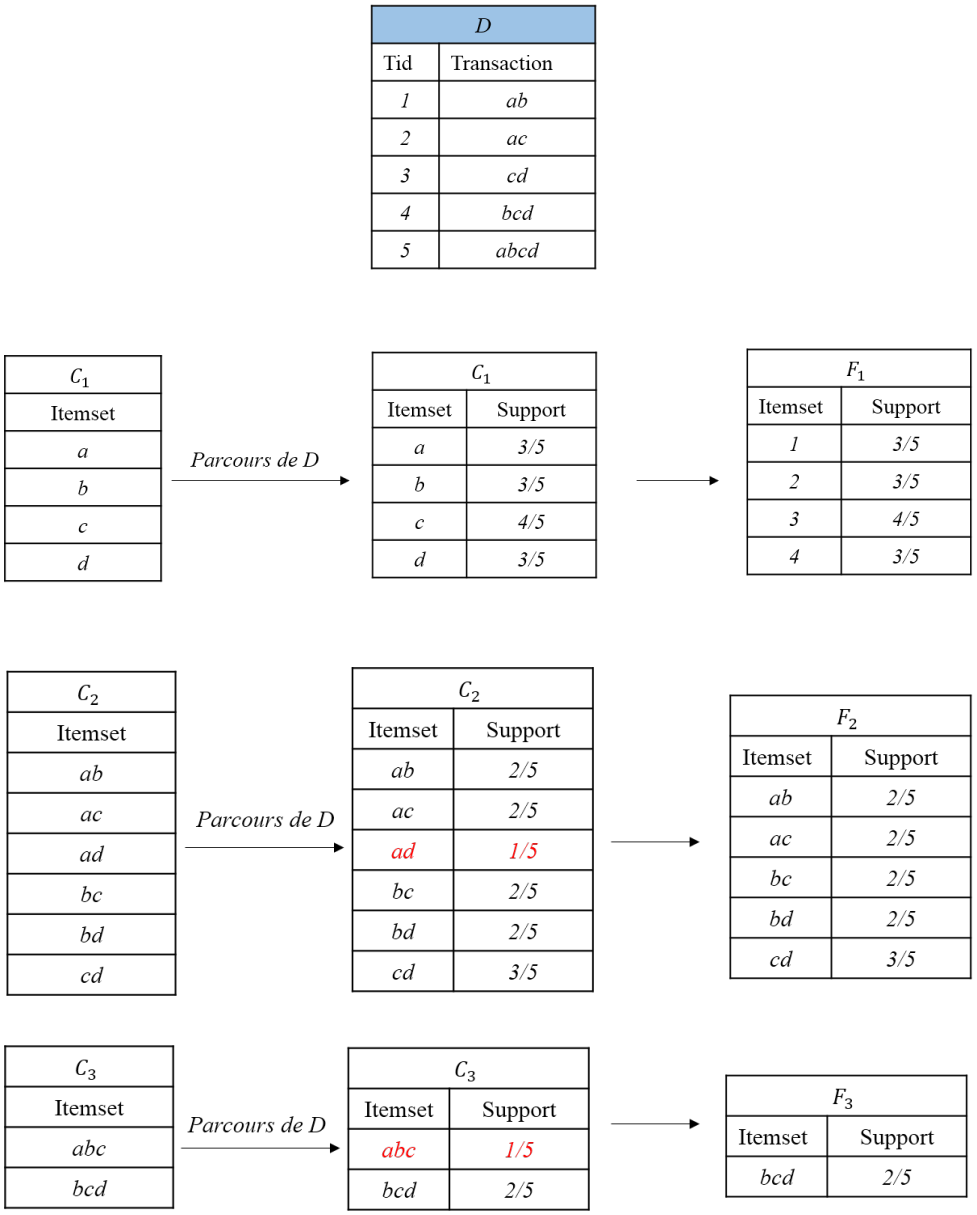

Figure 1.11 – Exemple d'application de l'algorithme Apriori

Dans la catégorie d'extraction des itemsets fréquents, nous trouvons les algorithmes ayant résolu le problème d'extraction des itemsets fréquents. Parmi ces algorithmes, on trouve ArioriTID [22], Partition [13], DIC [23], eclat [15], FP-growth [24], etc. Ces algorithmes sont basés sur la propriété d'antimonotonicité.

L'algorithme AprioriTID

Afin d'améliorer les performances de l'algorithme Apriori, les mêmes auteurs ont proposé dans [22] l'algorithme AprioriTID qui est basé sur le même principe d' Apriori, mais dans AproriTID à partir de deuxième passage, la BD n'est plus utilisée pour calculer les supports des itemsets candidats. Il utilise un ensemble C_k de la forme $(Tid, \{C_k\})$ ou C_k correspond à la liste des itemsets contenus dans la transaction identifiée par Tid. Pour $K = 1$, C_1 correspond à la base de transaction $\mathcal{D}$. L'amélioration qu'apporte cet algorithme par rapport au précédent est le fait de stocker à chaque itération les identificateurs des transactions contenant les sous-ensembles fréquents dans l'ensemble C_k par conséquent, on réduit le nombre de passages de la base de données, le pseudo-code de l'algorithme AprioriTID est donné dans la Figure 1.12.

Entrées : Base de transaction $\mathcal{D}$, Support minimum δ
Sorties : Ensemble F_k des itemsets fréquents

$F_1 \leftarrow \{1 - itemsets\ fréquents\}$;
$\overline{C}_1 \leftarrow \mathcal{D}$;
pour $(k \leftarrow 2, F_{k-1} \neq \emptyset; k++)$ **faire**
  $C_k \leftarrow Apriori - Gen(F_{k-1})$;
  $\overline{C}_1 \leftarrow \emptyset$;
  pour chaque *chaque objet* $o/o.OID \in \overline{C}_{k-1}$ **faire**
    $C_o \leftarrow \{c \in C_k | (c - C_k) \in o.liste - condidats \wedge (c - C_k) \in o.liste - condidats\}$;
    pour chaque $condidatc \in C_o$ **faire**
      $c.support++$;
    fin
    si $(s \notin \emptyset)$ **alors**
      $\overline{C}_k \leftarrow \overline{C}_k \cup \{(o.OID, C_k)\}$;
    fin
  fin
  $F_k \leftarrow \{c \in C_k | c.support \geqslant \delta\}$;
fin
retourner $\bigcup_k F_k$

Figure 1.12 – Extraction des itemsets fréquents à l'aide de l'algorithme AprioriTID

L'algorithme Partition

L'algorithme Partition proposée par Savasere et al. dans [13] considère seulement deux parcours de la base de transaction $\mathcal{D}$ pour l'extraction des itemsets fréquents. Comme il indique son nom le principe est de partitionner la base de transaction en m partition, $\{D_1, \ldots, D_m\}$ puis chercher des itemsets fréquents locaux sur chaque D_i et pour chaque

itemset trouvé, on calcule son support sur toute la base de transaction. Par exemple à partir de l'itemset a ayant pour Tidset 125 et l'itemset b de Tidset 145, on déduit le support de l'itemset ab par $125 \cap 145 = 15$ représentants les transactions contenant a et b. L'intérêt de cet algorithme est d'améliorer les performances de l'algorithme Apriori.

L'algorithme DIC

L'algorithme DIC (Dynamic Itemset Counting) proposé par Brin et al. [23] pour réduire le nombre de parcours de la base de transaction. L'idée de base est de partitionner la base de données en M blocs de transactions, après le parcours d'une partition de taille m on vérifier les k-itemsets candidates qui ont déjà atteint le support minimum pour générer les candidates de taille $(k+1)$. Par exemple, Apriori peut produire 3 parcours pour compter 3-itemsets tandis que DIC produit 1.5 parcours. Et dans le premier parcours avec 1-itemet, DIC peut compter certains itemsets qui sont 2-itemsets ou 3-itemsets. Le pseudo-code de l'algorithme est donné dans la Figure 1.13.

Entrées : Base de transaction $\mathcal{D}$, Support minimum δ, taille de la fenêtre de lecture $\mathcal{M}$
Sorties : Ensemble F_k des itemsets fréquents

$F_1 \leftarrow \{1 - itemsets\ avec\ situation = IP\}$;
pour *chaque objet* o **faire**
  pour chaque *chaque* $c \in o / c.situation = IP$ **faire**
    $c.support++$;
    insérer c *dans* P;
  fin
fin
pour *chaque condidate* $c \in P / c.situation = IP\ et\ c.support \geqslant \delta$ **faire**
  $c.situation \leftarrow FP$;
  pour chaque $sur - ensemble\ x\ de\ c / |x| = |c| + 1\ et\ x \notin C_{|x|}$ **faire**
    si ($\forall s\ sous - ensemble\ de\ x\ de\ taille\ |x| - 1\ nous\ avons\ s.situation = FC\ ou\ FP$) **alors**
      insérer x *dans* $C_{|k|}$ *avec* $x.situation = IP$;
    fin
  fin
fin
pour *chaque condidate* $c \in P / c.situation = IP\ ou\ FP$ **faire**
  $c.nombrelu \leftarrow c.nombrelu + \mathcal{M}$;
  si ($c.nombrelu = |\mathcal{D}|$ **alors**
    si ($c.support \geqslant \delta$) **alors**
      $c.situation \leftarrow FC$;
      sinon
        $c.situation \leftarrow IC$;
      fin
    fin
  fin
fin
si ($\exists c \in C | c.situation = FP\ ou\ IP$) **alors**
  lire $\mathcal{M}$ *objets dans* $\mathcal{D}$;
fin
$F_k \leftarrow \{c \in C_k | c.situation = FC\}$;
retourner $\bigcup_k F_k$

Figure 1.13 – Algorithme DIC : Extraction des itemsets fréquents

L'algorithme ECLAT

L'algorithme ECLAT (Equivalence Class clustering and bottom Up Lattice Transversal), proposé par Zaki et al. [15] utilise le format vertical de la base de données. En effet il effectue une recherche des itemsets fréquents en profondeur on se basant sur le concept

de classe d'équivalence (deux k-itemset appartenant à une même classe d'équivalence si ils ont en commun un préfixe de taille (K-1).

L'algorithme SSDM

SSDM (Semantically Similar Data Miner) proposé par Escovar et al. [25] est un algorithme de la famille d'Apriori auquel est rajouté des notions d'ensembles flous afin de renforcer la recherche d'associations sur des bases plus robustes on utilisant les matrices de similarités. L'algorithme nécessite de fixer une valeur minimale de similarité $minSim$ qui détermine si l'algorithme doit confondre deux éléments d'un même domaine en une seule entité. Le support du candidat flou est évalué selon le poids de ses similarités par la formule suivante :

$$poids_c = \frac{[poids(a)+poids(b)][1+sim(a,b)]}{2}$$

Où $poids_c$ est le poids du candidat C contenant les éléments a et b où $poids(a)$ correspond au nombre d'occurrences de a dans la base de données transactionnelle. La formule générale pour n éléments est donnée par la formule suivante :

$$poids = [\sum_{i=1}^{\infty} poids(item_i)][\frac{1+f}{2}]$$

L'algorithme FP-growth

Dans le but de répondre au problème posé par l'algorithme de base Apriori, Han et al. [24] propose FP-growth (Frequent Pattern growth) qui utilise une nouvelle technique pour générer les itemsets fréquents sans avoir extraire les itemsets candidates. Elle consiste d'abord à compresser la base de données en une structure compacte appelée FP-tree (Frequent Pattern tree), puis à diviser la base de données en sous-projections de la base de données appelées bases conditionnelles. Chacune de ces projections est associée à un item fréquent. La construction du FP-tree passe par deux parcours de $\mathcal{D}$ et se fait de la manière suivante : On effectue un premier parcours pour déterminer les items fréquents pour un support minimum donné, ces itemsets fréquents seront triés par la suite par ordre décroissant de support dans une liste L. Dans le deuxième parcours de $\mathcal{D}$ chaque transaction est triée selon l'ordre des items dans L. Au début, le nœud racine de l'arbre (*null*) est créé, ainsi une branche sera créée pour chaque transaction, les transactions ayant un même préfixe partagent le même début d'une branche de l'arbre. Les items sont traités de plus fréquent au moins fréquent pour une bonne structure compacte à laquelle les items fréquents sont proches de la racine et sont mieux partagés par les transactions.

Une fois le FP-tree construit, on commence par des itemsets suffixes de taille 1, l'extraction se fait récursivement sur cette nouvelle sous-structure. Ces itemsets fréquents sont obtenus par concaténation du suffixe de la base conditionnelle avec les itemsets fréquents de sous arbre conditionnel. Un exemple de FP-tree est donné dans la Figure 1.14.

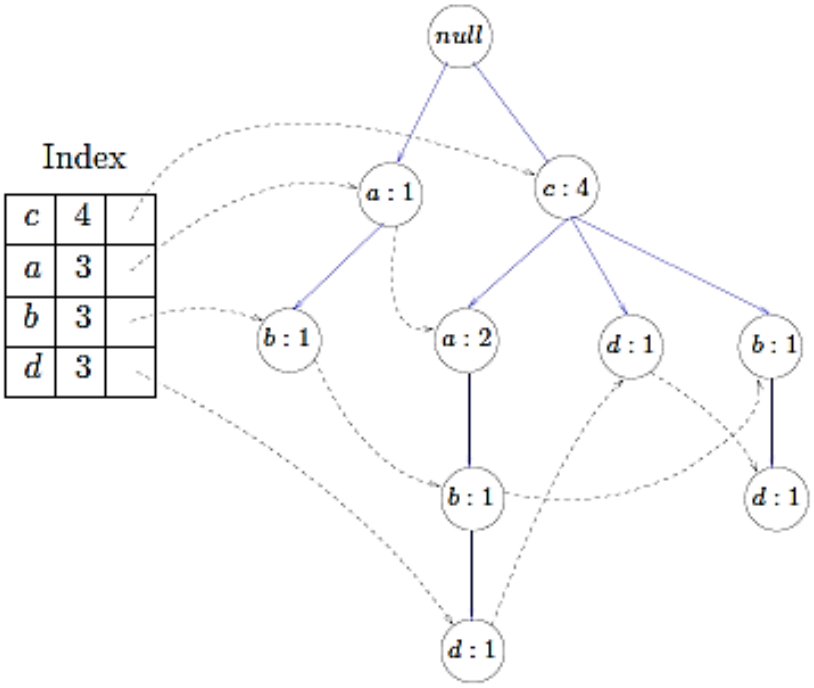

Figure 1.14 – FP-tree exemple

1.5 Fouille de données et extraction de la connaissance spatiale

Nous sommes souvent intéressés par l'analyse des situations complexes dans le but d'accroître la précision de prédire l'effet de certains phénomènes spatiaux. Une fois que son comportement est estimé par un modèle, le phénomène spatial peut être compris plus correctement. En ce moment des modèles spatiaux utilisés sont habituellement créés d'une manière très simple et ne représentent que la tendance générale. Pour donner le modèle d'une forme plus réaliste, des méthodes avancées d'analyse de données spatiales devraient être utilisées lorsqu'une représentation d'un phénomène spatiale existe.

Au cours des dernières années, les outils d'analyse statistique traditionnels éprouvent des difficultés pour la manipulation d'énormes volumes de données stockées dans les bases de données. De plus, les méthodes statistiques nécessitent une connaissance plus générale des données afin de définir une hypothèse principale pour l'analyse. L'extraction de connaissances est de plus en plus coûteuse et fastidieuse. Par conséquent, la statistique classique devient un outil inapproprié et inadapté pour l'analyse de données. L'exploration de données est introduite en tant que discipline qui s'intéresse à l'analyse des bases de données. L'objectif principal de la fouille de données est la recherche des informations cachées qui peut être transformée en connaissances utiles pour la prise de décisions. Cependant, la question principale concernant la fouille de données spatiales est la façon de traiter les relations spatiales intégrées aux bases de données géographiques.

1.5.1 Les Systèmes d'Informations Géographiques (SIG)

Un SIG est un système informatique de matériels, de logiciels et de processus, conçu pour permettre la collecte, la gestion, la manipulation et l'affichage de données à référence spatiale en vue de résoudre des problèmes d'aménagement et de gestion. On appelle donnée à référence spatiale toute donnée peuvent être localisée de façon directe (une école, une route, etc) ou indirecte (une adresse, un propriétaire, etc) à la surface de la Terre. Pour transformer un objet réel en une donnée à référence spatiale, on décompose le territoire en couches thématiques (relief, routes, bâtiments, etc) structurées dans des bases de données numériques (Figure 1.15 [26]). Les bases de données qui alimentent les SIG doivent être géoréférencées, c'est-à-dire partager un cadre commun de repérage appelé système de projection. Ce cadre commun est fixé légalement. Les références cartographiques utilisées au Maroc se réfèrent à la projection conforme de Lambert.

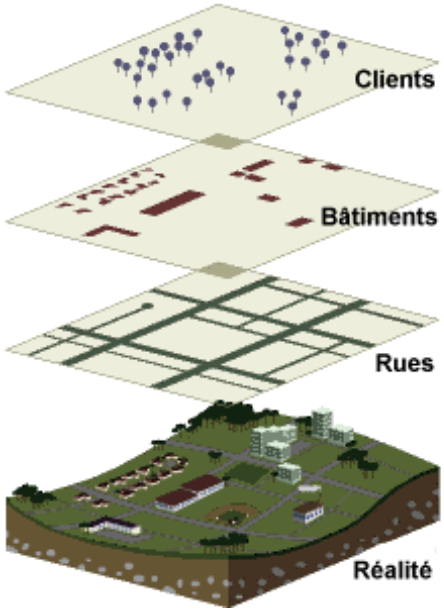

Figure 1.15 – Systèmes d'Informations Géographiques

1.5.2 Concepts de base de l'information géographique

Dans ce qui suit, nous présenterons les principaux concepts de l'information géographique à savoir, les caractéristiques d'un objet spatial, les couches thématiques et les relations spatiales.

Objet spatial

Un objet spatial est un type de données particulier pour les bases de données relationnelles. Ces objets peuvent être simples comme un point, une ligne, un polygone, ils représentent des données géographiques. Chaque objet spatial est caractérisé par des attributs spatiaux et non spatiaux.

Attributs non spatiaux
Des variables représentent une information classique non spatiale (nom, nombre d'étudiants, etc).

Attributs spatiaux
Des variables représentant des informations à référence spatiale. La position géographique de cet objet est repérée selon un système de projection. Il existe plusieurs manières de localiser un objet dans l'espace, par l'usage des coordonnées, ou adresses postales ou bien la localisation approximative.

Couche thématique
Est une carte qui représente uniquement un type précis de donnée pour mieux qualifier les objets et les phénomènes disposés dans l'espace que par leur simple forme matérielle. Il existe deux modes fondamentaux de représentation numérique des données géographiques, le mode raster et le mode vectoriel.
Le mode raster correspond à un tableau de valeurs numériques référencées géographiquement par rapport à un système de coordonnées, l'unité spatiale fondamentale est le pixel ou la cellule. L'espace est décomposé en une grille régulière et rectangulaire, organisée en ligne et en colonnes, chaque maille de cette grille à une intensité de gris ou une couleur. Le format vectoriel (Figure 1.16) utilise le concept d'objets géométriques (points, lignes et polygones) pour représenter les entités géographiques.

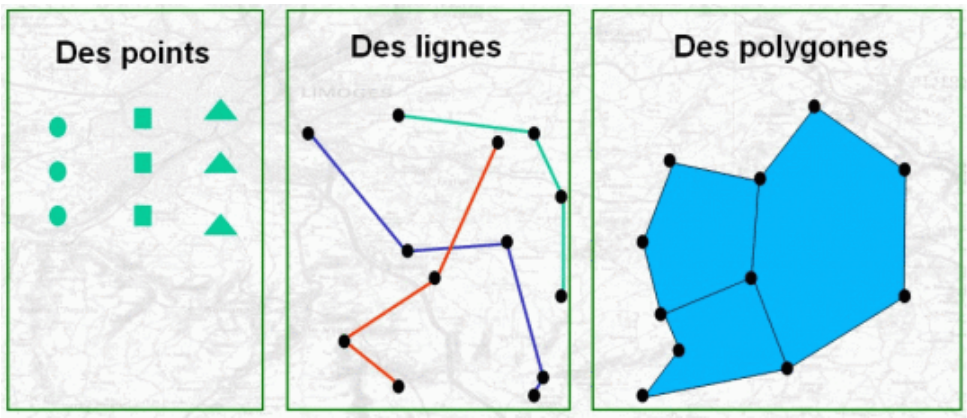

Figure 1.16 – Données Vecteur (Aspect géométrique)

Les deux schémas de la Figure 1.17 montrent la représentation raster et vecteur de la même zone. l'apparence en bloc est la représentation raster, alors que la représentation vecteur est faite de points, de lignes et de polygones.

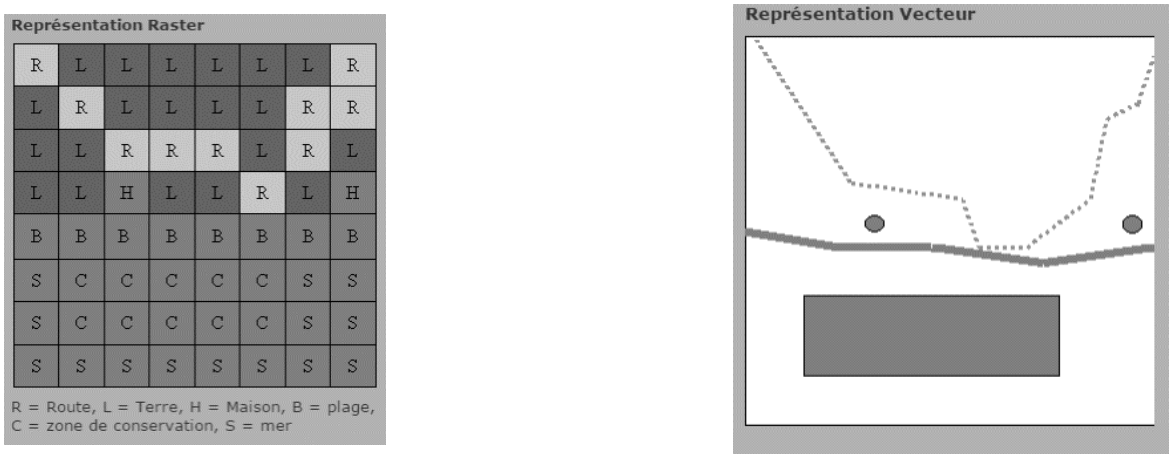

Figure 1.17 – Représentations numériques des données géographiques

Relations spatiales
Dans le monde réel, les objets d'une couche thématique peuvent avoir des relations avec les autres objets des autres couches thématiques. Ces relations peuvent être classées en trois catégories :

— Relations métriques basées sur la distance : près de, loin de, etc ;

— Relations topologiques : contenance, couverture, intersection, etc ;

— Relations d'orientation : à gauche, à droite, etc.

Toutes ces relations répondent bien à la première loi en Géographie [27] selon laquelle « Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés »

1.5.3 Les composantes d'un SIG

Un SIG possède cinq composantes majeures qui sont le matériel (ordinateur, serveur) qui servant à effectuer la numérisation et le stockage de données. Les données qui sont indispensables elles peuvent être de trois types : géographiques, attributaires ou méta-données. Les utilisateurs qui vont exploiter et maintenir les données géographiques, et les logiciels qui offrent les outils pour remplir les fonctionnalités connues par « les 5A » à partir d'une interface graphique, l'utilisateur va interroger une base de données afin d'analyser les données.

1.5.4 Les Fonctionnalités d'un SIG

Un Système d'Information géographique est constitué de 5 composantes majeures présentées comme suite :

Acquisition
Acquérir les données géographiques à travers la génération d'un lot de données complet

en passant par la saisie de quelques données spatiales. Il existe différentes techniques d'acquisition (numérisation, télédétection, acquisition manuelle...)

Assemblage des données géographiques
Il s'agit de fusionner les différentes sources de données, afin de construire une base de données géographique.

Archivage
C'est la structuration et le stockage des données géographique dans les systèmes de gestion de fichiers ou dans les bases de données. Le stockage peut être réalisé grâce aux différents systèmes de gestion de bases de données spatiales selon le modèle abstrait prédéfini.

Analyse
L'analyse spatiale géométrique est l'étude des formes, des positions et des relations entre les objets comme le calcul de distances, d'intersections ou d'exclusions par exemple. Parmi les outils d'analyse, nous trouvons les requêtes sémantiques (sur les attributs des objets) et les requêtes géométriques.

Affichage
C'est la représentation et la visualisation des résultats, notamment sous forme cartographique (cartes, tables, et documents textes).

1.5.5 La fouille de données spatiale

Grâce à l'évolution des techniques de collecte des données comme la télédétection, la numérisation, la surveillance météorologique et climatologique, etc., les bases de données géographiques contiennent une énorme quantité de données de divers types et attributs. L'analyse de ces données est un défi pour les méthodes classiques d'analyse qui sont principalement basées sur les statistiques. Depuis que les méthodes classiques d'exploration de données nous permettent de détecter des informations précieuses à partir de gros volume de données relationnelles, la fouille de données spatiales peut être une technique appropriée pour la détection des modèles intéressants dans les jeux de données géographiques. La fouille de données spatiales est définie comme l'extraction de connaissances implicites, de relations spatiales ou d'autres propriétés non explicitement stockées dans les bases de données spatiales [8]. En effet, sa spécificité par rapport à la fouille de données est qu'elle prend en compte les relations spatiales entre les objets.

1.5.6 Comparaison entre la fouille de données et fouille de données spatiales

L'extraction d'informations implicites à partir de bases de données géographiques apparaît, en comparaison avec les bases de données spatiales non traditionnelles plus difficile.

Les objets à référence spatiale sont représentés dans l'espace par des propriétés géométriques et topologiques. La topologie couvre les propriétés géographiques qui ne sont pas étroitement liées à la position réelle d'objets, c'est-à-dire qu'il représente les relations spatiales entre objets. Selon [28] la topologie est une branche de la géométrie qui traite les propriétés d'un objet qui restent inchangé, même lorsque l'objet est transformé [29]. L'emplacement est généralement décrit par des coordonnées euclidiennes ou la latitude et la longitude. La différence entre la fouille de données et la fouille de données spatiales peut être résumée en trois points essentiels :

Type de données
Pour la fouille de données spatiales, les données sont plus complexes, car elles incluent la forme géométrique de l'objet tel que les points, les lignes et les polygones.

Relations entre les données
Contrairement aux relations d'ordre, arithmétique, existantes dans la fouille de données, les relations entre les objets spatiaux sont implicites (relations d'orientation, relations topologiques, relations métriques).

Méthodes utilisées
Les méthodes permettant d'analyser les données sont issues du domaine des statistiques et de celui des bases de données. Le point commun de ces méthodes est l'exploitation des relations spatiales de voisinage.

Primitives des relations spatiales
L'analyse des données spatiales nécessite la prise en compte des relations spatiales qui sont explicites entre phénomènes qui fournissent des données nécessaires pour les algorithmes du Data Mining, nous disons que le DM spatial est une extension de DM classique.
Les objets spatiaux peuvent être des points ou des objets dans l'espace prolongé tels que des lignes, des polygones. L'influence mutuelle entre deux objets dépend des facteurs tels que la topologie, la distance ou la direction entre les objets. Par exemple, un nouvel ensemble industriel peut polluer son voisinage selon la distance et sur la direction principale du vent. Dans cette section, nous présentons trois types de bases de relations spatiales : Relations topologiques, de distance et de direction.

Les relations topologiques
Les relations topologiques sont des relations qui restent invariables sous des transformations topologiques, elles sont préservées si les deux objets sont changés d'échelle, translatés ou pivotés simultanément. Les définitions formelles sont basées sur les frontières, les intérieurs et les compléments des deux objets connexes. Ces relations selon les extensions SQL d'oracle spatial sont [30] : Inside, Touch, Covers, Equal, Contains, Disjoint, Coverded By, Overlap Boundary, (Figure 1.18).

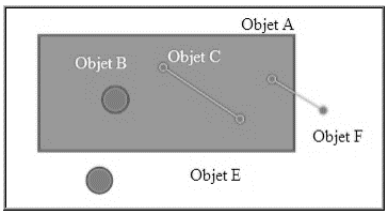

Figure 1.18 – Exemples de relations topologiques entre différents objets

Les relations de distance

Les relations de distance sont des relations comparant la distance de deux objets à une constante donnée en utilisant un des opérateurs arithmétiques. La distance entre deux objets, peut-être définie par la distance minimale entre leurs points. Nous citons selon SQL étendu d'oracle spatial [30] : Within Distance, Nearest Neighbor, (Figure 1.19).

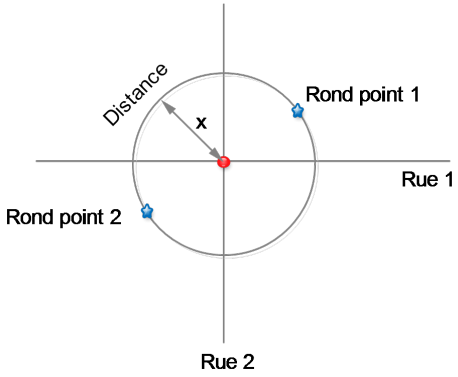

Figure 1.19 – Exemples de relation de distance

Les relations de direction

Pour définir la relation de direction $O_1 R O_2$ nous distinguons l'objet source O_1 de l'objet destination O_2 de la relation R de direction. Il y a plusieurs possibilités pour définir des relations de direction selon le nombre de points qu'ils considèrent dans la source et l'objet de destination. Nous définissons la relation de direction de deux objets dans l'espace prolongé utilisant un représentant de l'objet O_1 de source et tous les points de la destination d'objet O_2.

1.5.7 Techniques de la fouille de données spatiales

Différents types de modèles peuvent être extraites à partir de bases de données et peuvent être présentés en différentes formes. La catégorisation dépend souvent de l'arrière-plan d'un champ particulier de recherche. Si nous supposons qu'une personne soit intéressée à la visualisation des données, les critères de classification seront probablement dépendants de diverses techniques de visualisation, alors qu'un chercheur en informatique pourrait voir les principaux écarts dans l'utilisation de différents algorithmes. Un aperçu de différentes possibilités de classification des techniques d'extraction de données est donné dans [31]. Ester [32] divise les techniques d'exploration des données spatiales en

quatre grands groupes : les règles d'association spatiale, regroupement spatial, clustring et la classification. Les trois techniques les plus utilisées sont la classification, le clustering et les règles d'association [33].

1.5.8 Méthodes de la fouille de données spatiales

Avec le développement de la cartographie numérique, les chercheurs ont proposé des méthodes de découverte de connaissances à partir de base de données spatiale, parmi les méthodes principales de la fouille de données on trouve :

Le clustering spatial :
C'est une technique qui permet de découvrir des connaissances dans les bases de données spatiales. Il s'agit d'une classification automatique non supervisée d'objets, permettant de regrouper les objets de la base de données dans des sous-classes significatives, et ce, en respectant le principe de base de clustering : la similarité doit être maximale pour les objets de la même classe et elle doit être minimale pour les objets des classes différentes [34].

Exemple 1.13 :Soit x et y deux objets spatiaux et a une distance.

— Si $d(x, y) \prec a$ alors x et y doivent être mis dans la même classe.

— Si $d(x, y) \succ a$ alors x et y doivent être mis dans deux classes différentes.

La classification spatiale :
C'est une méthode qui vise à analyser les données spatiales, et ce, en recherchant des règles dites règles de classification. La recherche de ces règles vise à structurer un ensemble d'objets en classes d'objets ayant des propriétés communes. Dans le processus de la classification spatiale, nous cherchons les règles qui partagent ou divisent l'ensemble des objets dans des classes utilisant non seulement les propriétés non spatiales, mais aussi ses relations spatiales avec les autres objets de la base de données [35].

Règle d'association spatiale :
C'est une technique puissante de la fouille de données qui permet d'étudier les relations, corrélation et dépendances entre les objets afin d'extraire les connaissances. Le concept de règle d'association a été initié par Agrawal [4] pour l'analyse des grandes bases de données transactionnelles. Il consiste à rechercher des associations entre objets selon leurs caractéristiques. Il est présenté sous forme de motif : $Corps \longrightarrow Tête$. Nous nous intéressons plus particulièrement aux règles d'association spatiales qui sont une extension des règles d'association classique [10] elles sont de la forme suivante : $P_1 \wedge P_2 \wedge, ..., \wedge P_n \longrightarrow Q_1 \wedge Q_2 \wedge Q_m(S\%, C\%)$. Où au moins un des prédicats $P_2, P_2, ..., P_n$ *et* $Q_1, Q_2, \ldots Q_m$ est un prédicat spatial, $S\%$ est le support de la règle et $C\%$ est la confiance de la règle.

La découverte des règles d'association est une tâche importante, elle consiste à détecter des associations entre les objets selon les propriétés spatiales et non spatiales. Un exemple de règle d'association spatiale est $R2 : < est\ un, téléboutique > \wedge <$

$près\ de, faculté> \Longrightarrow < intérieure, cyber >$ $(80\%, 60\%)$. La règle $R2$ exprime que 80% des téléboutiques sont près de la faculté et sont à l'intérieur des cyber et que 60% des téléboutiques qui sont près de la faculté sont aussi à l'intérieur des cyber.

1.5.9 État de l'art des règles d'association spatiales

L'extension de l'extraction des règles d'association au contexte spatial est introduite par Koperski [10]. Cette extension a pris en considération la composante spatiale des objets, ces règles représentent des relations spatiales et non spatiales entre les objets spatiaux.
Parmi les méthodes de base qui ont traité la problématique d'extraction des règles d'association spatiales on trouve la méthode de Koperski [10] et celle de Salleb [7]. Pour la première méthode, proposée par Koperski, elle se base sur les techniques de recherche, avec approfondissement progressif, de haut en bas (top-down), au niveau d'une hiérarchie de concepts (Figure 1.20). Ces techniques entament d'abord la recherche des itemsets fréquents, pour le premier niveau de concept, ensuite, et uniquement pour chaque itemset fréquent, poursuivent la recherche à un niveau de concept inférieur. En ce qui concerne les attributs descriptifs, ils sont organisés sous forme de l'hiérarchies de concepts, l'extraction des règles d'association est alors guidée par ces hiérarchies en parcourant chacune d'elles niveau par niveau. Néanmoins, les informations appartenant à différents niveaux des hiérarchies ne peuvent être apprises. De plus, l'utilisation des attributs non hiérarchiques n'est pas mise en évidence dans son algorithme. Salleb a fait une extension du travail de Koperski en considérant les attributs hiérarchiques et non hiérarchiques et en permettant un mélange des niveaux entre les différentes hiérarchies.

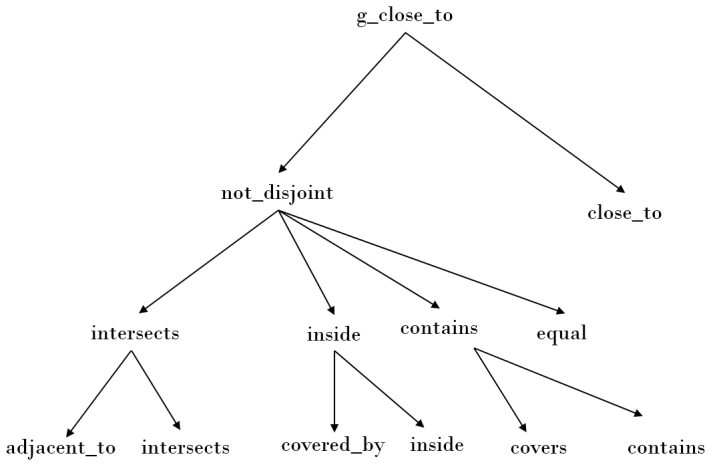

Figure 1.20 – Hiérarchie de concept des relations topologiques

Méthode de Koperski
Plusieurs types de prédicats spatiaux peuvent être utilisés dans les règles d'association spatiales (les relations topologiques, les relations d'orientations spatiales, et les rela-

tions métriques). Dans Koperski ces prédicats sont utilisés dans le processus du Data Mining avec deux concepts : g_close_to et $g\_coarse\_close\_to$. Le prédicat spatial : $g\_close\_to(X, Y)$ est satisfait par les objets X de type x_type et Y de type y_type si X et Y sont localisé à une distance seuil (d) spécifié pour des objets de type x_type et des objets de type y_type.
Le prédicat spatial : $g\_coarse\_close\_to$ est satisfait par les objets X de type x_type et Y de type y_type si les rectangles minimaux d'enveloppe MBR (Minimum Bounding Rectangle) pour les objets X et Y sont localisés à une distance seuil spécifiée pour les objets de type x_type et y_type.

Description de l'algorithme
L'algorithme proposé par koperski est composé des étapes suivantes (Figure 1.21) :

Entrées : Base de données spatiales, la requête concerne l'ensemble référencé S des objets à décrire, l'ensemble de descriptions des objets de S et l'ensemble des relations spatiales, Support minimum $minsupp$, la confiance minimale $minconf$.
Sorties : Les règles d'association spatiales multi-niveaux pour l'ensemble des objets spatiaux et des relations spatiales.

```
1  Procedure : find_frequent_predicate(D);
2  pour (l ← 1; L[l,1] ≠ ∅ et l < max_level; l++) faire
3  |  L[l,1] ← get_large_1_predicate_set(D, l);
4  |  pour (k ← 2; L[l,k−1] ≠ ∅; k++) faire
5  |  |  P_k ← get_condidate_set(L(l,k−l));
6  |  |  pour chaque objet o ∈ S faire
7  |  |  |  P_o ← get_subsets(P_k, o);
8  |  |  |  pour chaque condidate p ∈ P_o faire
9  |  |  |  |  p.support++;
10 |  |  |  fin
11 |  |  |  L[l,k] ← {p ∈ P_k | p.support ≥ minsupp[l]};
12 |  |  fin
13 |  |  LL[l] ← ⋃_k L[l,k];
14 |  fin
15 fin
16 retourner gen_association_rules(LL[l])
```

Figure 1.21 – Algorithme de Koperski : Extraction des règles d'association spatiales

A chaque niveau de concept, pour le calcul des k-prédicats fréquents (pour tous les k), après la recherche des k-prédicats fréquents, l'ensemble des règles d'association spatiales peut être dérivé, pour chaque niveau de concept l, en se basant sur la confiance minimale

fixée pour ce niveau ($minconf$).

En résumé l'algorithme de koperski, en considérant $LL[l]$ comme étant le tableau contenant l'ensemble des prédicats fréquents au niveau l, et $L[l,k]$ comme étant le tableau contenant l'ensemble des k-prédicats fréquents au niveau l. Ces règles ainsi générées nécessitent d'être examinées par un expert ou passées à travers un système automatique testant la qualité d'une règle, dans le but d'éliminer quelques règles redondantes, et de garder celles exprimant de l'intérêt pour les utilisateurs.

Méthode de Salleb

L'algorithme ARGIS (Association Rules in GIS), proposé par Salleb et al. [7] est un algorithme itératif basé sur la notion de table de liens, notée T_l. Il a défini deux catégories de couches thématiques (couche de référence et couche descriptive), Figure 1.22.

Principe

Étant donne une couche de référence C_r, et un ensemble de couches descriptives C_d, nous cherchons alors l'ensemble des règles d'association spatiales décrivant les objets de la couche de référence C_r par rapport aux objets de la couche descriptive en question. Autrement dit, les règles d'association découvertes dans ARGIS considèrent à chaque fois deux et seulement deux couches thématiques, il n'est pas possible d'extraire des règles regroupant plus de deux couches thématiques. Une table de liens, définie sur trois champs, donne pour chaque objet de la couche de référence identifie par une clé id_R, un ou plusieurs objets de la couche descriptive identifié par une clé unique id_D, qui lui sont reliés par une relation spatiale. Pour chaque couche descriptive, nous calculons donc sa table de liens avec la couche de référence.

Entrées :
- base de données géographique($\mathcal{S}, \mathcal{T}, \mathcal{H}$);
- couche de référence C_r;
- couches descriptives $\{C_{d_i}/i = 1, ..., n\}$;
- ensemble d'attributs non spatiaux pour C_r et pour tous les C_d;
- ensemble de prédicats spatiaux à calculer SP;
- buffer (km),$minsupp$, $minconf$;

Sorties : $\mathcal{R}$:Règles d'association spatiales, multi-niveaux, solides

```
1  pour tous les (i = 1, ..., n) faire
2  |  création des tables de lien;
3  |  calcul dans T_{L_i}, SP entre C_r et C_{d_i};
4  |  calcul des prédicats fréquents;
5  |  pour chaque exemple E de clé j de T_{L_i} faire
6  |  |  calculer prédicats fréquents sur E selon minsupp;
7  |  fin
8  |  Génération de règles R_i sur ces prédicatss selon minconf;
9  fin
10 retourner ⋃_{i=1}^{n} R_i
```

Figure 1.22 – ARGIS : Algorithme d'extraction de règles d'association dans un SIG

Méthode de Marghoubi

Dans ce contexte, Marghoubi et al. [9] proposé une nouvelle approche pour l'extraction des règles d'association spatiale basée sur les fondements mathématiques du concept de Treillis de Galois notamment la fermeture de la connexion de Galois [9]. Afin de résoudre le problème posé par ARGIS, concernant la prise en considération de deux et seulement deux couches thématiques à chaque comparaison, Maroubi a utilisé un système de codification des objets spatiaux permettant d'identifier la couche thématique de n'importe quel objet spatial.

Principe

L'algorithme proposé peut être résumé en quatre étapes, la première est relative à la préparation des données selon les choix effectués par les décideurs. La deuxième a pour but d'extraire l'ensemble des générateurs fermés fréquents du contexte spatial, selon le choix de décideur et la valeur du support minimum. La troisième est relative à l'extraction des règles d'association spatiales à partir de l'ensemble des générateurs extraits dans la deuxième étape. Quant à la dernière, elle est relative à la visualisation des résultats à l'aide de la plateforme Galicia [9].

1.6 Mesures de qualités des règles d'association

Depuis les travaux d'Agrawal [4], l'extraction des règles d'association est devenue aujourd'hui l'une des tâches les plus populaires de la fouille de données. Il a pour but de découvrir des relations intelligibles entre les attributs dans une base de données. Le processus de l'extraction des règles d'association passe par deux étapes importantes, la première est l'extraction des motifs fréquents ainsi que la deuxième étape concerne la génération des règles d'association à partir des motifs fréquents précédemment extraites. Cette dernière génère une quantité importante des règles qui ne permettent pas aux utilisateurs de faire leur choix des règles les plus pertinentes. Pour filtrer les règles d'association issue d'un contexte de la fouille de données, on utilise des critères communément appelés mesures de qualité de règles. Le support et la confiance sont deux mesures utilisées habituellement pour évaluer la solidité des règles d'association, ces deux mesures ont été critiquées, car ils ne peuvent réellement exprimer la corrélation entre itemestes d'une manière efficace. C'est pourquoi d'autres mesures ont été proposées [36, 23, 37] permettent de déterminer la précision des règles d'association.

Définition 1.14 (Redondance d'une règle d'association)
une règle d'association $R : X \rightarrow Y'$ est redondante s'il existe une autre règle $R' : X' \rightarrow Y'$, telle que $X \rightarrow X'$, $Y \rightarrow Y'$ et, le support et la confiance de R et R' sont identiques.

Étant donnée une règle $A \rightarrow B$, elle est courant de l'analyser en se rapportant à une matrice de contingence croisant de deux variables binaires A et B. On assimile les ensembles A et B à deux évènements. $P(A)$ dénote alors la probabilité qu'un évènement A arrivé, estime par le nombre d'occurrences de l'itemset A dans D, et de même pour B.

On obtient alors une description de la règle sous l'une des formes listées dans le tableau 1.2 et la Figure 1.23, n_x corresponds au nombre de transactions contenant X (ou fréquence absolue), et p_x la probabilité de X (ou fréquence relative). La Figure 1.23 illustre ces notations en fréquences absolues et relatives. Par convention, on note multiplicativement l'union de deux itemsets afin d'alléger l'écriture (ainsi, le nombre de transactions contenant $X \cup Y$ est n_{xy}). Le premier tableau présente ces notations en fréquences absolues, la seconde en fréquences relatives. Les cellules du tableau de contingence sont liées par les relations suivantes :

1. $P_a + P_{\overline{a}} = 1$
2. $P_b + P_{\overline{b}} = 1$
3. $P_{ab} = P_a - P_{\overline{a}b} = P_a - P_{a\overline{b}}$
4. $P_{a\overline{b}} = P_a - P_{ab} = P_{\overline{b}} - P_{\overline{ab}}$
5. $P_{\overline{a}b} = P_b - P_{ab} = P_{\overline{a}} - P_{\overline{ab}}$
6. $P_{\overline{ab}} = P_{\overline{b}} - P_{a\overline{b}} = P_{\overline{a}} - P_{\overline{a}b}$

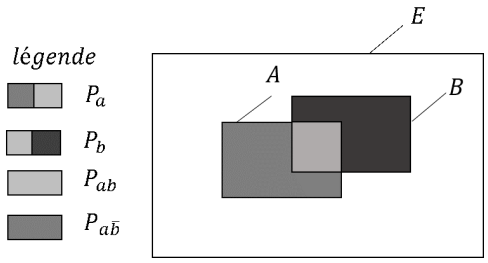

Figure 1.23 – Notations usuelles associées à une règle $A \rightarrow B$

Tableau 1.2 – Notations usuelles associées à une règle d'association $A \rightarrow B$

A B	0	1	Total
0	$P_{\overline{ab}}$	$P_{\overline{a}b}$	$P_{\overline{a}}$
1	$P_{a\overline{b}}$	P_{ab}	P_a
$Total$	$P_{\overline{b}}$	P_b	1

Soit A et B deux motifs positifs. Quatre types de règles d'association peuvent être obtenus à partir de A et B :

— Une règle dite positive de la forme $A \rightarrow B$ ou $B \rightarrow A$;
— Une règle dite négative à droite de la forme $A \rightarrow \overline{B}$ où $B \rightarrow \overline{A}$;
— Une règle dite négative à gauche de la forme $\overline{A} \rightarrow B$ où $\overline{B} \rightarrow A$;
— Une règle dite bilatéralement négative de la forme $\overline{A} \rightarrow \overline{B}$ où $\overline{B} \rightarrow \overline{A}$.

La validité d'une règle d'association est évaluée à partir d'une mesure de qualité. Nous donnons ci-dessous la définition d'une mesure de qualité.

Définition 1.15 (Mesure de qualité)
Une mesure de qualité des règles est une fonction μ de l'ensemble des règles d'association à valeurs dans R telle que pour toute règle d'association $A \rightarrow B$, $\mu(A \rightarrow B)$ est fonction exclusivement de quatre paramètres $n, p(A'), p(B')$ et $p(A' \cap B')$ ou p désigne la probabilité discrète uniforme sur l'espace probabilisable $(E, P(E))$.

Intelligibilité d'une mesure
Une mesure doit être intelligible [38, 39] i.e., elle doit être facile à interpréter. C'est le cas des mesures support et confiance, elles ont un sens « concret », elles sont facilement interprétables par l'utilisateur. Soit $A \rightarrow B$ et $C \rightarrow D$ deux règles d'association ayant le même support et telles que $conf(A \rightarrow B) = 2\ conf(C \rightarrow D)$ la seule connaissance de la confiance permet à l'utilisateur de savoir que la règle $A \rightarrow B$ est deux fois plus fiables que la règle $C \rightarrow D$.

Nous présentons dans ce qui suit quelque mesures de qualité les plus utilisées dans la littérature :

Le Support

Le support [4] est une mesure de qualité indique la proportion d'entités vérifiant à la fois l'antécédent et le conséquent de la règle. Il prend ses valeurs dans l'intervalle [0, 1] est défini par la formule 1.1 :

$$Supp(A \rightarrow B) = Supp(A \cup B) = P(A \cap B) \tag{1.1}$$

La Confiance

La confiance [4], est une mesure exprimée à l'aide de la probabilité conditionnelle d'avoir l'évènement B sachant que l'évènement A s'est produit, c'est une mesure descriptive qui prend ses valeurs dans l'intervalle [0, 1]. Il est défini par la formule 1.2 :

$$Conf(A \rightarrow B) = P(B|A) = \frac{A \cup B}{P(A)} \tag{1.2}$$

Le Lift

Le Lift [36], est une mesure statistique, symétrique, représente le rapport d'indépendance entre l'antécédent et le conséquent de la règle. Il prend ses valeurs dans l'intervalle $[0, +\infty[$. Elle est défini par la formule 1.3 :

$$Lift(A \rightarrow B) = \frac{Conf(A \cup B)}{Supp(B)} \tag{1.3}$$

La Conviction

Est une mesure non symétrique qui mesure le dégrée d'implication de la règle, elle est proposée dans [36] pour pallier au limite de la confiance. La conviction varie dans l'intervalle $[0, +\infty[$, plus cette valeur est grande le conséquent apparait avec l'antécédent. Elle est donné par la formule 1.4 :

$$Conv(A \rightarrow B) = \frac{1 - Supp(B)}{1 - conf(A \rightarrow B)} \tag{1.4}$$

Nous avons aussi les trois cas suivants :
si $Conv(A \rightarrow B) = 1$ alors A et B sont indépendants
si $Conv(A \rightarrow B) > 1$ alors A et B sont positivement dépendants
si $Conv(A \rightarrow B) < 1$ alors A et B sont négativement dépendants

Pearl

La mesure de Pearl [40] est une mesure statistique, symétrique, implicative et sensible à la taille de données. Elle permet de mesurer l'intérêt d'une règle par rapport à l'indépendance entre l'antécédent et le conséquent. Elle prend ses valeurs dans l'intervalle [0, 1] et donné par la formule 1.5 :

$$Pearl(A \rightarrow B) = p(A)|p(B|A) - p(B) \tag{1.5}$$

φ-coefficient

φ-coefficient [41], est une mesure statistique, symétrique non implicative est sensible à la

taille de données, elle mesure l'association entre l'antécédent et le conséquent. Elle prend ses valeurs dans l'intervalle [-1, 1] et définie par la formule 1.6 :

$$\varphi(A \to B) = \frac{leve(A \to B)}{\sqrt{(Supp(A) \times Supp(B) \times (1 - Supp(A)) \times (1 - Supp(B))}} \tag{1.6}$$

La Confiance centrée

La mesure de confiance centrée [38] permet de mesurer l'influence de réalisation du conséquent par apport à celle de l'antécédent d'une règle. Une valeur voisine de zéro implique que la règle proche de la situation d'indépendance entre l'antécédent et le conséquent. Elle prend ses valeurs dans l'intervalle [-1, 1] et définie par la formule 1.7 :

$$Conf_{centree} = p(A)|p(B|A) - p(B) \tag{1.7}$$

La Loevinger

Loevinger [42], est une mesure de qualité non symétrique qui normalise la mesure de confiance centrée par le nombre d'entités ne vérifiant pas le conséquent de la règle. Elle prend ses valeurs dans l'intervalle $]-\infty, 1]$ et définie par la formule 1.8 :

$$loevinger(A \to B) = \frac{p(A)|p(B|A) - p(B)}{p(\overline{B})} \tag{1.8}$$

Pietetsky-Shapiro

Est une mesure symétrique [43] qui évalue l'intérêt d'une règle par rapport à son écart à l'indépendance. Elle prend ses valeurs dans l'intervalle [-n, n] et définie par la formule 1.9 :

$$Piatetsky(A \to B) = np(B)(p(B|A) - p(B)) \tag{1.9}$$

La Nouveauté

Comme la mesure de Pietetsky-Shapiro, la Nouveauté [44] est une mesure symétrique, qui évalue l'intérêt d'une règle par rapport à son écart à l'indépendance. Elle prend ses valeurs dans l'intervalle [-1, 1] et définie par la formule 1.10 :

$$Nouveaute(A \to B) = p(A \cap B) - p(A)p(B) \tag{1.10}$$

Laplace

C'est une estimation de confiance qui prend en compte le support [45] elle varie dans l'intervalle [0, 1[et défini par la formule 1.11 :

$$Laplace(A \to B) = \frac{Supp(A \cap B) + 1}{Supp(A) + 2} \tag{1.11}$$

Jaccard

Le coefficient de Jaccard prend des valeurs dans [0, 1] et évalue la distance entre l'antécédent et la conséquence et défini par la formule 1.12 :

$$Nouveaute(A \to B) = \frac{Supp(A \cap B)}{Supp(A) + Supp(B) - Supp(A \cap B)} \tag{1.12}$$

Le classement des règles d'association par utilisation des mesures de qualités est un domaine de recherche qui a attiré de nombreux auteurs dans la littérature. Deux catégories de mesures sont identifiées : mesures subjectives et objectives. Nous avons examiné certaines propriétés importantes qui sont largement discutées pour donner un panorama général sur ce problème. Nous rassemblons dans la Annexe (Figure A.1) les principales mesures de qualité d'une règle auxquelles nous nous référons par la suite.

1.7 Conclusion

Le cout d'extraction des règles d'association dépend fortement de la phase d'extraction des itemsets fréquents. L'algorithme Apriori pose trois problèmes majeurs, le premier est relatif au temps de calcul (plusieurs parcours de la base de données, 2^n avec n la taille de la base de données), le deuxième concerne la qualité des règles d'association extraites et le troisième concerne le stockage en mémoire pour les bases de données massives (Big data).

La recherche des connaissances pertinentes dans les bases de données est un problème primordial dans le domaine de la fouille de données en particulier au niveau de l'analyse des règles d'association. Pour ce faire, les mesures de qualités servant à évaluer les règles d'association selon leur utilité et en fonction des préférences des décideurs. Dans ce chapitre, nous avons décrit les trois travaux qui ont été faits pour l'extraction des règles d'association spatiales. Ces travaux utilisent, pour l'extraction des itemset fréquents, l'algorithme Apriori qui présente des limitations au niveau de temps de réponse, au niveau de l'espace mémoire et au niveau des résultats obtenus (plus de redondance). Ensuite, Nous avons présenté les mesures de qualité des règles d'association proposées dans la littérature. Dans le chapitre suivant, nous présentons les fondements et les méthodes d'analyse multicritère et leurs apports au processus d'extraction des règles d'association.

Chapitre 2

L'analyse multicritère et la logique floue

«If you do not know how to ask the right question, you discover nothing.»

W.Edwards Deming

Dans ce chapitre, nous allons nous intéresser aux outils méthodologiques qui seront utilisés dans la démarche d'aide multicritère à la décision constituant la base de notre approche qui sera présentée dans la contribution. Nous présenterons également les méthodes d'analyse multicritère, ainsi que les concepts de base de processus de décision. Nous présenterons ensuite, les fondements mathématiques de la théorie des ensembles flous et son apport au processus d'extraction des règles d'association.

Sommaire

2.1 L'analyse multicritère

2.1.1 Introduction

L'analyse multicritère se présente comme une alternative aux méthodes classiques basées sur la définition d'une fonction unique, souvent exprimée en terme économique et qui reflète la prise en compte de plusieurs critères. L'intérêt des méthodes multicritères est de considérer un ensemble de critères de différente nature (exprimés en unités différentes), sans nécessairement les transformer en critères économiques ni en une fonction unique. Il ne s'agit pas de rechercher un optimum, mais une solution compromise qui peut prendre diverses formes : choix, affectation ou classement. Plusieurs méthodes existent dans la littérature, ces méthodes multicritères connaissent de multiples applications dans différents domaines tels que le transport, agriculture, le management de territoire, la politique et la conception de réseaux de télécommunication. Dans ce chapitre nous allons définir le cadre théorique et les aspects méthodologiques des méthodes multicritères, ensuite nous allons illustrer leurs approches en étudiant deux types de méthodes multicritère : ELECTRE TRI et PROMETHEE.

2.1.2 L'aide à la décision

Selon Roy [46], « L'aide à la décision est l'activité de celui qui prenant appui sur des modèles clairement explicités, mais non nécessairement complètement formalisés, aide à obtenir des éléments de réponses aux questions que se pose un intervenant dans le processus de décision, éléments concourant à éclairer la décision et normalement à prescrire, ou simplement à favoriser un comportement de nature à accroître la cohérence entre l'évolution du processus d'une part, les objectifs et le système de valeurs au service desquels cet intervenant se trouve placé d'autre part ». Autrement dit, l'aide à la décision est un processus qui utilise un ensemble d'information à un instant donné pour formuler un problème en fonction d'un ensemble d'alternatives à l'aide des méthodes d'analyse multicritères selon les quatre problématiques de base : problématique de choix, noté P_α, la problématique de tri notée P_β, la problématique de rangement noté P_γ et la problématique de description notée P_δ, voir la Figure 2.1.

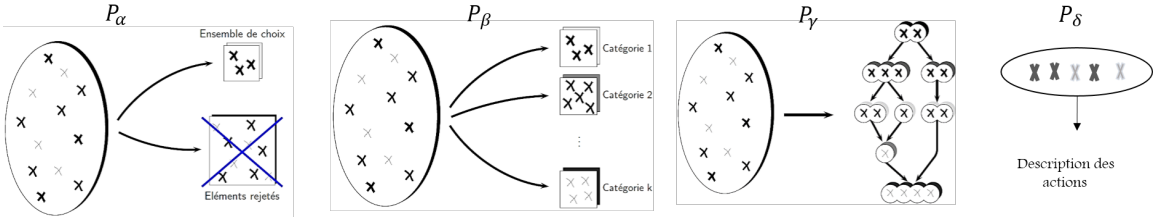

Figure 2.1 – Les problématiques d'analyse multicritère

2.1.3 Processus de décision

Simon [47] à proposé que le processus d'aide à la décision puisse être décrit selon trois phases :

— Intelligence : cette phase commence par identifier les éléments structurants le problème ;

— Désigne : formulation du problème et description des solutions potentielles ;

— Choix : à partir de l'évaluation de chaque solution, le décideur choisit la meilleure d'entre elles.

2.1.4 Terminologie

Cette section décrit les notions de base de l'analyse multicritères, action, critère, poids, préférence, matrice de décision, etc.

Définition 2.1 (Action). *Une action est une représentation d'une éventuelle contribution à la décision globale susceptible, eu égard à l'état d'avancement du processus de décision, d'être envisagée de façon autonome et de servir de point d'application à l'aide à la décision [48].*

Définition 2.2 (Critère). *Un critère est une fonction définie sur l'ensemble des actions représentant les préférences de l'utilisateur selon son point de vue [48].*

Définition 2.3 (Poids). *Le poids mesure l'importance d'un critère par rapport aux autres, selon les préférences de décideur [48].*

Définition 2.4 (Matrice de décision). *La matrice de décision représente généralement, les valeurs des critères par rapport aux actions. Elle est constituée en ligne des actions et en colonnes, des critères, des valeurs qui remplissent ce tableau sont des évaluations d'actions selon les critères.*

Les seuils et les poids sont donnés par les décideurs, pour chaque action considérée, et pour chaque critère un seuil de préférence p, d'indifférence q et un seuil de veto v sont estimés. Chaque critère a un poids k traduisant sa contribution dans la décision finale. Le résultat de l'analyse des conséquences est présenté dans un tableau de performances, voir le Tableau 2.1.

Tableau 2.1 – Tableaux des performances

Critères	g_1	g_2	...	g_j	...	g_n
Poids	k_1	k_1	...	k_j	...	k_n
Seuils	p_1	p_1	...	p_j	...	p_n
	q_1	q_1	...	q_j	...	q_n
	v_1	v_1	...	v_j	...	v_n
Actions	g_1	g_1	...	g_j	...	g_n
A_1	$g_1(a_1)$	$g_2(a_1)$	...	$g_j(a_1)$	...	$g_n(a_1)$
A_2	$g_1(a_2)$	$g_2(a_2)$	...	$g_j(a_2)$	...	$g_n(a_2)$
...	...	...	...	...	...	...
A_n	$g_1(a_n)$	$g_2(a_n)$	...	$g_j(a_n)$	...	$g_n(a_n)$

2.1.5 Les étapes d'une méthodologie d'aide à la décision

Afin de formaliser le processus d'aide à la décision associé à un problème décisionnel, nous présenterons dans la Figure 2.2 les différentes étapes principales à suivre. Dans un premier temps nous identifions le problème, cette étape commence par l'analyse et l'identification des acteurs concernés avec leurs rôles au sein du processus. Ensuite, nous déterminons les alternatives à évaluer en fonction d'un ensemble de critères, cette étape permet de choisir une méthode multicritère d'aide à la décision. Une fois la méthode est choisi, nous déterminons les différents seuils et poids de l'ensemble des critères choisis pour évaluer les alternatives. Une fois le modèle construit, une dernière étape peut être nécessaire. Elle consiste à construire une recommandation à partir des résultats fournis par le modèle.

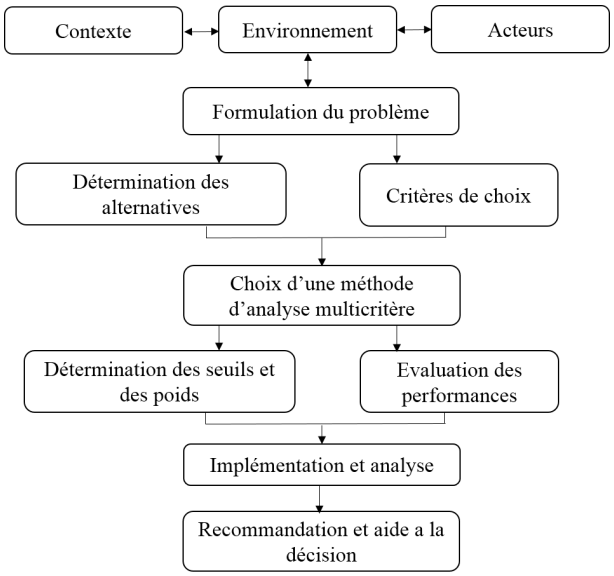

Figure 2.2 – Les étapes d'une méthodologie d'aide à la décision

2.1.6 Problématiques multicritères de décision

Roy distingue quatre problématiques en aide à la décision multicritère [48] : problématique de choix, problématique de tri, problématique de rangement et problématique de description :

- Problématique de choix α : Il s'agit de la problématique classique dans laquelle, l'aide à la décision est orientée de telle sorte à ce que le résultat soit une sélection d'un ensemble de « bonnes » alternatives.
- Problématique du tri β : Consiste à poser le problème en termes de tri des alternatives par catégories. Le choix de cette catégorie est justifié par le type de jugement que l'on voudrait porter sur les actions et par les traitements que l'on souhaiterait faire.
- Problématique du rangement γ : Consiste à ranger les alternatives selon leurs mérites à partir de l'évaluation de chaque alternative.
- Problématique de la description δ : Aider à décrire les actions et/ou leurs conséquences de façon systématique et formalisée ou à élaborer une procédure cognitive. Ce type de problématique est approprié lorsque le décideur rencontre des difficultés à définir le problème.

2.1.7 Les méthodes d'analyse multicritère

Dans la littérature, il existe deux grandes familles de méthodes d'analyse multicritère. La première représente les méthodes utilisant un critère unique de synthèse dont le principe consiste à agréger les performances d'une alternative en un seul critère, tels que les méthodes TOPSIS, AHP, MAUT, MAVT, UTA, SMART, etc. La deuxième famille représente les méthodes du surclassement dont le principe consiste à comparer les alternatives par paires, comme les méthodes ELECTRE TRI et PROMETHEE. Le classement des méthodes d'analyse multicritère selon le type de problématiques décisionnelles est donné dans la Figure 2.3.

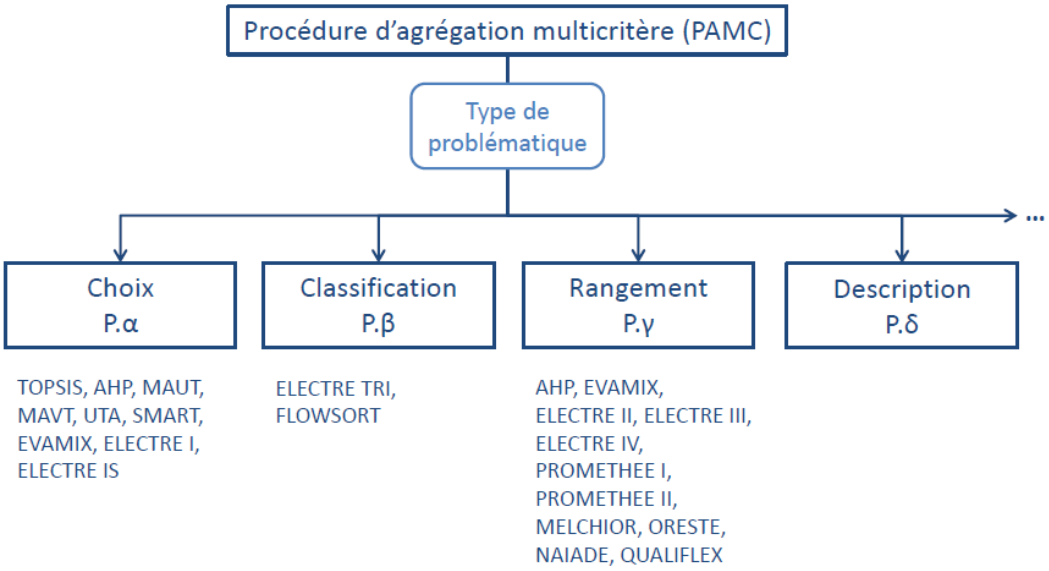

Figure 2.3 – Classement des méthodes selon le type de problématique multicritère de décision

Dans le cadre de cette thèse, nous avons choisi d'utiliser les méthodes de surclassement. Le principe de ces méthodes consiste à comparer les alternatives par paires au moyen d'une relation de surclassement S. Cette relation permet de répondre à l'une des problématiques (choix, tri, rangement) mentionnées dans la Figure 2.3. L'objectif de ces méthodes n'est pas de prescrire une solution au décideur, mais plutôt d'éclairer son choix. Ces méthodes comprennent deux phases : celle de la construction de la relation de surclassement S et celle de l'exploitation de cette dernière, parmi ces méthodes nous pouvons citer la méthode ELECTRE TRI et PROMETHEE. Ces méthodes peuvent être utilisé pour l'évaluation des règles d'association issu d'un processus de fouille de données.

Définition 2.5 (Relation de surclassement). *On dit qu'une alternative a surclasse une alternative b et on note aSb si il y a suffisamment d'arguments pour admettre que a est au moins aussi bonne que b et qu'il n'y a pas d'arguments importants prétendant le contraire [49].*

La méthode ELECTRE TRI

Electre Tri est une méthode d'analyse multicritères, son but est de résoudre des problèmes multicritères en vue d'un tri d'affectation des alternatives à des classes bien définies. Le principe de cette méthode est d'assigner un ensemble de m alternatives (Actions) notées $A = \{a_1, a_2, a_3, \ldots, a_m\}$ sur lesquelles se base la décision à des catégories ou classes bien définies. On note $F = \{1, 2, \ldots, n\}$ l'ensemble des indices des critères. Chaque action de l'ensemble A sera évaluée par une fonction réelle, exprimant l'évaluation de l'action pour un critère donné, on note $G = \{g_1, g_2, \ldots, g_n\}$ l'évaluation de l'action pour les critères considérés [50, 51].

Les alternatives qui constituent l'objet de la décision ne sont pas comparées entre elles, mais à des seuils traduisant la frontière entre les h classes prédéfinies, notées $C = \{C_1, C_2, \ldots, C_h\}$. Chaque alternative sera comparée aux frontières de chaque catégorie, formant un profil $B = \{b_1, b_2, b_3, \ldots, b_h\}$. La Figure 2.4 illustre la problématique de tri ou d'affectation.

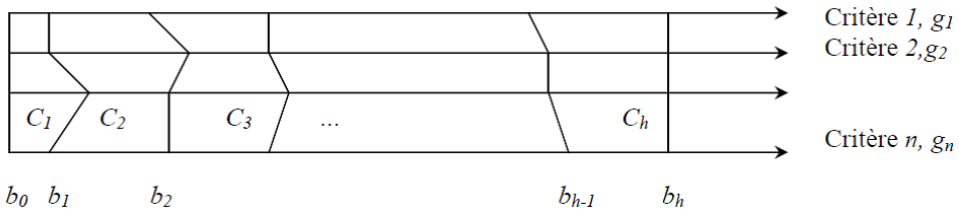

Figure 2.4 – Illustration de la problématique de Tri

L'affectation des actions dans les catégories se base sur le concept de surclassement. Une action a de l'ensemble A surclasse b_h, noté aSb_h, si a est aussi bon que b_h sur tous les critères et a n'est pas mauvais que b_h sur la majorité des critères (a_h peut-être mauvais que b_h sur certains critères).

La méthode ELECTRE-TRI s'appuie sur les étapes suivantes :

— Calcul des indices de concordance, formule 2.1 :

$$C_j(a, b_h) = \begin{cases} 0 \ si \ g_j(b_h) - g_j(a) \succeq p_j(b_h) \\ 1 \ si \ g_j(b_h) - g_j(a) \preceq q_j(b_h) \\ sinon \ \ \frac{p_j(b_h)+g_j(a)}{p_j(b_h)-q_j(b_h)} \end{cases} \tag{2.1}$$

— Calcul de l'indice de concordance global, formule 2.2 :

$$C(a_h, b_h) = \frac{\sum_{j \in F} K_j C_j(a, b_h)}{\sum_{j \in F} K_j} \tag{2.2}$$

avec
K_j : Poids de critère j
$C_j(a, b_h)$: Indice de concordance de critère j

— Calcul de l'indice de discordance $d_j(a, b_h)$, formule 2.3 :

$$d_j(a, b_h) = \begin{cases} 0 \ si \ g_j(a_h) \preceq g_j(b_h) + p_j(b_h) \\ 1 \ si \ g_j(a_h) \succ g_j(b_h) + v_j(b_h) \\ sinon \ \ \in [0, 1] \end{cases} \tag{2.3}$$

— Calcul de l'indice de crédibilité et définition de la relation de surclassement, formule 2.4 :

$$\sigma(a, b_h) = C(a, b_h) \prod_{j \in \overline{F}} \frac{1 - d_j(a, b_h)}{1 - C(a_h, b_h)} \tag{2.4}$$

avec
$\overline{F} = \{j \in F : d_j(a, b_h) \succ C(a, b_h)\}$
$C(a, b_h)$: Indice de concordance globale
$d_j(a, b_h)$: Indice de discordance

On définit l'indice de coupe λ comme le paramètre qui détermine la situation de préférence entre a et b_h. La relation de surclassement définie se base sur l'indice de crédibilité $\sigma(a, b_h)$ et l'indice de coupe λ tel que :

- $\sigma(a, b_h) \succeq \lambda$ et $\sigma(b_h, a) \succeq \lambda \Rightarrow aSb_h$ et $b_hSa \Rightarrow aIb_h$, a est indifférent de b_h
- $\sigma(a, b_h) \succeq \lambda$ et $\sigma(b_h, a) \prec \lambda \Rightarrow aSb_h$ et b_h ne surclasse pas $a \Rightarrow a$ surclasse b_h
- $\sigma(a, b_h) \prec \lambda$ et $\sigma(b_h, a) \succeq \lambda \Rightarrow a$ ne surclasse pas b_h et $b_hSa \Rightarrow b_h$ surclasse a
- $\sigma(a, b_h) \prec \lambda$ et $\sigma(b_h, a) \succeq \lambda \Rightarrow a$ ne surclasse pas b_h et b_h ne surclasse pas a, dans ce cas, a et b sont incomparables

Les procédures d'affectation
Deux procédures d'affectation sont possibles lors de l'application de la méthode Electre Tri, procédure pessimiste et optimiste.

Procédure pessimiste

- Comparer successivement a à b_i, tel que $i = p, p-1, ..., 0$
- Si aSb_h, a est assigné à la catégorie C_{h+1}

Procédure optimiste

- Comparer successivement a à b_i, tel que $i = 1, 2, ..., p$
- Si b_hSa, a est assigné à la catégorie C_h

La Méthode PROMETHEE

PROMETHEE (Preference Ranking Organization METHod for Enrichment Evaluation) est une Méthode développée par Brans [52, 53], elle a été appliquée dans plusieurs situations grâce à sa capacité de simplifier et de résoudre les problèmes de classement complexes. Elle permet de définir des relations de surclassement, d'indifférence, et d'incomparabilité entre deux scénarios du meilleur au moins bon. Cette méthode est appropriée pour traiter les problèmes multicritères du type : $max f_1(a), ..., f_n(a) | a \in A$, en respectons les étapes suivantes :

Tout d'abord, il est nécessaire de déterminer une matrice de k critères en fonction de n différentes alternatives, soit $A = \{a_1, \ldots, a_n\}$ l'ensemble de n alternatives et $j = \{f_1, \ldots, f_k\}$ l'ensemble de q critères, voir le Tableau 2.2.

Tableau 2.2 – Le tableau d'évaluation

Critères	f_1	f_2	...	f_k	...	f_q
A_1	$f_1(a_1)$	$f_2(a_1)$	...	$f_k(a_1)$	...	$f_q(a_1)$
A_2	$f_1(a_2)$	$f_2(a_2)$	...	$f_k(a_2)$	...	$f_q(a_2)$
...	...	...	...	...	...	...
A_n	$f_1(a_n)$	$f_2(a_n)$	...	$f_k(a_n)$	...	$f_q(a_n)$

$$\forall a_i, a_j \in A : d_k(a_i, a_j) = f_k(a_i) - f_k(a_j) \tag{2.5}$$

L'action a est bon que l'action b selon le critère f, si $f(a) \succ f(b)$. La fonction de préférence peut prendre des valeurs dans l'intervalle de [0, 1]. Toutes les comparaisons entre toutes les paires d'actions peuvent être réalisées pour tous les critères.

— Calcul de préférence pour chaque couple d'action, formule 2.6 :

$$\pi_k(a_i, a_j) = p_k[d_k(a_i, a_j)] \tag{2.6}$$

Avec W_k, le poids associé au critère k, $\forall k \in [1, ..., n]$.

— Calcul de la matrice de préférence, formule 2.7 :

$$\forall a_i, a_j \in A : \pi(a_i, a_j) = \sum_{k=1}^{q} W_k \pi_k(a_i, a_j) \tag{2.7}$$

Par conséquent :
$\pi(a_i, a_i) = 0$
$\pi(a_i, a_j) \succeq 0$
$\pi(a_i, a_j) + \pi(a_j, a_i) \preceq 1$

— Calcul des flux, formule 2.8 :

$$\begin{aligned} \phi^+(a_i) &= \frac{1}{n-1} \sum_{b \in A} \pi(a_i, b) \\ \phi^-(a_i) &= \frac{1}{n-1} \sum_{b \in A} \pi(b, a_i) \\ \phi(a_i) &= \phi^+(a_i) - \phi^-(a_i) \end{aligned} \tag{2.8}$$

Par conséquent :
$\phi(a_i) \in [-1, 1]$ et $\sum_{a_i \in A} \phi(a_i) = 0$

— Classement complet basé sur le flux net, formule 2.9 :

$$a_i P a_j \Longleftrightarrow \phi(a_i) \succ \phi(a_j) \quad et \quad a_i I a_j \Longleftrightarrow \phi(a_i) = \phi(a_j) \tag{2.9}$$

— Classement partiel sur la base des flux positifs et négatifs, formule 2.10 :

$$\begin{aligned} a_i P a_j &\Longleftrightarrow [\phi^+(a_i) \succ \phi^+(a_j)] \wedge [\phi^-(a_i) \preceq \phi^-(a_j)] \\ a_i P a_j &\Longleftrightarrow [\phi^+(a_i) \succeq \phi^+(a_j)] \wedge [\phi^-(a_i) \prec \phi^-(a_j)] \\ a_i I a_j &\Longleftrightarrow [\phi^+(a_i) = \phi^+(a_j)] \wedge [\phi^-(a_i) = \phi^-(a_j)] \\ &\qquad\qquad Sinon\ a_i J a_j \end{aligned} \tag{2.10}$$

Le Plan GAIA (Geometrical Analysis for Interactive Aid)

PROMETHEE GAIA permet au décideur de visualiser les principales caractéristiques d'un problème de décision. Elle permet ainsi d'identifier facilement les conflits ou les synergies entre les critères, d'identifier des groupes d'actions et de mettre en évidence des performances remarquables. Elle calcule le flux de préférence positif et négatif pour chaque alternative, le flux positif exprime la dominance d'une alternative par rapport aux autres, et le flux négatif exprime la dominance des autres alternatives par rapport à une alternative, voir la Figure 2.5.

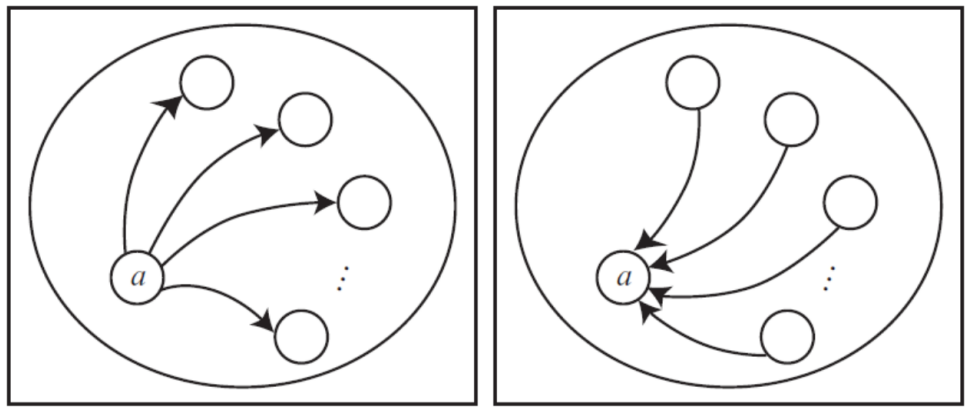

Figure 2.5 – Le classement par la méthode PROMETHEE

Le flux positif (sortant) exprime la dominance d'une action a par rapport à toutes les autres actions ; plus le flux est élevé l'action est meilleure. Le flux négatif (entrant) exprime la dominance de toutes les actions par rapport à une action a ; plus le flux négatif est bas l'action est meilleure.

Chaque alternative peut être représentée par un vecteur dans un espace de q dimensions. $\phi_1(a_i) = [\phi_1(a_i), ..., \phi_q(a_i)]$

L'analyse multicritère se présente comme une alternative aux méthodes d'optimisation classiques basée sur la définition d'une fonction unique. L'intérêt de l'AMC est de considérer un ensemble de critères de différente nature pour chercher une solution compromis qui peut prendre diverses formes : choix, affectation ou classement.

2.2 La logique floue

2.2.1 Introduction

La logique floue est une extension de la logique booléenne créée par Lotfi Zadeh en 1965 [54] en se basant sur sa théorie mathématique des ensembles flous, qui est une généralisation de la théorie des ensembles classique. En introduisant la notion de degré d'appartenance pour la vérification d'une condition, ce qui rend possible la prise en compte des imprécisions. Des chercheurs, dans différents domaines scientifiques, utilisant la logique floue, leur but principal consiste à implémenter un savoir-faire humain, sous forme d'un programme informatique. Les régulateurs flous modélisent l'expérience humaine sous forme de règles linguistiques $Si \;\; Alors$.

2.2.2 La théorie des ensembles flous

Dans la théorie des ensembles classiques, seules deux situations sont acceptables pour un élément donné, appartenir ou ne pas appartenir à un sous-ensemble. La théorie des ensembles flous proposée par Zadeh [54] introduit la notion d'appartenance afin de permettre des graduations dans l'appartenance d'un élément à un sous-ensemble. Soit X un ensemble de référence et soit x un élément quelconque de X. Un sous-ensemble flou A de X est défini comme l'ensemble des couples :

$$A = \{(x, \mu_A(x)), x \in X\} \; avec \; \mu_A : X \to [0, 1] \tag{2.11}$$

Un sous-ensemble flou A de X est caractérisé par une fonction d'appartenance $\mu(x)$ qui associe, à chaque point x de X un réel dans l'intervalle [0, 1], $\mu(x)$ représente le degré d'appartenance de x à A, ainsi on peut considérer les cas suivants :

$$\begin{cases} si \; x \; n'appartient \; pas \; à \; A \; \mu_A(x) = 0 \\ 1 \; si \; x \; appartient \; partiellement \; à \; A \; 0 \prec \mu_A(x) \prec 1 \\ si \; x \; appartient \; entirement \; à \; A \;\; \mu_A(x) = 1 \end{cases} \tag{2.12}$$

On remarque que si A est un sous-ensemble classique, la fonction d'appartenance qui lui est associée ne peut prendre que les valeurs extrêmes 0 et 1. On a dans ce cas :

$$\mu_A(x) = \begin{cases} 0 \ si \ x \notin A \\ 1 \ si \ x \in A \end{cases} \tag{2.13}$$

2.2.3 Caractéristiques d'un sous-ensemble flou

Pour pouvoir définir les caractéristiques des ensembles flous, nous redéfinissons et étendons les caractéristiques usuelles des ensembles classiques.

Le support
Le support d'un sous-ensemble flou de A de X, noté $supp(A)$ est l'ensemble de tous les éléments qui lui appartiennent. Autrement dit, c'est l'ensemble, noté :

$$Supp(A) = \{x \in X | \mu_A(x) \succ 0\}$$

La hauteur
La hauteur du sous-ensemble flou A de X, noté $h(A)$ correspond à la borne supérieure de l'ensemble d'arrivée de sa fonction d'appartenance, noté :

$$h(A) = supp_{x \in X} \mu_A(x)$$

Le Noyau
Le noyau de A est l'ensemble des éléments de X appartenant totalement à A. autrement dit, c'est l'ensemble, noté :

$$noy(A) = \{x \in X | \mu_A(x) = 1$$

La Cardinalité
La cardinalité d'un sous-ensemble flou A de X, noté A, est le nombre d'éléments appartenant à A pondéré par leur degré d'appartenance, noté :

$$|A| = \sum_{x \in X} \mu_A(x)$$

ρ-Coupe
Le $\rho - Coupe$ de A est le sous-ensemble classique des éléments ayant un degré d'appartenance supérieur ou égal à ρ, noté :

$$\rho - Coupe = \{x \in X | \mu_A(x) \succeq \rho$$

Variables linguistiques
La logique classique utilise la logique binaire. La logique floue permet d'associer une plage de valeurs (un ensemble flou) à des variables linguistiques. Une variable linguistique est une variable floue. Par exemple : La tension est haute, la variable linguistique «tension» prend la valeur linguistique «élevée». La plage de valeurs linguistiques possibles d'une règle représente l'univers de cette variable.

Fonctions d'appartenance
Chaque sous-ensemble flou peut être représenté par une fonction d'appartenance. Elle

peut être triangulaire, trapézoïdale (Figure 2.6). En général la forme de fonctions d'appartenance dépend de l'application.

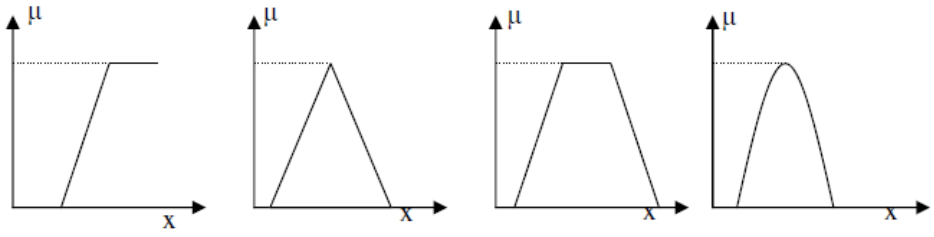

Figure 2.6 – Différentes formes de fonctions d'appartenance

Comme le montre la Figure 2.7 ci-dessous, en logique classique, une température de 22.5 °C est considérée comme élevée. En logique floue, une température de 22.5 °C appartient au groupe «moyenne» avec un dégrée d'appartenance de 0.16, et appartient au groupe «élevée» avec un degré d'appartenance de 0.75.

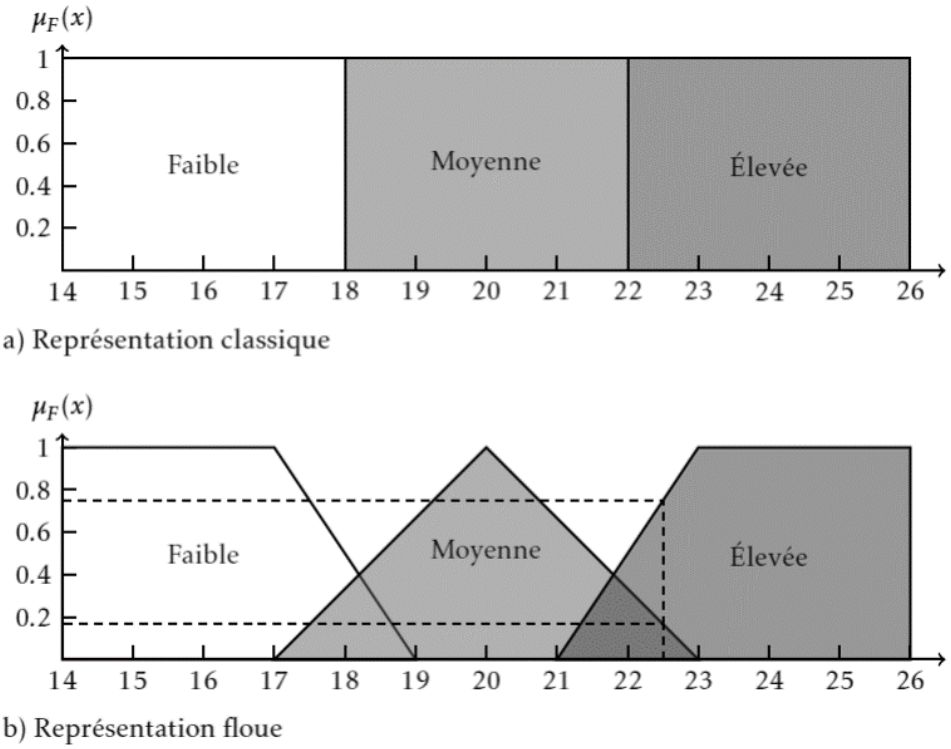

Figure 2.7 – Comparaison de l'appartenance de la température en logique classique vs la logique floue

Les variables floues faible, moyenne et élevée sont représentées par des fonctions linéaires. D'autres fonctions auraient pu être utilisées, comme des trapézoïdes, des paraboles, etc. Cependant, les fonctions linéaires sont beaucoup plus faciles à implémenter de façon pratique, et donnent de bons résultats.

2.2.4 Opérateurs de sous-ensembles flous

Afin de pouvoir manipuler aisément les ensembles flous, nous redéfinissons les opérateurs de la théorie des ensembles classiques afin de les adapter aux fonctions d'appartenance propres à la logique floue permettant des valeurs strictement entre 0 et 1. Les opérations les plus couramment utilisées sont présentées dans la Figure 2.8.

Complément
Le complémentaire d'un sous-ensemble flou A de X noté $\neg A$ est défini par :

$$\forall x \in X, \mu_{\neg A}(x) = 1 - \mu_A(x)$$

Intersection
L'intersection de deux sous-ensembles flous A et B de X est le sous-ensemble flou constitué des éléments de X affectés du plus petit des degrés avec lesquels ils appartiennent à A et B, noté :

$$\forall x \in X, \mu_{A \cap B}(x) = min[\mu_A(x), \mu_B(x)]$$

Union
L'union de deux sous-ensembles flous A et B de X est le sous-ensemble flou constitué des éléments de X affectés du plus grand des degrés avec lesquels ils appartiennent à A et B, noté :

$$\forall x \in X, \mu_{A \cup B}(x) = max[\mu_A(x), \mu_B(x)]$$

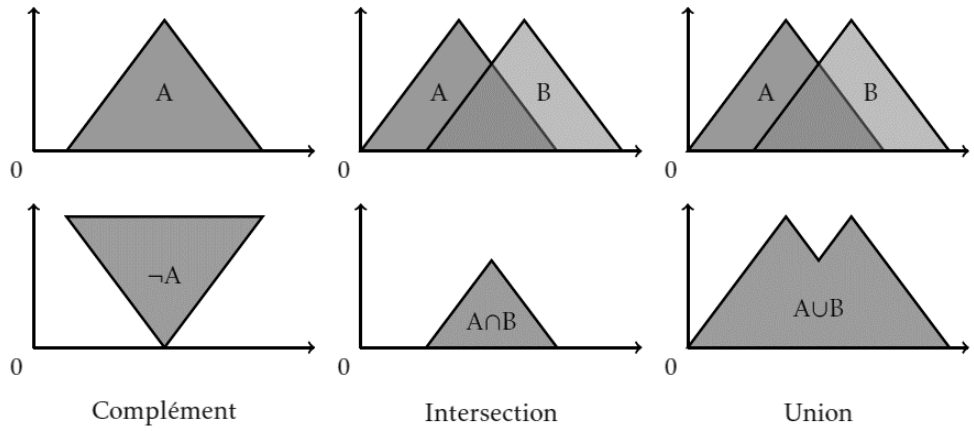

Figure 2.8 – Exemple d'opérations sur des ensembles flous

2.2.5 Processus du système flou

Un système à base de la logique floue est composé de quatre blocs principaux à savoir :

— Fuzzification : La fuzzification est l'opération de rendre une entrée classique en valeur linguistique. Des valeurs d'entrée sont traduites en concepts linguistiques représentés comme des ensembles flous.

— Base de connaissances floue : Il s'agit tout simplement de donner les règles qui lient les données aux actions. C'est ici qu'intervient tout l'intérêt de la logique floue, car ces règles sont établies par un décideur.

— Inférence floue : Les inférences lient les grandeurs mesurées et les variables de sorties par des règles linguistiques. Ces règles sont combinées en utilisant les connections et/ou.

— Défuzzication : La défuzzification est le processus de convertir une valeur floue en valeur nette. Il y a plusieurs méthodes de defuzzification proposée dans la littérature comme l'appartenance maximale, la méthode du centroïde, et la méthode des moyennes pondérées.

2.3 Conclusion

Nous avons vu, dans ce chapitre, les différents outils méthodologiques disponibles dans la littérature, qu'offrent l'aide multicritère à la décision et la logique floue. Ces outils seront utilisés dans différentes étapes de notre approche. Nous utiliserons la méthodologie présentée pour formaliser le processus de décision associé à notre problème d'extraction des règles d'association pertinentes. Nous verrons aux chapitres suivants comment ont été appliqués ces outils dans les différentes étapes de nos approches.

Deuxième partie

Contributions

Chapitre 3

Approche basée sur l'analyse multicritère pour l'extraction des règles d'association pertinentes (ERA-AMC)

«You can have data without information, but you cannot have information without data.»

Daniel Keys Moran

Ce chapitre présente notre approche pour la fouille de données et plus particulièrement l'extraction des règles d'associations pertinentes en utilisant l'analyse multicritère.

Sommaire

3.1 Introduction

Les règles d'association jouent un rôle primordial dans le processus d'extraction de connaissances dans les bases de données. Dans ce contexte, la tâche la plus difficile est l'extraction des règles d'association utiles et non redondantes, en fait, dans la plupart des cas, le jeu de données produit un très grand nombre de règles, ce qui ne permet pas aux décideurs de sélectionner les règles les plus pertinentes.

L'extraction des règles d'association est une technique qui consiste à extraire des corrélations entre les objets de base de données. Elle a été utilisée avec succès dans de nombreux domaines, à savoir le diagnostic médical, l'amélioration des processus de télécommunications, le transport, l'analyse d'images, le texte Mining, etc. Dans ce contexte, au niveau de l'extraction des règles d'association la phase finale de validation des règles d'association ne permet pas aux décideurs de choisir des plus pertinentes, vu le grand nombre de ces règles. Par conséquent, il est nécessaire d'aider le décideur dans la tâche de validation en mettant en œuvre une étape d'évaluation des règles extraites. Cette tâche doit prendre en compte à la fois les préférences des décideurs et les mesures de qualités [55, 56] pour extraire des règles d'association potentiellement intéressantes. C'est dans ce cadre, que s'inscrit cette contribution qui consiste à une proposition d'une approche basée sur l'analyse multicritère.

3.2 Approche basée sur l'analyse multicritère (AMC) pour l'extraction des règles d'association pertinentes

Aujourd'hui le monde ultra-connecté génère des volumes massifs de données stockés dans de grandes bases de données, ces données doivent être analysées afin d'extraire des connaissances utiles et valides pour l'aide à la décision. Pour ce faire, les techniques d'extraction des règles d'association se pressentent comme outils puissants, pour découvrir des corrélations et des relations entre les variables dans une base de données, ces techniques sont basées sur l'analyse statistique et l'intelligence artificielle, et utilisent des algorithmes pour trouver tous les motifs fréquents, puis générer les règles d'association en satisfaisant certains paramètres comme le support minimum et la confiance minimale. Cependant, dans la plupart des cas, ces algorithmes produisent un très grand nombre de règles extraites, ce qui ne permet pas aux décideurs de faire leurs choix des règles les plus pertinentes.

Dans ce travail, nous avons utilisé l'AMC pour le surclassement des règles d'association les plus pertinentes selon les critères proposés [57, 58]. Dans la littérature quatre types de problématiques multicritères, à savoir la sélection, le tri, l'arrangement et la description, notés respectivement $P_\alpha, P_\beta, P_\gamma,$ *et* P_δ. Vu le grand nombre de règles d'association généralement extraites, nous avons choisi une méthode qui couvre un nombre important des alternatives (règle d'association), nous avons choisi la méthode ELECTRE TRI [59] qui supporte un très grand nombre d'alternatives. Cette méthode a été appliquée dans plusieurs situations grâce à sa capacité à simplifier et à résoudre des problèmes de déci-

sion multicritère de classement. Le principe de cette méthode est d'attribuer un ensemble A de règles notées $A = \{a_1, a_2, a_3, ..., a_m\}$ à une catégorie des plus pertinentes. Notons $F = \{1, 2, ..., n\}$ l'ensemble des indices de critères. Chaque règle de l'ensemble A sera évaluée par une fonction réelle exprimant l'évaluation de la règle pour un critère donné, on note $G = \{g_1, g_2, g_3, ..., g_n\}$ l'évaluation des règles pour les critères considérés [60]. Les règles ne sont pas comparées les unes avec les autres, mais avec des seuils reflétant la limite entre les classes h prédéfinis notés $C = \{C_1, C_2, C_3, ..., C_h\}$. Chaque règle sera comparée aux limites de chaque catégorie formant un profil $B = \{b_1, b_2, b_3, ..., b_h\}$. L'affectation des règles dans les catégories est basée sur le concept de classification. Une règle a de l'ensemble A surclasse b_h noté aSb_h, si A est bon que b_h sur tous les critères.

L'intégration de l'AMC, en particulier l'utilisation de la méthode ELECTRE TRI permet de classer les règles extraites de la plus intéressante à la moins intéressante. Dans ce cadre, l'algorithme proposé est constitué de deux étapes principales, la première est l'extraction des règles d'association en utilisant l'algorithme Apriori et la deuxième consiste à l'évaluation des règles associations extraites en utilisant l'analyse multicritère, voir la Figure 3.1.

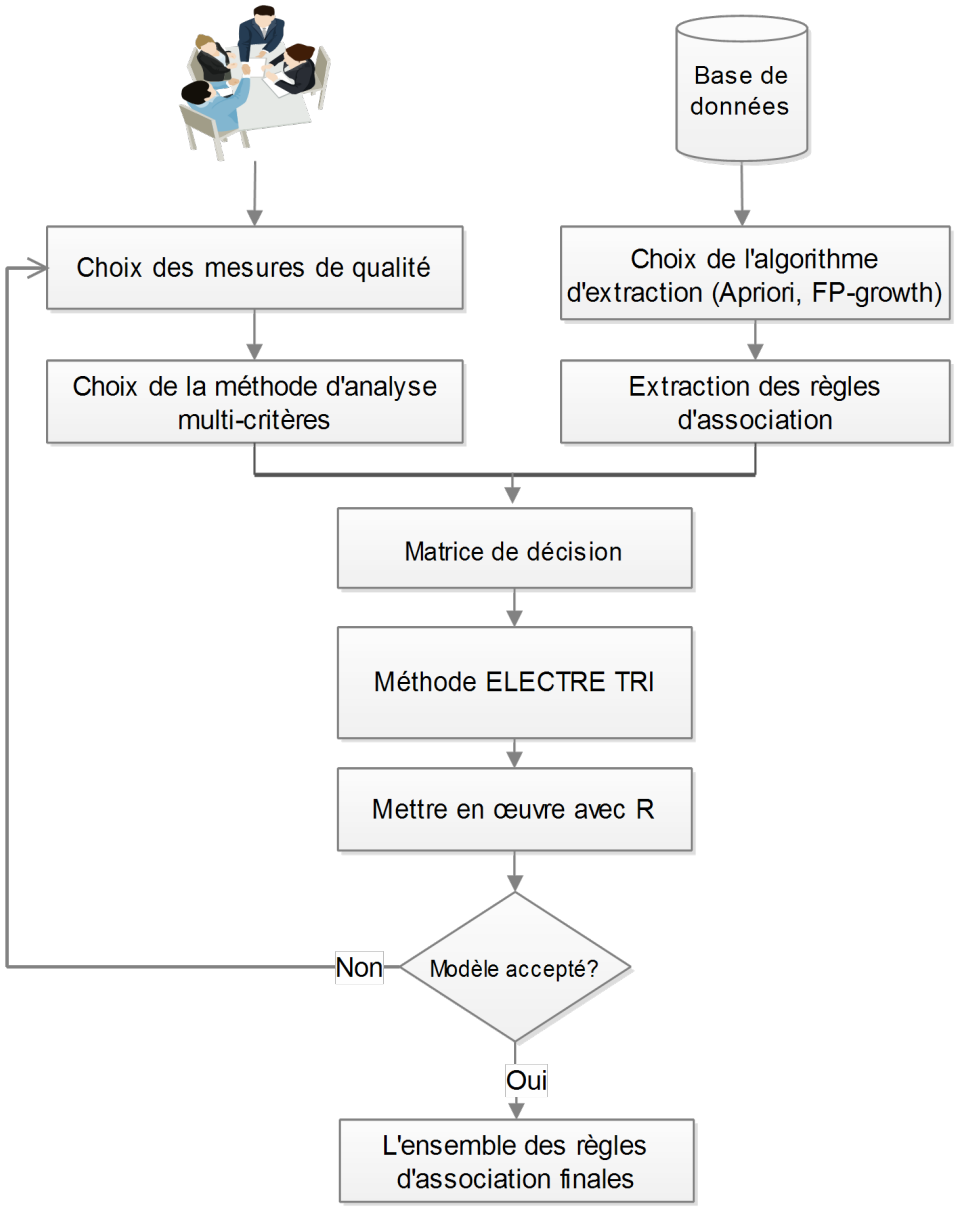

Figure 3.1 – Organigramme de l'algorithme proposé

L'approche proposée [60, 61] est décrite à travers cinq principaux modules qui sont présentés dans la Figure 3.2. Le premier est le module de préparation de données. Le deuxième présente l'extraction des itemsets fréquents à l'aide de l'algorithme Apriori. Le troisième permet de générer les règles d'association à partir des itemsets fréquents extraites dans l'étape précédente. Le quatrième décrit la visualisation interactive des règles d'association extraites. Le cinquième permet d'évaluer les règles extraites à l'aide de la méthode ELECTRE TRI. Quant au dernier, il présent les règles d'association pertinentes et prise de décision. Les détails de l'approche proposée sont présentés comme suite, Figure 3.2.

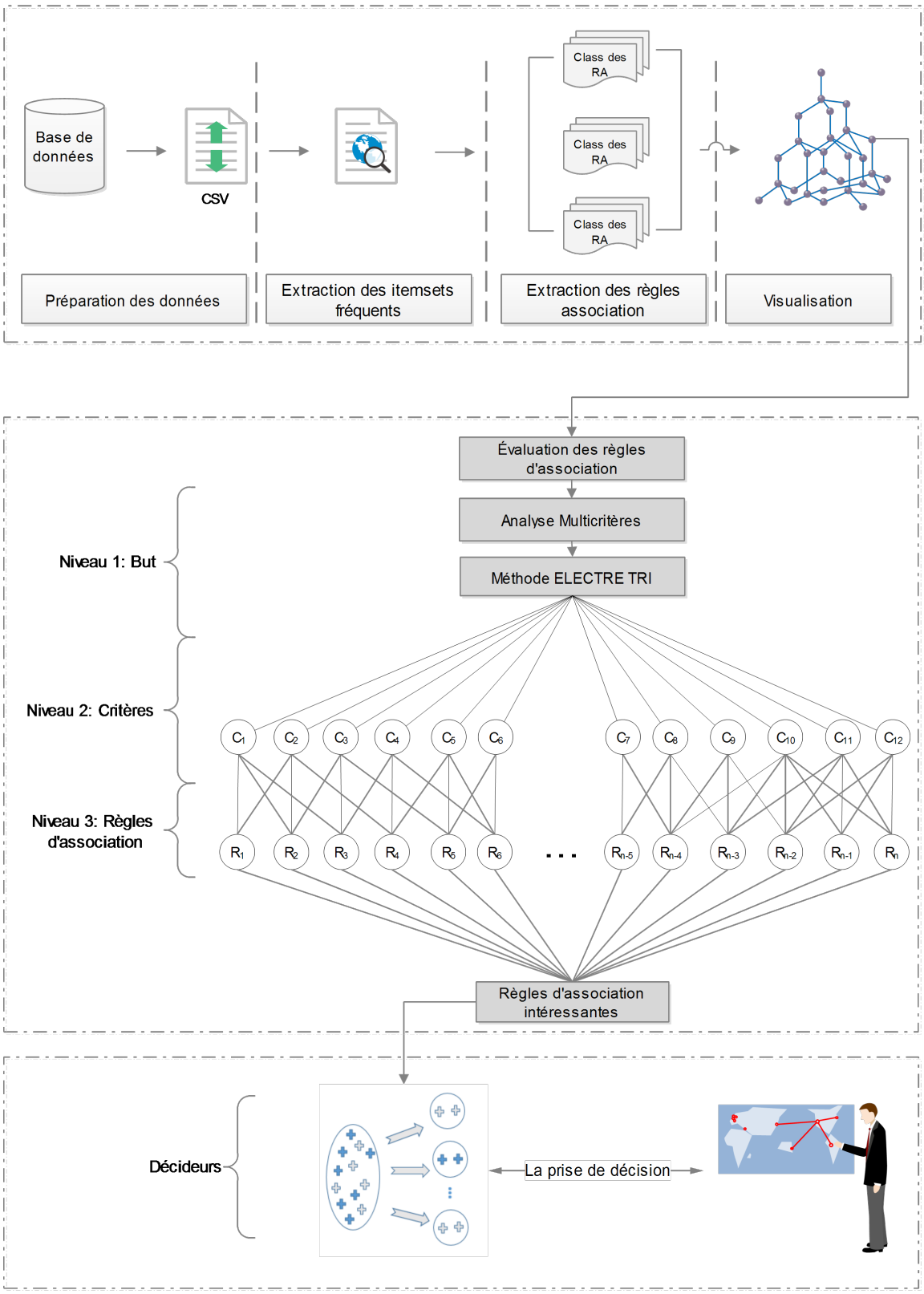

Figure 3.2 – Approche proposée

3.2.1 Préparation de données

Afin de préparer les données, nous nous référons à un ETL (Extraction Transformation Loading) pour préparer et nettoyer les données relatives aux accidents routiers en transformant les données en un format approprié et en sélectionnant seulement certains échantillons à utiliser.

Les données des accidents routiers ont été recensées auprès du Ministère de l'Équipement, de Transport et de la Logistique [12]. Les accidents de la route sont enregistrés comme des enregistrements dans la base de données, ces enregistrements comportent plusieurs attributs importants concernant les accidents de la route. Nous avons sélectionné un ensemble d'enregistrements pertinents comme entrée pour l'algorithme. Pour identifier les principaux facteurs qui affectent les accidents de la route, 18 variables ont été utilisées (voir le tableau 3.1) [61]. Ces variables décrivent les caractéristiques liées à l'accident (type de collision, usagers de la route, blessures, etc.) ; conditions du trafic (vitesse maximale, réglementation de priorité), conditions environnementales (météo, conditions de lumière, etc.) ; conditions de la route (surface de la route, obstacles, etc.) ; les conditions humaines (fatigue, alcool, etc.) et les conditions géographiques (emplacement, caractéristiques physiques).

Tableau 3.1 – Attributs et facteurs des accidents routiers

Nom d'attribut	Valeur	Description
Accident_ID	Entier	Identifiant d'accident
Accident_Type	Fatal, Blessure, Dommages à la propriété	Type d'accident
Driver_Age	$<$ 20, [21-27], [28-60] $>$ 61	Age du chauffeur
Driver_Sex	M, F	Sexe du chauffeur
Driver_Experience	$<$1, [2-4], $>$5	Niveau d'expérience du chauffeur
Vehicle_Age	[1-2], [3-4], [5-6] $>$ 7	Année du service du véhicule
Light_Condition	Crépuscule, Éclairage public, Nuit	Conditions de la lumière
Weather_Condition	Météo normale, Pluie, Brouillard, Vent, Neige	Condition de la météo
Road_Condition	Autoroute, Route de glace Route non pavée	Conditions de la route
Road_Geometry	Horizontal, Alignment, Bridge, Tunnel	Géométrie de la route
Road_Age	[1-2], [3-5], [6-10], [11-20] $>$ 20	Age de la route
Time	[00-6], [6-12], [12-18],[18-00]	Le temps
City	Marrakesh, Casablanca, Rabat...	Nom de la ville où un accident s'est produit
Particular_Area	Ecole, Marché, Boutique...	Local où l'accident s'est produit
Season	L'automne, Printemps, Été, Hiver	Saisons de l'année
Accident_Causes	Effets de l'alcool, Fatigue, Perte de contrôle Vitesse, Poussé par un autre véhicule, Frein	Causes de l'accident
Number_of_injuries	1, [2-5], [6-10],$>$ 10	Nombre de blessures
Number_of_death	1, [2-5], [6-10],$>$ 10	Nombre de décès
Victim_Age	$<$ 1, [1-2], [3-5] $>$ 5	Age de victime

Le modèle de données utilisé (voir 3.3) présente une description des données relatives aux accidents, et contient des attributs pertinents liés aux accidents de la route.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Accident_Type	Drive_Age	Drive_Sex	Drive_Exp	Vehicle_Age	vehicle_Type	Light_Condition	Weather_Condition	Road_Condition	Road_Geometry	Time	Season	Day	Causes
2	Fatal	<20	M	<1	<2	Car	Day	Clear	Collapse road	Horizontal	[6-12]	Spring	Md	Loss of Control
3	Injury	[21-27]	F	>6	<5	Car	Day	Run	Highway	Crossing	[12-18]	Summer	S	Alcohol effects
4	Injury	[28-60]	F	>7	<10	Car	Night	Clear	Collapse road	Alignment	[18-00]	Autumn	W	Speed
5	Injury	>60	F	<1	<15	Car	Day	Run	Highway	Horizontal	[12-18]	Summer	Sa	Speed
6	Injury	<21	F	<2	<10	Truck	Day	Clear	Unpaved road	Alignment	[12-18]	Summer	T	Brake Failure
7	Injury	[21-27]	F	<3	<5	Car	Day	Wind	Highway	Alignment	[6-12]	Winter	Md	Speed
8	Property damage	[28-60]	M	[2-6]	<15	Car	Day	wind	Collapse road	Horizontal	[12-18]	Summer	T	Loss of Control
9	Injury	<21	F	[2-6]	<10	Truck	Day	wind	Unpaved road	Alignment	[12-18]	Autumn	S	Speed
10	Injury	[21-27]	F	[2-6]	<5	Truck	Day	Clear	Highway	Alignment	[12-18]	Summer	W	Pushed by another vehicle
11	Injury	[28-60]	F	[2-6]	<15	Pedestrian	Day	Clear	Collapse road	Crossing	[6-12]	Autumn	Md	Alcohol effects
12	Injury	>61	F	>6	<5	Truck	Day	Clear	Unpaved road	Alignment	[6-12]	Summer	S	Speed

Figure 3.3 – Le modèle de données

3.2.2 Extraction des itemsets fréquents

L'extraction des règles d'association est l'un des problèmes les plus intensément étudiés en termes de développement informatique et algorithmique, il constitue la technique principale pour extraire les règles d'association. Cette étape est généralement coûteuse en terme de temps en raison des parcours multiples de la base de données (voir la Figure 3.4).

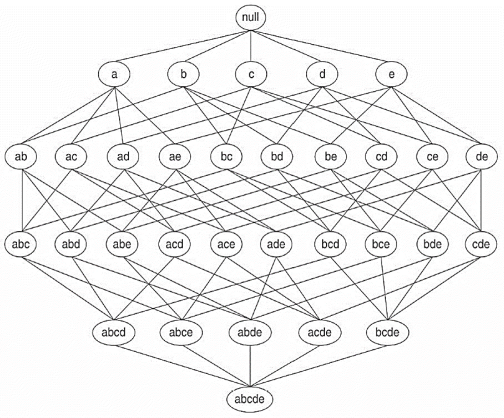

Figure 3.4 – Treillis d'un itemset

3.2.3 Extraction des règles d'association

Dans le domaine de la fouille de données, la recherche des règles d'association est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans des bases de données [4]. Elle est constituée généralement de deux étapes importantes, la première est l'extraction des itemsets fréquents, et la deuxième consiste à l'extraction des règles d'association à partir des motifs fréquents précédemment extraits. Cette tâche est coûteuse en termes de temps d'exécution et de quantité des règles extraites. Dans ce module, nous avons utilisé l'algorithme Apriori pour extraire un nombre important de règles d'association afin de montrer l'apport de l'approche proposée.

3.2.4 Visualisation des règles d'association

La visualisation est un ensemble de méthodes de représentation graphique, en deux ou trois dimensions, elles permettent aux décideurs de voir les analyses présentées visuelle-

ment afin qu'ils puissent identifier de nouveaux modèles d'une manière plus simple. Dans cette approche nous avons utilisé un ensemble de techniques de visualisation [62] à savoir la visualisation matricielle, matrice groupée, graphe, etc.

3.2.5 Évaluation des règles d'association

Les deux mesures de qualité, le support et la confiance permet l'extraction des règles d'association, mais produit un très grand nombre de règles, qui ne permet pas aux décideurs de choisir celles qui sont intéressantes. Alors, il doit y avoir d'autres mesures pour compléter le support et la confiance, ces mesures peuvent jouer un rôle clé pour filtrer les règles extraites selon des critères adaptés aux besoins de l'utilisateur.
Vu le nombre important des mesures de qualité proposée dans la littérature, nous avons choisi les mesures de support, de confiance et de lift. Elles sont utilisées comme des critères de sélection des règles d'association. Quand le décideur choisit tous les critères qui déterminent leur choix favorable, on convertit les appréciations attribuées par les décideurs à une valeur précise, et enfin, on détermine le poids de chaque critère [63].

Dans ce module, nous avons choisi la méthode ELECTRE TRI, dans laquelle nous considérons l'ensemble de règles d'association extraites comme action et l'ensemble de mesures de qualité comme critère. Le processus principal consiste à évaluer les règles d'association en utilisant la méthode choisie. Ce processus produit un ensemble de règles d'association pertinentes selon les préférences des décideurs, en cas de satisfaction, nous obtenons l'ensemble final des règles d'association pertinentes, dans le cas contraire, nous mettons à jour l'ensemble des seuils pour répondre aux préférences des décideurs.

3.2.6 Interprétation et prise de décision

Au cours de la dernière étape, les connaissances sont prêtes pour la prise de décision, les décideurs peuvent disposer de suffisamment de règles et d'informations pour prendre des décisions appropriées, telles que l'amélioration de la sécurité routière en évitant les zones dangereuses, minimiser le coût du transport, la planification des nouveaux réseaux routiers, la gestion des retards, etc.

3.3 Résultats et discussions

Après la préparation de données, nous avons sélectionné un ensemble d'enregistrements significatifs, pour identifier les facteurs liés aux accidents de la route. Ensuite, la première étape de l'approche proposée consiste à extraire les itemsets fréquents à l'aide de l'algorithme Apriori avec un seuil de support minimum égal à 0.33 (voir la Figure 3.5). Le nombre des itemsets fréquents dépend du support minimum proposé au début. Ces itemsets fréquents extraits sont utilisés à la deuxième étape, de notre approche, pour générer les règles d'associations. Pour notre cas d'étude, les règles d'association extraites sont présentées dans le tableau 3.2.

Tableau 3.2 – Les règles d'association extraites

N	Règle d'association	Support	Confiance	lift
1	$\{\} \rightarrow Light\_Condition = Day$	0.850	0.850	1.000
2	$Road\_Geometry = Horizontal \rightarrow Light\_Condition = Day$	0.300	1.000	1.176
3	$Drive\_Age = [21-27] \rightarrow Light\_Condition = Day$	0.300	1.000	1.176
4	$Day = Monday \rightarrow Light\_Condition = Day$	0.300	1.000	1.176
5	$Road\_Condition = Unpavedroad \rightarrow Light\_Condition = Day$	0.300	0.857	1.008
6	$Causes = Speed \rightarrow Road\_age = [11-20]$	0.300	0.857	1.905
7	$Victim\_Age = [2-5] \rightarrow Light\_Condition = Day$	0.300	0.857	1.008
8	$Number\_of\_injuries = 1 \rightarrow Light\_Condition = Day$	0.350	1.000	1.176
9	$Number\_of\_injuries = 1 \rightarrow Light\_Condition = Day$	0.300	0.857	1.008
10	$Time = [6-12] \rightarrow Light_Condition = Day$	0.350	1.000	1.176
11	$Road\_age => 20 \rightarrow Season = Summer$	0.300	0.750	1.364
12	$Road\_age = 20 \rightarrow Light\_Condition = Day$	0.350	0.875	1.029
13	$Accident\_Type = Fatal \rightarrow Weather\_Condition = Clear$	0.300	0.750	1.364
14	$Accident\_Type = Fatal \rightarrow Drive\_Sex = M$	0.400	1.000	1.818
15	$Drive\_Sex = M \rightarrow Accident\_Type = Fatal$	0.400	0.727	1.818
16	$Accident\_Type = Fatal \rightarrow Light\_Condition = Day$	0.350	0.875	1.029
17	$vehicle\_Type = Car \rightarrow Light\_Condition = Day$	0.350	0.778	0.915
18	$Road\_age = [11-20] \rightarrow Light\_Condition = Day$	0.350	0.778	0.915
19	$Drive\_Sex = F \rightarrow Accident\_Type = Injury$	0.450	1.000	2.000
20	$Accident\_Type = Injury \rightarrow Drive\_Sex = F$	0.450	0.900	2.000
21	$Drive\_Sex = F \rightarrow Light\_Condition = Day$	0.400	0.889	1.046
22	$Victim\_Age => 5 \rightarrow Light\_Condition = Day$	0.400	0.889	1.046
23	$Time = [12-18] \rightarrow Season = Summer$	0.450	0.900	1.636
24	$Season = Summer \rightarrow Time = [12-18]$	0.450	0.818	1.636
25	$Time = [12-18] \rightarrow Light\_Condition = Day$	0.500	1.000	1.176
26	$Number\_of\_injuries = [2-5] \rightarrow Road\_Geometry = Alignment$	0.350	0.700	1.273
27	$Number\_of\_injuries = [2-5] \rightarrow Light\_Condition = Day$	0.350	0.700	0.824
27	...	...	...	...
53	$Time = [12-18] \wedge Season = Summer \rightarrow Light\_Condition = Day$	0.450	1.000	1.176
54	$Light\_Condition = Day \wedge Time = [12-18] \rightarrow Season = Summer$	0.450	0.900	1.636
55	$Light\_Condition = Day \wedge Season = Summer \rightarrow Time = [12-18]$	0.450	0.818	1.636
56	$Season = Summer \wedge Number\_of\_death = [2-5] \rightarrow Light\_Condition = Day$	0.300	1.000	1.176
57	$Light\_Condition = Day \wedge Number\_of\_death = [25] \rightarrow Season = Summer$	0.300	0.750	1.364
58	$Weather\_Condition = Clear \wedge Road\_Geometry = Alignment \rightarrow Light\_Condition = Day$	0.300	0.857	1.008
59	$Light\_Condition = Day \wedge Road\_Geometry = Alignment \rightarrow Weather\_Condition = Clear$	0.300	0.750	1.364
60	$Weather\_Condition = Clear \wedge Season = Summer \rightarrow Light\_Condition = Day$	0.350	1.000	1.176
61	$Light\_Condition = Day \wedge Weather\_Condition = Clear \rightarrow Season = Summer$	0.350	0.700	1.273
62	$Drive\_Sex = M \wedge Season = Summer \rightarrow Light\_Condition = Day$	0.300	1.000	1.176
63	$Drive\_Sex = M \wedge Weather_Condition = Clear \rightarrow Light\_Condition = Day$	0.300	1.000	1.176
64	$Accident\_Type = Fatal \wedge Drive_Sex = M \wedge Weather_Condition = Clear \rightarrow Light\_Condition = Day$	0.300	1.000	1.176
65	$Accident\_Type = Fatal \wedge Light_Condition = Day \wedge Weather\_Condition = Clear \rightarrow Drive\_Sex = M$	0.300	1.000	1.818
66	$Accident\_Type = Fatal \wedge Drive_Sex = M \wedge Light\_Condition = Day \rightarrow Weather\_Condition = Clear$	0.300	0.857	1.558
67	$Drive\_Sex = M \wedge Light\_Condition = Day \wedge Weather\_Condition = Clear \rightarrow Accident\_Type = Fatal$	0.300	1.000	2.500

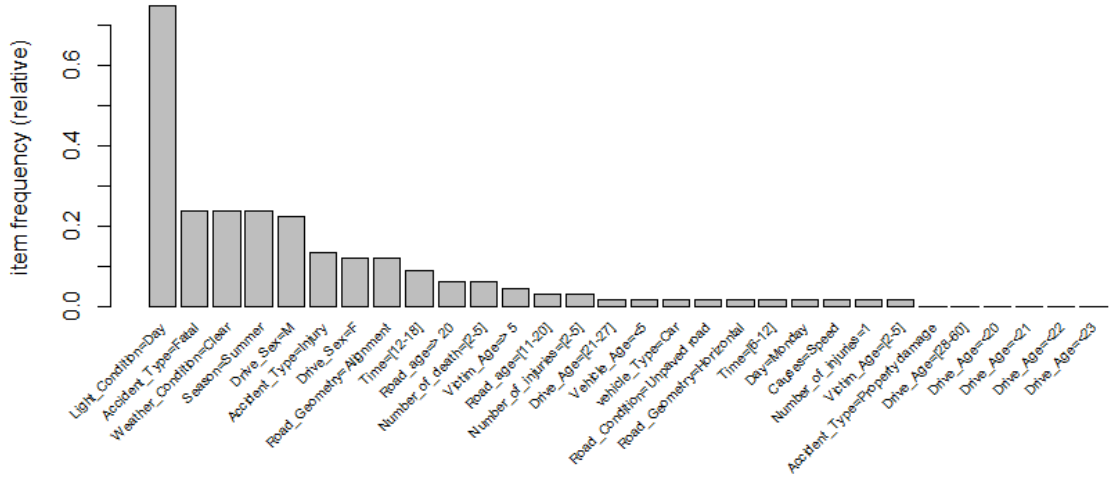

Figure 3.5 – Les itemsets fréquents

Pour visualiser les règles d'association extraites, nous avons utilisé arulesViz [62], ce paquetage propose plusieurs techniques de visualisation connues et nouvelles telles que la visualisation matricielle, graphique, matrice groupée, etc. La technique de visualisation matricielle présente les antécédents (LHS) et les conséquents (RHS) sur les axes X et Y, cette technique est renforcée par une matrice groupée en regroupant les règles extraites par clustering, la visualisation basée sur la matrice groupée est présentée dans la Figure 3.6. De plus, la visualisation graphique utilise des sommets et des arêtes, les sommets représentent généralement des objets et les arêtes indiquent une relation entre les règles, la Figure 3.7 illustre la représentation graphique des règles extraites.

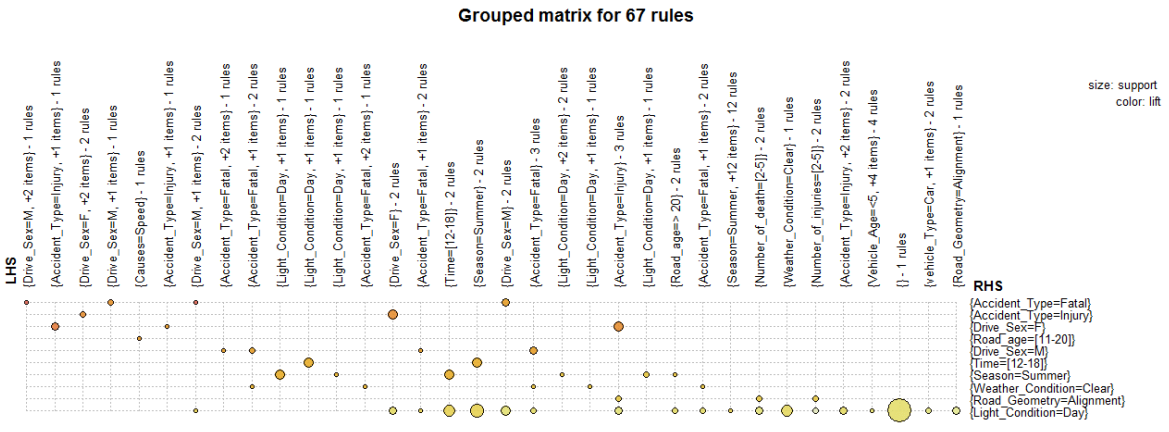

Figure 3.6 – Visualisations à l'aide de matrice groupée

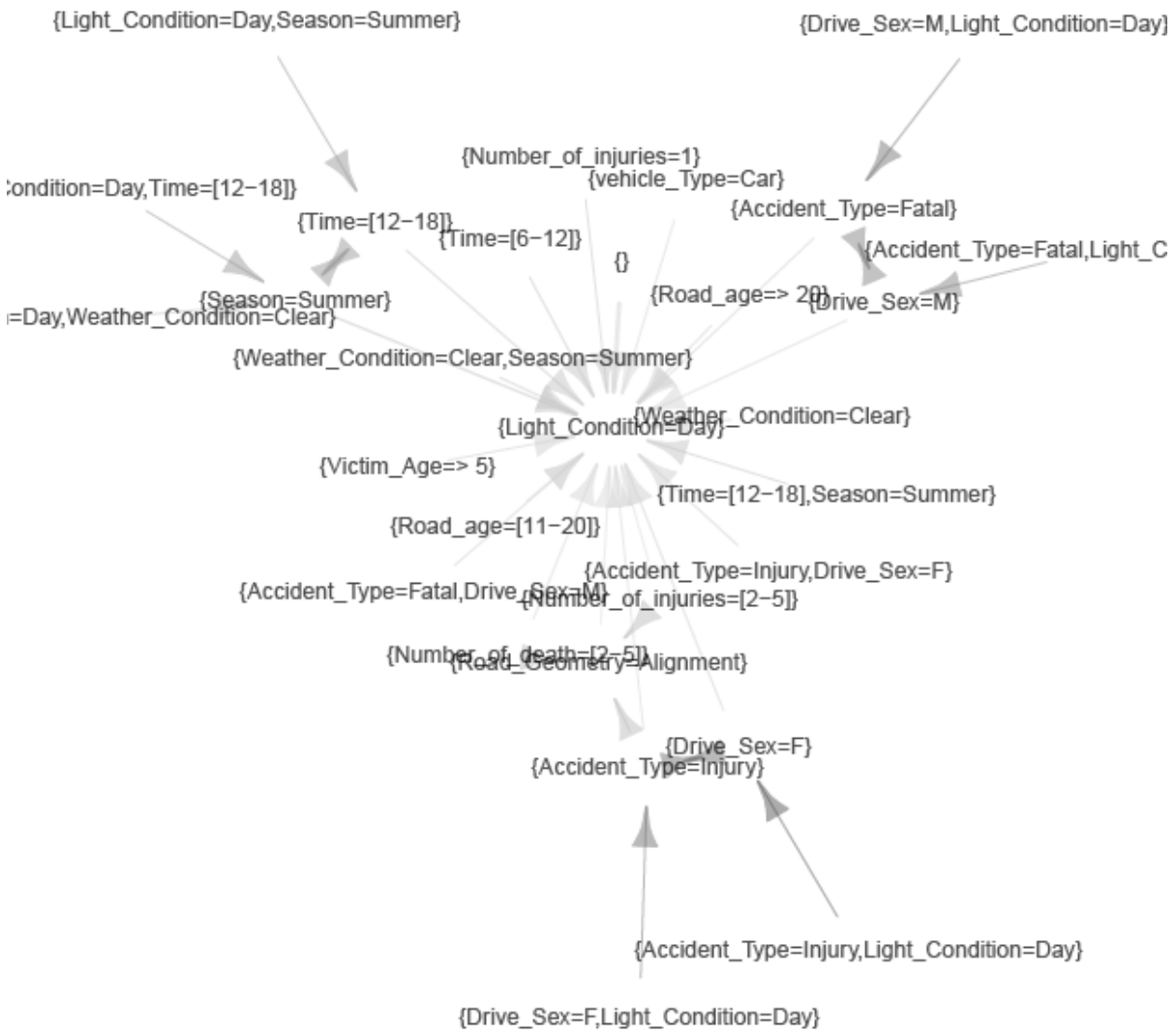

Figure 3.7 – Visualisations graphiques

L'algorithme Apriori et ses dérivés fournissent une solution substantielle à l'extraction des règles d'association. Cependant, ces algorithmes produisent un très grand nombre de règles, ce qui ne permet pas aux décideurs de choisir les règles les plus intéressantes. Pour résoudre ce problème, l'intégration d'une approche d'analyse multicritères pour le classement des règles extraites est pratiquement utile pour les décideurs [61, 60]. Dans ce contexte, nous avons choisi la méthode ELECTRE TRI, vu sa performance de supporter un très grand nombre d'alternatives et sa capacité à résoudre les problèmes complexes de tri. Nous considérons l'ensemble de règles extraites comme alternatives et le support (Cr_1), la confidence (Cr_2) et le lift (Cr_3) comme mesures de qualité des règles d'association.

Dans la première étape, les règles d'association extraites à l'aide de l'algorithme Apriori sont présentées dans le tableau 3.2, ce tableau contient 67 règles extraites dans lesquelles certaines d'entre elles sont redondantes comme les règles 20, 22 et 24, autrement dit, certaines règles ne sont pas intéressantes. Par conséquent, la prochaine étape consiste à

appliquer notre approche sur ces règles extraites en prenant en compte les préférences des décideurs. Nous avons utilisé l'ensemble des règles préalablement extraites comme alternatives à évaluer en fonction des mesures de qualités choisies. Vu le nombre important de mesures proposées dans la littérature, nous avons choisi trois mesures : le support, la confiance et le lift respectivement comme critères Cr_1, Cr_2 et Cr_3. Lorsque plusieurs options sont possibles, il est généralement utile d'analyser les avantages et les inconvénients de chacun avant de prendre la décision, un moyen efficace de mener cette analyse comparative est d'établir une matrice de décision qui traverse toutes les options disponibles avec différents critères. Cette matrice contient les préférences des décideurs à prendre en compte lors de la prise de décision. La matrice de décision est donnée dans le tableau 3.3.

Tableau 3.3 – La matrice de décision

Règle	Support	Confiance	Lift
Rule1	0.85	0.85	1.00
Rule2	0.30	1.00	1.17
Rule3	0.30	1.00	1.17
Rule4	0.30	1.00	1.17
Rule5	0.30	0.85	1.00
Rule6	0.30	0.85	1.90
Rule7	0.30	0.85	1.00
Rule8	0.35	1.00	1.17
Rule9	0.30	0.85	1.00
Rule10	0.35	1.00	1.17
Rule11	0.30	0.75	1.36
Rule12	0.35	0.87	1.02
...	...	...	...
Rule63	0.30	1.00	1.17
Rule64	0.30	1.00	1.81
Rule65	0.30	0.85	1.55
Rule66	0.30	0.85	2.50
Rule67	0.30	0.85	1.55

Après la définition de la matrice de décision, l'étape suivante consiste à définir l'ensemble des seuils des profils qui vont être comparés avec les règles extraites (Tableau 3.4) et l'ensemble des seuils de préférence, d'indifférence, de veto et de poids des critères (Tableau 3.5).

Tableau 3.4 – Définition des profils

Profile	Support	Confiance	Lift
b_1	0.5	1.0	1.2
b_{12}	0.4	0.9	1.0

Tableau 3.5 – Définitions des poids et les seuils de préférence, l'indifférence et le veto

Profile	Support	Confiance	Lift
$weight(k_j)$	0.5	1.0	1.2
$q_j(b_1)$	0.4	0.9	1.0
$p_j(b_1)$	0.5	1.0	1.2
$v_j(b_1)$	0.4	0.9	1.0
$q_j(b_2)$	0.5	1.0	1.2
$p_j(b_2)$	0.4	0.9	1.0
$v_j(b_2)$	0.5	1.0	1.2

Après la définition des seuils, la troisième étape est le calcul des indices de concordance $c_j(a, b_h)$ (Formule 2.1), et les indices de discordance $d_j(a, b_k)$ (Formule 2.3) pour obtenir les relations de surclassement qui fournissent une relation de préférence entre les règles par rapport aux profils. En outre, l'indice de coupe λ détermine la situation de préférence entre une règle d'association et un profil b_h, sa valeur par défaut est $\lambda = 0.76$. L'évaluation et l'affectation des règles d'association aux différentes catégories sont données dans le Tableau 3.6.

Tableau 3.6 – Affectations des règles d'association aux différentes catégories

Règle	C1	C2	C3
Rule1			×
Rule2	×		
Rule3	×		
Rule4	×		
Rule5	×		
Rule6			×
Rule7	×		
Rule8	×		
Rule9	×		
Rule10			×
Rule11		×	
Rule12		×	
...	...	...	...
Rule63	×		
Rule64			×
Rule65		×	
Rule66		×	
Rule67		×	

Le tableau 3.6 présente le résultat d'affectation des règles d'association à des catégories $C1$, $C2$ et $C3$ sachant que la catégorie la plus pertinente est $C1$. Ainsi, les règles d'association extraites indiquent que des accidents mortels sont produits principalement dans les situations suivantes :

— La première cause d'accidents la plus fréquente est la vitesse, la vitesse influence à la fois le risque d'accident et ses conséquences.

— La plupart des accidents surviennent lorsque l'éclairage existe.

— Le nombre de décès et de blessures augmente surtout en été.

Dans cette approche, l'intégration de l'analyse multicritères dans le processus d'extraction des règles d'association dans les bases de données a bien produit des connaissances utiles. Nous avons obtenu 33 règles importantes après avoir éliminé les règles non intéressantes. Le reste des règles appartient aux autres catégories moins d'intérêt que la première. Enfin, les règles les plus pertinentes sont présentées dans le tableau 3.7.
Il existe dans la littérature des travaux riches [64, 65] dans lesquels les auteurs ont appliqué des techniques d'exploration de données pour extraire les règles d'association, les résultats ont été bien réalisés. Cependant, la taille de base de données entraîne un très grand nombre de règles d'association extraites. Le résultat de notre approche confirme non seulement une association entre différentes variables, mais montre également que l'intégration de l'AMC permet aux décideurs de faire leur propre choix de règles d'association les plus pertinentes selon leurs préférences.

L'intégration de l'analyse multicritères dans le processus d'extraction des règles d'association a contribué globalement à l'amélioration de qualité des règles extraites, et une meilleure compréhension de la dynamique des accidents routiers et pourrait fournir des informations significatives qui peuvent aider les décideurs, à améliorer la sécurité routière. L'approche proposée présente les avantages suivants :

— Extraction et visualisation des règles d'association pertinentes ;

— Mesure de qualité des règles d'association ;

— Réduction de nombre des règles extraites ;

— Analyse des accidents routiers ;

— Amélioration de la sécurité routière.

Tableau 3.7 – Les règles d'association pertinentes

N°	Règle
Rule2	$Road\_Geometry = Horizontal \rightarrow Light\_Condition = Day$
Rule3	$Drive\_Age = [21 - 27] \rightarrow Light\_Condition = Day$
Rule4	$Day = Monday \rightarrow Light\_Condition = Day$
Rule5	$Road\_Condition = Unpavedroad \rightarrow Light\_Condition = Day$
Rule7	$Victim\_Age = [2 - 5] \rightarrow Light\_Condition = Day$
Rule8	$Number\_of\_injuries = 1 \rightarrow Light\_Condition = Day$
Rule9	$Vehicle\_Age =< 5 \rightarrow Light\_Condition = Day$
Rule11	$Road\_age = 20 \rightarrow Season = Summer$
Rule12	$Road\_age = 20 \rightarrow Light\_Condition = Day$
Rule13	$Accident\_Type = Fatal \rightarrow Weather\_Condition = Clear$
Rule16	$Accident\_Type = Fatal \rightarrow Light\_Condition = Day$
Rule17	$vehicle\_Type = Car \rightarrow Light\_Condition = Day$
Rule18	$Road\_age = [11 - 20] \rightarrow Light\_Condition = Day$
Rule24	$Season = Summer \rightarrow Time = [12 - 18]$
Rule26	$Number\_of\_injuries = [2 - 5] \rightarrow Road\_Geometry = Alignment$
Rule27	$Number\_of\_injuries = [2 - 5] \rightarrow Light\_Condition = Day$
Rule28	$Accident\_Type = Fatal \wedge Accident\_Cause = Sleep \rightarrow Drive\_Sex = M$
Rule29	$Accident\_Type = Fatal \wedge Drive\_Sex = M \rightarrow Accident\_Type = Fatal$
Rule30	$Accident\_Type = Fatal \wedge Drive\_Sex = M \rightarrow Accident\_Cause = Sleep$
Rule32	$Accident\_Cause = Sleep \wedge Drive\_Nationality = M \rightarrow Accident\_Type = Fatal$
Rule33	$Accident\_Type = Fatal \wedge Drive\_Nationality = M \rightarrow Accident\_Cause = Sleep$
Rule34	$Accident\_Cause = Sleep \wedge Drive\_Sex = M \rightarrow Drive\_Nationality = M$
Rule35	$Accident\_Cause = Sleep \wedge Drive\_Nationality = M \rightarrow Drive_Sex = M$
Rule38	$Accident\_Type = Fatal \wedge Light\_Cond = Day \rightarrow Drive\_Sex = M$
Rule50	$Drive\_Sex = M, Light_Cond = Day \rightarrow Drive_Nationality = M$
Rule52	$Accident\_Type = Injury \wedge Drive\_Sex = F \rightarrow Light\_Cond = Day$
Rule55	$Accident\_Type = Injury \wedge Drive_Sex = F \rightarrow Drive\_Nationality = M$
Rule56	$Drive\_Sex = F \wedge Drive\_Nationality = M \rightarrow Accident\_Type = Injury$
Rule58	$Drive\_Sex = F \wedge Light_Cond = Day \rightarrow Drive\_Nationality = M$
Rule59	$Drive\_Sex = F \wedge Drive\_Nationality = M \rightarrow Light\_Cond = Day$
Rule61	$Drive\_Nationality = M \wedge Weather_Cond = Run \rightarrow Light_Cond = Day$
Rule63	$Accident\_Type = Injury \wedge Drive\_Nationality = M \rightarrow Light\_Cond = Day$

3.4 Conclusion

De nombreux algorithmes ont été proposés pour l'extraction des règles d'association. Ces algorithmes produisent un très grand nombre de règles, ce qui ne permet pas aux décideurs de choisir les plus pertinentes. Afin de résoudre ce problème, nous avons proposé une approche basée sur l'analyse multicritères pour l'évaluation des règles extraites. Cette approche est appliquée au domaine de la sécurité routière. En effet, nous avons généré 67 règles par application de l'algorithme Apriori. Ensuite, nous avons classé ces règles en utilisant la méthode ELECTRE TRI et en considérant trois mesures de qualité, à savoir le support, la confiance et le lift. Ainsi, nous avons obtenu trois catégories (classes). Pour ce cas, nous avons constaté que la première catégorie, contenant 33 règles correspond aux règles les plus pertinentes en réponse aux préférences des décideurs. En plus, nous avons constaté que les préférences des décideurs ont un impact sur l'ordre et la sélection des règles pertinentes.

Cependant, cette approche présente des limites au niveau de la prise en compte de l'aspect spatial de données. Dans ce contexte, nous allons détailler dans le chapitre qui suit une nouvelle approche basée sur la théorie des ensembles flous pour l'extraction des règles d'association pertinentes en prenant en considération les données spatiales.

Chapitre 4

Approche basée sur la logique floue pour l'extraction des règles d'association spatiales (ERAS-LF)

«Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés.»

Tobler

Ce chapitre présente une proposition d'une approche d'extraction des règles d'association spatiales basée sur la logique floue.

Sommaire

4.1 Introduction

Aujourd'hui, des volumes massifs de données spatiales ont été générés et stockés dans les bases de données, ces grandes quantités de données nécessitent des méthodes et des outils sophistiqués pour extraire des connaissances utiles, ce qui conduit à un domaine émergent, appelé la fouille de données spatiales. Ce domaine est apparu pour répondre au besoin d'exploitation dans un but décisionnel à référence spatiale, ce domaine englobe à la fois les techniques des bases de données spatiales, et celles de la fouille de données. Toutefois, l'extraction des connaissances à partir des ensembles de données spatiales reste difficile par rapport à l'extraction des connaissances à partir des données numériques traditionnelles. Cette difficulté est due à la complexité des types de données spatiales, de relations spatiales et de l'auto-corrélation spatiale [66, 33].

Les techniques de la fouille de données spatiales peuvent être classées en deux grandes catégories à savoir les techniques descriptives et les techniques prédictives. Cette classification s'adapte également aux taches de fouille de données. La fouille de données spatiales descriptives décrit les données spatiales et les phénomènes spatiaux. Elle explore également les relations parmi les données spatiales ou non spatiales, et identifie le modèle de la distribution spatiale. Par ailleurs, la fouille de données spatiales prédictives, basées sur l'état actuel de données spatiales, essaie de développer des modèles pour prévoir le futur de l'état de données spatiales et prévoir la tendance du changement de ces modèles [67].

L'exploration de données spatiales est une tâche importante pour la compréhension et l'utilisation de données spatiales en découvrant des connaissances intéressantes dans les bases de données géographiques, elle se réfère à l'extraction des connaissances implicites, des relations spatiales stockées dans des bases de données [10]. Cependant, les méthodes existantes dans la littérature, par exemple la méthode de Koperski [10], Salleb (ARGIS) [7], et Marghoubi [67] présentent, les limites suivantes :

— L'utilisation de l'algorithme Apriori, pour les deux méthodes [67, 7], pour l'extraction des règles d'association spatiales qui sont très couteuses en termes de temps de calcul et de l'espace mémoire. Il est avantageux dans le cas d'un jeu de données faiblement corrélées. Or, dans un contexte spatial, généralement les données sont denses et fortement corrélées ;

— L'algorithme ARGIS proposé par Salleb relatif à la découverte des règles d'association spatiales considère deux et seulement deux couches thématiques à chaque comparaison. Autrement dit, il n'est pas possible d'extraire des règles regroupant plus de deux couches thématiques ;

— L'algorithme proposé par Marghoubi relatif à la découverte des règles d'association regroupant plus de deux couches thématiques produit des règles d'associations imprécises.

l'évaluation des prédicats spatiaux présente une complexité intrinsèque liée à la distance métrique entre les objets des couches thématiques. Alors, il convient de bien choisir le formalisme de représentation de la connaissance pour gérer cette imprécision. Pour ces

raisons, nous avons choisi d'utiliser la logique floue. Ainsi, nous proposons une approche basée sur la logique floue, cette approche est reposée sur un algorithme d'extraction des règles d'association spatiales pertinentes à travers trois étapes principales à savoir la préparation de données spatiales, le calcul des relations métriques et l'extraction des règles d'association.

4.2 Extraction des règles d'association spatiales

Les règles d'association spatiales sont des règles qui indiquent certaines relations d'association entre un ensemble d'attributs spatiaux et éventuellement non spatiaux d'objets géographiques, par exemple, une règle comme « la plupart des stations de gaz sont proches de l'autoroute » est une règle spatiale de la forme : $is(X, station\ de\ gaz) \rightarrow close\_to(X, Autoroute)$. Les prédicats spatiaux peuvent représenter des relations topologiques entre des objets spatiaux, tels que disjoints, intersectés, adjacents à, etc., ils peuvent également contenir des informations sur l'orientation spatiale comme gauche, nord, est, etc., ou spécifier une distance, par exemple près, à l'intérieur, se croisent, etc. [10]. Comme les règles d'association classiques, les règles d'association spatiales sont également associées à un support minimum et à une confiance minimale. Le support d'une règle peut être défini comme le nombre d'objets spatiaux qui satisfont ce modèle. La confiance d'une règle peut être définie comme la probabilité que le motif conséquent se produit si un antécédent se produit.

4.3 La logique floue

Les évaluations linguistiques de la perception humaine peuvent être incohérentes, incomplètes, vagues et même imprécises. Alors que, l'utilisation de l'intervalle de jugement serait pratiquement préférable au jugement des valeurs exactes [68].
La logique floue repose sur la Théorie des Ensembles Flous (TEF) introduite par Zadeh [69, 54]. Cette théorie mathématique étend la théorie des ensembles classique en permettant la gestion de l'imprécision et de l'erreur. Un ensemble flou A d'un univers X est caractérisé par une fonction d'appartenance $\mu_M \forall x \in X \mu_M \in [0, 1]$. L'ensemble A est défini par $A = \{(x, \mu_M(x)) | x \in X\}$. Les Nombres Flous Triangulaires (NFT) sont les formes les plus utilisées, représentés par des triplets (a, b, c) tels que $a \preceq b \preceq c$ comme montré dans la Figure 4.1.

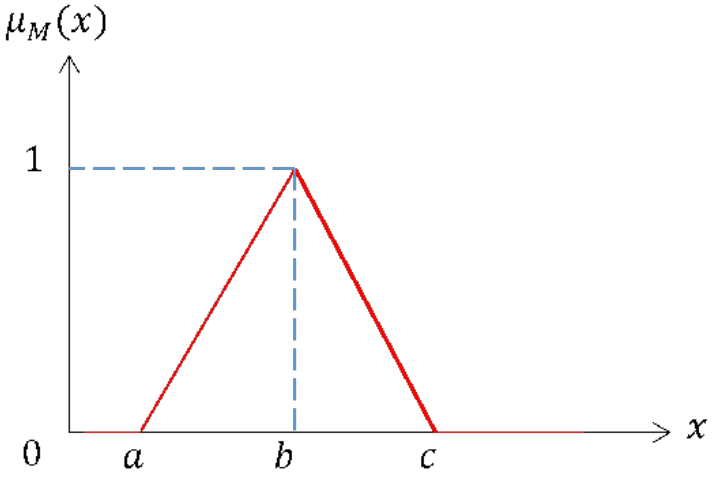

Figure 4.1 – Nombres flous triangulaire positive

La fonction d'appartenance est définie comme suite formule 4.1 :

$$\mu_M(x) = \begin{cases} 0 \ si \ x \prec a \\ (x-a)/(b-a) \ si \ a \preceq x \preceq b \\ (x-b)/(c-b) \ si \ b \preceq x \preceq c \\ 0 \ si \ x \succ c \end{cases} \tag{4.1}$$

Le terme b représente la valeur la plus possible, a et c représentent les limites inférieure et supérieure respectivement utilisées pour refléter le flou des préférences. En utilisant la TEF, le concept d'expression linguistique peut être quantifié en utilisant des nombres flous. Dans notre approche proposée, les variables linguistiques ont été considérées pour représenter les préférences des décideurs, et les nombres flous triangulaires positifs (NFT) ont été utilisés pour quantifier les variables linguistiques pour le calcul de la distance métrique entre les objets des couches thématiques.

4.4 Approche proposée

L'approche proposée dans ce chapitre repose sur l'extension de l'approche de Marghoubi [67]. En effet, nous visons à extraire des règles d'association spatiales en utilisant la logique floue. Cette approche est composée de trois modules qui sont, la préparation de base de données spatiales, la gestion de l'incertitude des préférences humaines telles que la distance entre les objets de différentes couches thématiques et l'extraction des règles d'association. Les différents modules constituant l'approche sont présentés dans la Figure 4.2 comme suit :

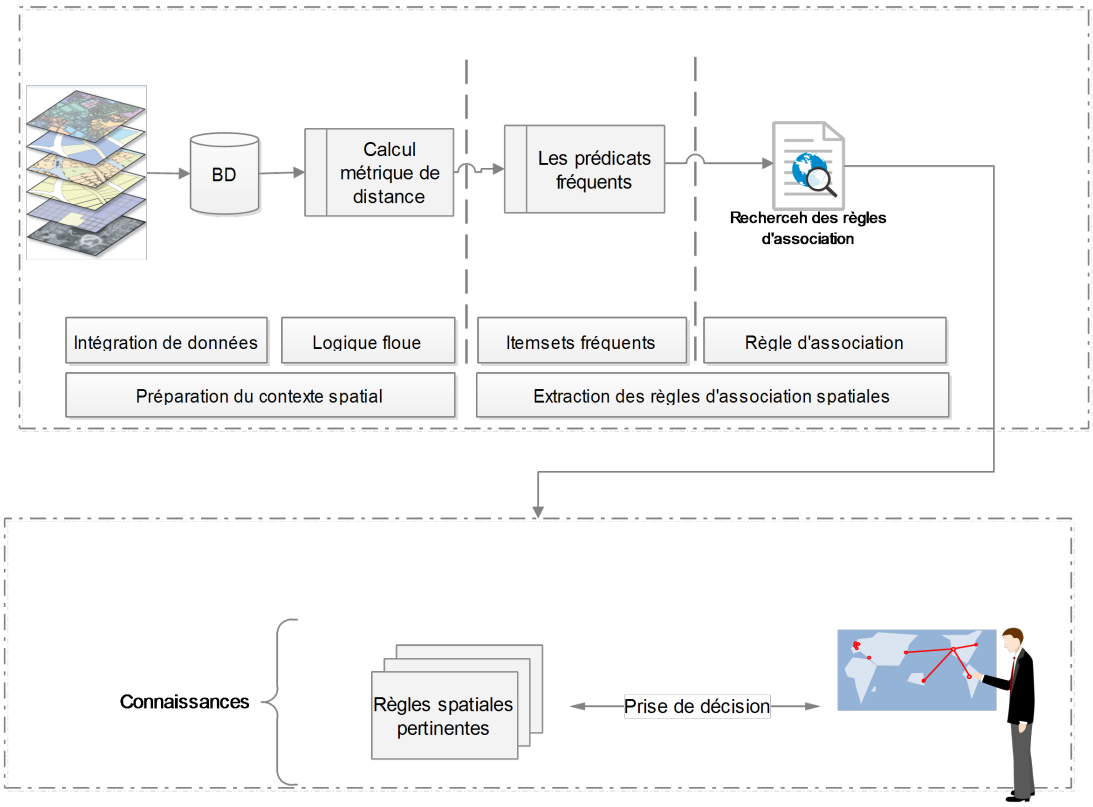

Figure 4.2 – Approche proposée

4.4.1 Préparation de base de données spatiales

Étape 1 : Intégration de données

Dans cette étape (Figure 4.3), nous nous référons à un outil ETL (Extraction Transformation Loading) pour préparer, nettoyer et transformer les sources de données en un format approprié pour l'extraction. Le contexte spatial d'extraction représente le résultat du calcul des relations spatiales de chaque objet de couche thématique avec les autres objets spatiaux en tenant compte des attributs non spatiaux.

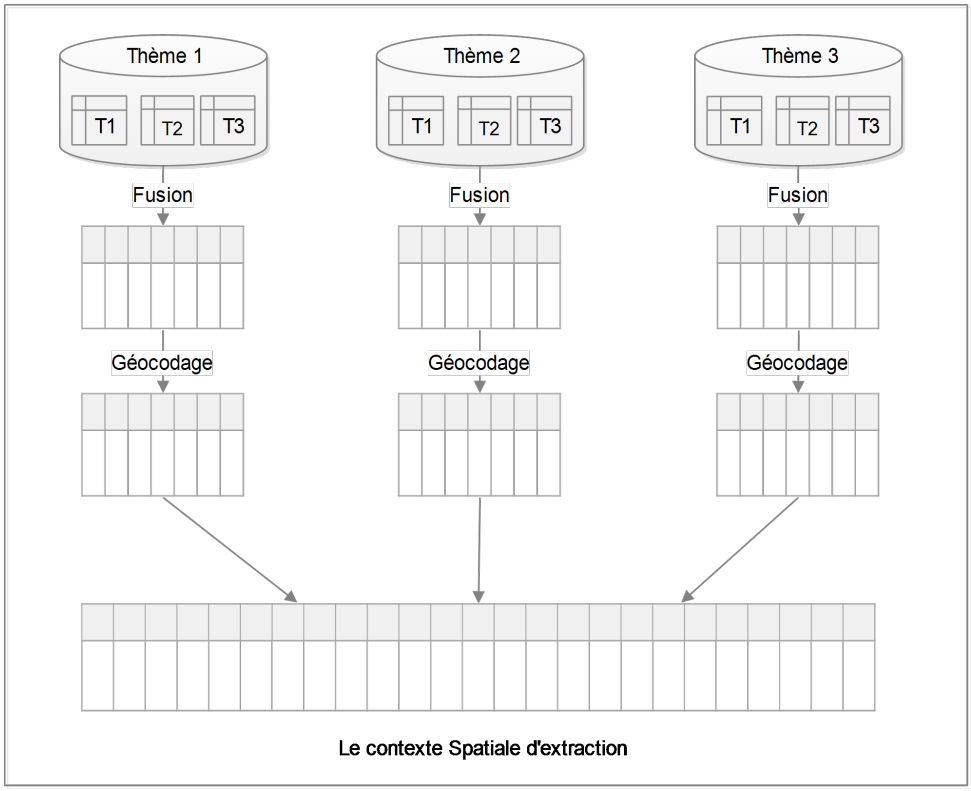

Figure 4.3 – Étapes d'élaboration de base de données spatiales

Dans cette étape nous avons considéré trois sources de données comme suite :

— *Tableau de fusion de la source de données 1 :* Soit T_r le tableau qui représente la fusion de la source de données des accidents et plus précisément, la route (catégorie Route).

— *Tableau de fusion de la source de données 2 :* Soit T_e le tableau qui représente la fusion de la source de données des institutions (École, université, faculté).

— *Tableau de fusion de la source de données 3 :* Soit T_{tr} le tableau qui représente la fusion de la source de données des territoires.

Les objets des tableaux de fusion sont codés de la manière suivante [67] :

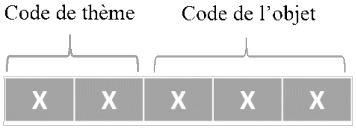

Figure 4.4 – Principes de codification

le premier caractère est réservé au code de la couche thématique, les autres sont réservés au code de l'objet, voir la Figure 4.4.

Dans cette approche, nous avons défini trois couches thématiques, la route, le territoire et les établissements (voir la figure 4.5, tableau 4.1), qui contiennent les objets spatiaux que nous avons utilisé. Ensuite, nous avons défini des prédicats flous (Tableau 4.2) en donnant la possibilité aux utilisateurs de choisir la distance entre différentes couches et d'affecter un prédicat approprié.

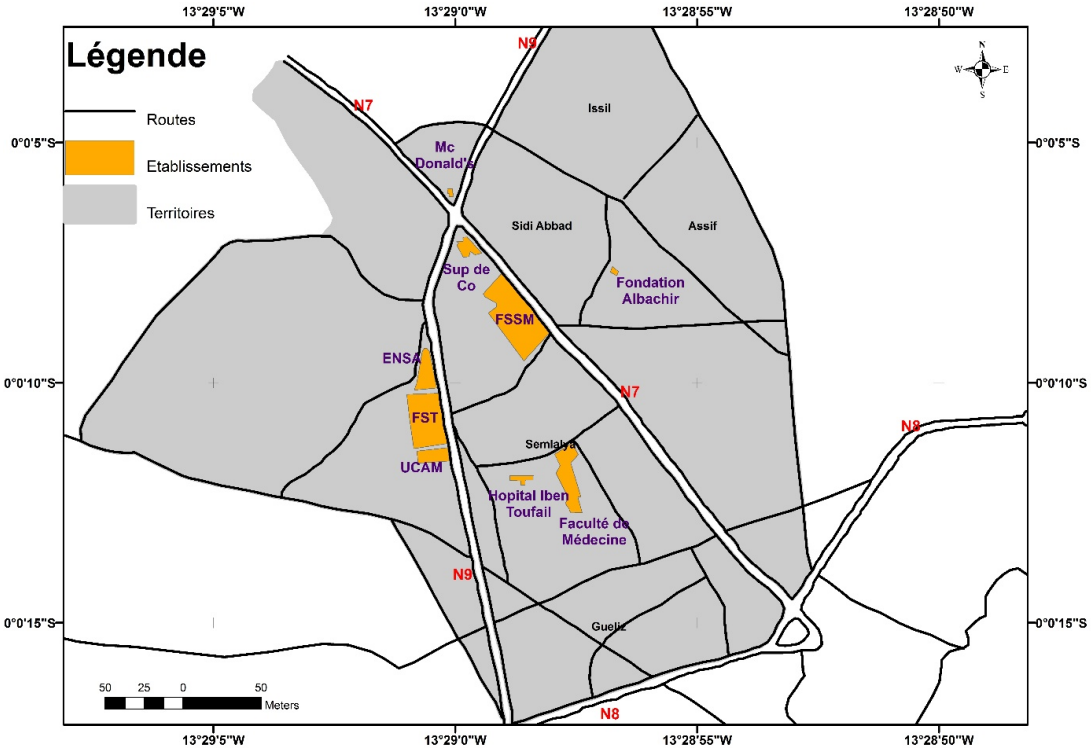

Figure 4.5 – Extrait de carte géographique de la ville Marrakech avec des couches thématiques utilisées

Tableau 4.1 – Table de codification des couches thématiques

Id	Nom	Exemple
1	Route	N7, N8, N9
2	Territoire	ISSIL, ASSIF, SEMLALIA
3	Établissement	FSSM, FST, UCAM

Les variables linguistiques ont été considérées comme des préférences des décideurs, et les nombres flous triangulaires positifs (NFT) ont été utilisés pour quantifier les variables linguistiques. Pour cela nous devons tout d'abord déterminer les prédicats spatiaux. Ensuite, ajouter les prédicats non spatiaux relatifs à chaque objet spatial. La Figure 4.3 présente le scénario pour l'élaboration de base de données spatiales de notre prototype.

Tableau 4.2 – Codification des prédicats

Id prédicat	Prédicat	Nombres Flous Triangulaires
P01	Proche de	(0, 10,15)
P02	Contient	(10, 15,20)
P03	Loin de	(15, 20,30)

Dans cette approche, nous sommes intéressés par l'exploration des données spatiales, en particulier l'extraction des règles d'association, ainsi nous présentons un algorithme pour extraire des prédicats fréquents à partir de données spatiales et non spatiales liées aux accidents routiers. Dans cette étude, les données ont été collectées à partir des rapports annuels publiés par le Ministère de l'Équipement, du Transport et de la Logistique [12, 70].

Pour identifier les principaux facteurs contribuant aux accidents routiers, nous avons utilisé 5 facteurs et 18 attributs [60, 61, 71] (voir le Tableau 4.3). Ces facteurs décrivent les attributs liés aux accidents (type, cause, temps, nombre de décès, nombre de blessures), humains (âge, sexe, expérience), véhicules (âge, type), environnements (géométrie de la route, conditions météorologiques, Âge de la route).

Tableau 4.3 – Les facteurs contribuent aux accidents de la route

Facteur	Attribut	Valeur	Description
Accident	*ID*	Entier	Identifiant
	Type	Fatal, Blessure	Type d'accident
	Accident_Causes	Effets de l'alcool, Fatigue Vitesse, Poussé par un autre véhicule, Défaillance des freins, perte de contrôle	
	Number_of_injuries	1, [2-5], [6-10],> 10	Nombre de blessures
	Number_of_deaths	1, [2-5], [6-10],> 10	Nombre de décès
	Victim_Age	$<$ 1, [1-2], [3-5] $>$ 5	Âge de la victime
	Day_and_Time	[00-6], [6-12], [12-18],[18-00]	Date
Géolocalisation	*Longitude*		
	Latitude		
Humaine	*Driver_Age*	$<$ 20, [21-27], [28-60] $>$ 61	Âge du chauffeur
	Driver_Sex	M, F	Sex du chauffeur
	Driver_Experience	<1, [2-4], >5	Niveau d'expérience du chauffeur
Véhicule	*Vehicle_Age*	[1-2], [3-4], [5-6] $>$ 7	Année de service de véhicule
	Vehicle_Type	Voiture, Camions, motocyclettes, Autre	Type de véhicule
Environnement	*Light_Condition*	Jour, crépuscule, Éclairage public, Nuit	Conditions de la lumière
	Weather_Condition	Météo normale, Pluie, Brouillard, Wind, Snow	Conditions de la météo
	Road_Condition	Autoroute, Route effondrer, Route non pavée	Conditions de la route
	Road_Geometry	Horizontal, Alignement, Pont, Tunnel	Géométrie de la route
	Road_Age	[1-2], [3-5], [6-10], [11-20] $>$ 20	Âge de la route

Étape 2 : Les prédicats spatiaux et le calcul des relations métriques
La détermination des prédicats spatiaux pour les différentes couches thématiques est faite en donnant à l'utilisateur la possibilité de choisir les distances entre différentes couches

thématiques et d'attribuer le prédicat spatial correspondant en utilisant la logique floue. Par exemple : si la distance entre les objets spatiaux de « station de gaz » et « l'autoroute » est de $20m$ alors le « station de gaz » est proche de «l'autoroute».

Exemple 4.1 :
Si la **distance** entre **accident** et **École** égale **500 m** alors l'accident est proche de l'école.
A la fin de cette étape, nous allons avoir le tableau 4.4 :

Tableau 4.4 – Détermination des prédicats

Route	Territoire	Établissement	Prédicat	Distance floue
O_{R001}	O_{T001}	O_{I001}	P01	(0, 10,15)
O_{R001}	O_{T002}	O_{I002}	P02	(10, 15,20)
O_{R003}	O_{T002}	O_{I001}	P03	(15, 20,30)
O_{R002}	O_{T003}	O_{I003}	P01	(15, 20,30)

Le calcul des différentes distances se fait de la manière suivante : pour chaque objet spatial, nous allons calculer sa distance avec les autres objets, dont les indices sont supérieurs à l'indice de l'objet choisi. L'indice représente le numéro d'ordre de l'objet dans le tableau qui regroupe les différents objets des différentes couches thématiques, voir la Figure 4.6.

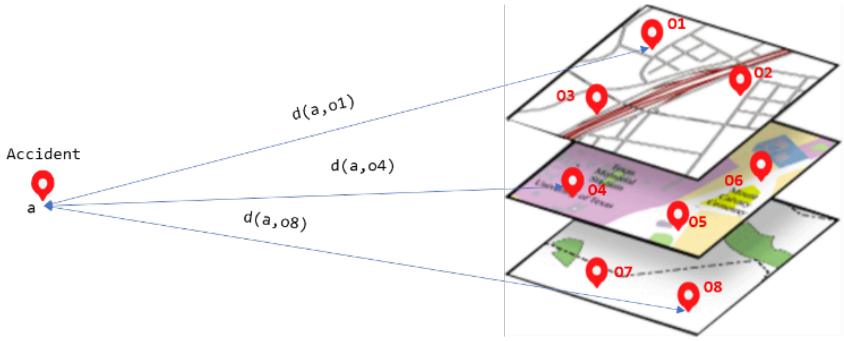

Figure 4.6 – Le calcul des distances métriques

Une fois les distances calculées, nous allons associer à chaque objet spatial dans un premier temps, les prédicats correspondants, et ce en faisant un parcours de Tableau 4.4. Ensuite nous allons choisir à partir des différentes sources de données les attributs non spatiaux répondant à l'objectif défini par le décideur, chaque couche thématique est décrite par un ensemble d'attributs spatiaux et non spatiaux (Tableaux 4.5, 4.6, 4.7).

Tableau 4.5 – Objets de couche thématique Route

ID	Forme	Géométrie	Âge
O_{R001}	Line	Horizontal	[6-10]
O_{R002}	Line	Alignement	[3-5]
O_{R003}	Line	Pont	[6-10]

Tableau 4.6 – Objets de couche thématique Territoire

ID	Forme	Population
O_{T001}	Polygone	5000
O_{T002}	Polygone	10000
O_{T003}	Polygone	15000

Tableau 4.7 – Objets de couche thématique Établissement

ID	Forme	Type	Âge
O_{R001}	Polygone	École	[6-10]
O_{R002}	Polygone	Banque	[3-5]
O_{R003}	Polygone	Faculté	[10-20]

La préparation de base de données spatiales (Figure 4.7) est constituée de quatre étapes, la première relative à la sélection des différentes couches thématiques selon le but de l'extraction, la deuxième a pour but de fusionner les sources de données de chaque couche thématique choisie. La troisième a pour but de déterminer les prédicats spatiaux selon les distances floues, et ce, en fusionnant à chaque fois pour deux couches thématiques. Enfin, le contexte spatial est généré en calculant pour chaque objet spatial ses distances avec tous les autres objets spatiaux.

L'algorithme proposé (Figure 4.7) pour cette étape est décrit comme suit : de la ligne (1) à la ligne (3) l'algorithme permet au décideur de choisir les couches thématiques selon le but de l'extraction. De la ligne (4) à la ligne (6), l'algorithme permet de fusionner les objets de différentes couches définies précédemment dans un tableau. De la ligne (7) à la ligne (10), l'algorithme permet au décideur de déterminer les prédicats spatiaux et les différentes relations spatiales entre les couches thématiques choisies. De la ligne (12) à la ligne (17), l'algorithme calcule la distance pour chaque objet spatial par rapport aux autres objets spatiaux et affecte le prédicat approprié en fonction de certains seuils de distance en utilisant la logique floue. La ligne (19) construit la base de données spatiales en faisant l'union de tous les objets spatiaux après l'affectation des prédicats spatiaux.

Entrées : Tableau des Couches Thématiques (TCT), Tableau des Prédicats Spatiaux (TPS), Tableau des Objets Spatiaux (TOS)

pour $(i \leftarrow 1, F_{k-1} \neq \emptyset; i++)$ **faire**
  $TCT \leftarrow AjouterCT(TCT(i))$;
fin
pour $(j \leftarrow 1, TCT \neq \emptyset; j++)$ **faire**
  $TOS \leftarrow IntegrationDonnee(TCT(j))$;
fin
pour $(k \leftarrow 1, TCT \neq \emptyset; k++)$ **faire**
  pour $(l \leftarrow k, TCT \succeq 0; l++)$ **faire**
    $TPS \leftarrow Predicat(TCT(k), TCT(l), DistanceFloue(TCT(k), TCT(l)), Operateur, PS)$;
  fin
fin
pour $(l \leftarrow 1, TPS \neq \emptyset; l++)$ **faire**
  pour $(m \leftarrow l+1, Taille \succeq 0; m++)$ **faire**
    $Distance \leftarrow Distance(TOS(l), TOS(m))$;
    $Predicat(TOS(l)) \leftarrow AffectationPredicat(Distance, TPS)$;
  fin
  $C_s \leftarrow C_s \cup (TOS(l), TOS(m), Distance, Predicat(TOS(l)))$;
fin
retourner C_s *(Base de données Spatiales)*

Figure 4.7 – Préparation de base de données spatiales

4.4.2 Extraction des règles d'association spatiales

Étape 1 : Extraction des prédicats fréquents
Une fois que les distances sont calculées, l'algorithme relie chaque couche thématique avec d'autres en affectant le prédicat approprié en fonction de la distance floue (Tableau 4.4). Une fois que les prédicats sont associés à chaque objet spatial, nous utilisons l'algorithme Apriori [4] (Figure 4.8). À partir de la ligne (1) à la ligne (11), l'algorithme Apriori prend en compte la base de données spatiales préalablement préparée et le support minimum pour extraire les prédicats fréquents.

Pour extraire l'ensemble des itemsets fréquents (Figure 4.9), en considérant un ensemble d'attributs non spatiaux selon les objectifs fixés par les décideurs.

Entrées : Base de donnée Spatiale C_s, Support minimum δ

$F_1 \leftarrow \{1 - itemsets\ frquents\}$;
pour $(k \leftarrow 2, F_{k-1} \neq \emptyset; k++)$ **faire**
  $C_k \leftarrow Apriori - Gen(F_{k-1})$;
  pour chaque *objet* $o \in \mathcal{D}$ **faire**
    $C_o \leftarrow Subset(C_k, o)$;
    pour chaque *condidat* $c \in C_o$ **faire**
      $c.support++$;
    fin
  fin
  $F_k \leftarrow \{c \in C_k | c.support \geqslant \delta\}$;
fin
retourner $\bigcup_k F_k$ *(Ensemble* F_k *des prédicats fréquents)*

Figure 4.8 – Extraction des prédicats fréquents

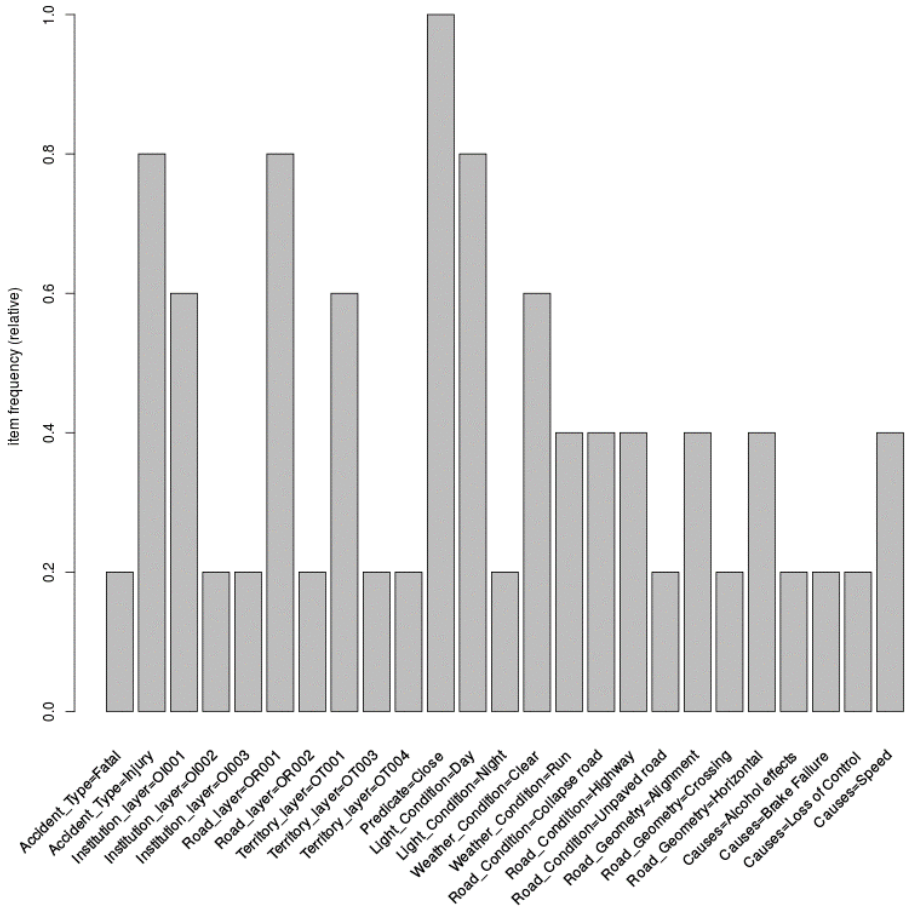

Figure 4.9 – Extrait des prédicats fréquents

Étape 2 : Extraction des règles d'association spatiales

La deuxième étape consiste à générer les règles d'association à partir des itemsets fréquents précédemment extraits en utilisant la confiance minimale (Figure 4.10). En effet, les règles extraites sont données dans la Figure 4.11, et pour la visualisation graphique nous avons utilisé arulesViz [62], voir la Figure 4.12.

Entrées : k-prédicats fréquents l_k, H_m de m-prédicat conséquences, $minconf$
Sorties : Ensemble AR_s de règles d'association spatiales valides générées à partir de l_k

si $(k \succ m+1$ **alors**
  $H_{m+1} \leftarrow Apriori - Gen(H_m)$;
  pour chaque $h_{m+1} \in H_{m+1}$ **faire**
    $confiance(r) \leftarrow support(l_k)/(l_k - h_{m+1})$;
    si $(confiance(r) \succeq minconf$ **alors**
      $AR \leftarrow AR \cup \{r : (l_k - h_{m+1}) \rightarrow h_{m+1}\}$;
      sinon
        Supprimer h_{m+1} *de* H_{m+1};
      fin
    fin
  fin
  $Gen - Rules(l_k, H_{m+1})$;
fin

Figure 4.10 – Extraction des règles d'association spatiales

	lhs		rhs	support	confidence	lift
[1]	{Institution_layer=OI001}	=>	{Light_Condition=Day}	0.6	1.0	1.25
[2]	{Territory_layer=OT001}	=>	{Road_layer=OR001}	0.6	1.0	1.25
[3]	{Weather_Condition=Clear}	=>	{Road_layer=OR001}	0.6	1.0	1.25
[4]	{Institution_layer=OI001,Predicate=Close}	=>	{Light_Condition=Day}	0.6	1.0	1.25
[5]	{Territory_layer=OT001,Predicate=Close}	=>	{Road_layer=OR001}	0.6	1.0	1.25
[6]	{Predicate=Close,Weather_Condition=Clear}	=>	{Road_layer=OR001}	0.6	1.0	1.25
[7]	{Road_layer=OR001}	=>	{Predicate=Close}	0.8	1.0	1.00
[8]	{Predicate=Close}	=>	{Road_layer=OR001}	0.8	0.8	1.00
[9]	{Light_Condition=Day}	=>	{Predicate=Close}	0.8	1.0	1.00
[10]	{Predicate=Close}	=>	{Light_Condition=Day}	0.8	0.8	1.00
[11]	{Accident_Type=Injury}	=>	{Predicate=Close}	0.8	1.0	1.00
[12]	{Predicate=Close}	=>	{Accident_Type=Injury}	0.8	0.8	1.00
[13]	{Institution_layer=OI001}	=>	{Predicate=Close}	0.6	1.0	1.00
[14]	{Territory_layer=OT001}	=>	{Predicate=Close}	0.6	1.0	1.00
[15]	{Weather_Condition=Clear}	=>	{Predicate=Close}	0.6	1.0	1.00
[16]	{Institution_layer=OI001,Light_Condition=Day}	=>	{Predicate=Close}	0.6	1.0	1.00
[17]	{Road_layer=OR001,Territory_layer=OT001}	=>	{Predicate=Close}	0.6	1.0	1.00
[18]	{Road_layer=OR001,Weather_Condition=Clear}	=>	{Predicate=Close}	0.6	1.0	1.00
[19]	{Road_layer=OR001,Light_Condition=Day}	=>	{Predicate=Close}	0.6	1.0	1.00
[20]	{Accident_Type=Injury,Road_layer=OR001}	=>	{Predicate=Close}	0.6	1.0	1.00
[21]	{Accident_Type=Injury,Light_Condition=Day}	=>	{Predicate=Close}	0.6	1.0	1.00

Figure 4.11 – Les règles d'association extraites

Figure 4.12 – Visualisation graphique des règles extraites

Considérons par exemple ces deux règles suivantes, si la distance entre deux couches (O_{I001} et O_{R001}) est égale à $10m$, on peut dire que l'accident dans la route O_{R001} est proche de O_{I001}. Mais, si la distance est égale à $12m$, on ne peut pas dire que l'accident est loin de O_{I001}, cette imprécision peut être manipulée en utilisant la logique floue. Comme nous l'avons expliqué dans les règles précédentes, l'utilisation de la logique floue peut produire deux règles différentes : si la distance entre les accidents sur la route O_{R001} est égale à $12m$, nous pouvons extraire deux règles spatiales $R1$: l'accident est proche de O_{I001} avec un pourcentage de 10% et $R2$: l'accident est loin de O_{I001} avec un pourcentage de 90%.

Les travaux de recherches précédents ont trouvé une association entre les comportements des chauffeurs et les conditions météorologiques. Cependant, l'utilisation des valeurs fixes du jugement ou les préférences humaines peuvent être vagues et même imprécises. Le résultat de cette approche confirme non seulement une association entre différentes variables, mais aussi l'intégration de la logique floue pour le calcul des prédicats spatiaux à amélioré la précision et la qualité des règles d'association spatiales extraites.

4.5 Conclusion

Dans ce chapitre, nous avons proposé une approche pour l'extraction des règles d'association spatiales. D'une part, nous avons présenté les limites des travaux proposés par Koperski [10], Salleb [7], et Marghoubi [67], ainsi, nous avons décrit notre approche pour l'extraction des règles d'association spatiales en utilisant la théorie des ensembles flous pour traiter l'incertitude de calcul des distances entre les objets de différentes couches thématiques. D'autre part, nous avons proposé un algorithme qui peut être résumé en trois étapes, la première relative à la préparation de données spatiales selon le choix de décideur. La deuxième a pour but d'extraire les itemsets et prédicats fréquents selon la valeur du support minimum. La troisième est relative à la génération des règles d'association spatiales à partir de l'ensemble des itemsets fréquents établis dans la deuxième étape.

Les méthodes utilisées pour l'extraction des itemsets fréquents présentent généralement des limites en termes de temps de calcul, et d'espace mémoire dans le contexte des bases de données spatiales. Alors, il est indispensable de faire recours au partitionnement de la base de données pour réduire le temps de réponse des algorithmes d'extraction et traiter les bases de données volumineuses. L'approche que nous proposons par la suite se base sur les calculs distribués dans un environnement du Big Data pour effectuer des traitements complexes à grande échelle.

Chapitre 5

Extraction des règles d'association pertinentes dans les bases de données massives

«Information is the oil of the 21st century, and analytics is the combustion engine.»

Peter Sondergaard

Ce chapitre présente une proposition d'une approche d'extraction des règles d'association dans le contexte du Big Data. Dans un premier lieu, nous rappelons la problématique d'extraction en termes de temps de réponse et l'espace mémoire. Nous présenterons par la suite notre approche et son application dans le domaine de la sécurité routière.

Sommaire

5.1 Introduction

La fouille de données se base sur plusieurs techniques pour l'extraction de connaissances à partir de données, ces connaissances ont un grand intérêt pour les décideurs, puisqu'elles sont non identifiables par les méthodes classiques d'analyse. En générale, les techniques utilisées sont issus de l'intelligence artificielle et les statistiques. Dans ce chapitre, nous nous intéresserons aux techniques d'exploration des règles d'association dans le contexte du Big Data. En effet, cet extraction vise à rechercher les éventuelles corrélations entre les données. Après avoir fait ses preuves en Data Mining, les techniques d'extraction des règles d'association se sont vues introduite en Big Data pour explorer les bases de données massives et tirer profit des nouvelles connaissances.

Comme son expression l'indique, le Big Data se caractérise par la taille ou la volumétrie de données. Mais d'autres attributs, notamment la vélocité, la variété et la valeur, sont aussi à considérer. En ce qui concerne la variété, le Big Data est souvent rattaché au contenu non structuré ou semi-structuré, ce qui peut représenter un défi pour les environnements classiques de stockage et de calcul. Les données non structurées et semi-structurées sont partout : contenu web, posts twitter ou commentaires client en format libre. Par vitesse on entend la rapidité avec laquelle les informations sont créées. Grâce à ces nouvelles technologies, il est maintenant possible d'analyser et d'utiliser l'importante masse de données fournie par les fichiers logs des sites web, l'analyse d'opinions des réseaux sociaux, et même les vidéos en streaming et les capteurs environnementaux. Le Big Data vise à proposer une alternative aux solutions traditionnelles de bases de données et d'analyse (serveur SQL, plate-forme de Business Intelligence).

Dans la littérature, de nombreux algorithmes ont été conçus pour extraire les règles d'association. Néanmoins, le nombre élevé de ces algorithmes est lui-même un obstacle à la capacité de choix d'un décideur. Dans ce contexte, nous avons mené une étude comparative basée sur la méthode ELECTRE [72, 73] pour le choix des algorithmes les plus appropriés parmi ceux proposés. Le résultat de cette étude comparative est présenté dans la Figure 5.1.

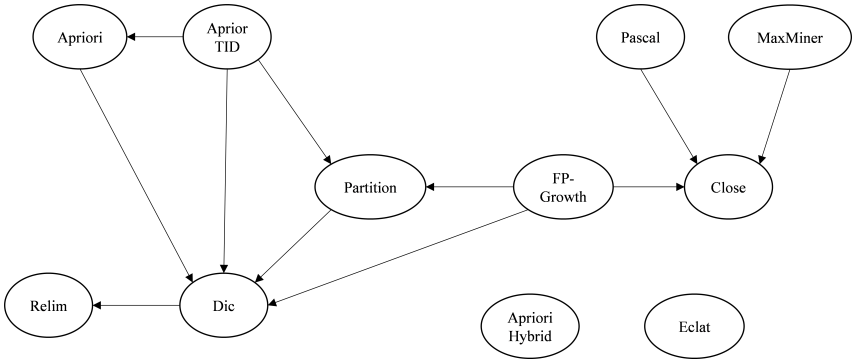

Figure 5.1 – Graphe de surclassement des algorithmes

D'après cet graphe de surclassement on remarque que les algorithmes Apriori-Hybrid et Eclat sont incomparables, par contre AprioriTID surclasse Partition, Apriori et Dic, FP-growth surclasse Close et Dic. Dic surclasse Relim, Pascal surclasse Close et MaxMiner surclasse Close. À partir de cette étude comparative, FP-growth est l'algorithme le plus approprié selon les préférences des décideurs.

Ainsi, FP-growth [24] est l'algorithme le plus efficace en terme du temps de réponse et l'espace mémoire pour l'extraction des itemsets fréquents sans passer par la génération des itemsets candidates. Malgré la performance de cet algorithme, il présente de nombreux inconvénients tels que la complexité d'extraction des itemsets fréquents dans les bases de données massives et la production d'un grand nombre de règles d'association. Pour faire face à ce défi, l'adaptation de cet algorithme d'une manière distribuée dans le contexte du Big data constitue une solution puissante pour la réduction du temps de réponse et le stockage en mémoire, de plus, la phase finale de la validation des règles permet à l'utilisateur de faire face à une difficulté majeure : comme l'extraction des règles les plus pertinentes. Par conséquent, il est nécessaire d'aider l'utilisateur dans la tâche de validation en mettant en œuvre une étape préliminaire du post-traitement des règles extraites. La tâche de post-traitement vise à réduire le nombre de règles potentiellement intéressantes en utilisant les techniques d'analyse multicritères. Cette tâche doit prendre en compte à la fois les préférences des décideurs et les mesures de la qualité des règles d'association.

5.2 Outils de traitement de données massives

De nombreux outils open source du Big Data, tels que Hadoop, Pig, Hive, et Spark sont disponibles. Celle-ci propose, contrairement aux logiciels propriétaires, de nombreux avantages : innovation continue, réduction des coûts, interopérabilité et la capabilité.

5.2.1 Apache Hadoop

Apache Hadoop [74] est un framework open source pour le stockage d'ensembles de données extrêmement volumineuses dans un environnement informatique distribué. Permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Ainsi chaque nœud est constitué de machines standard regroupées en grappes. Avec Hadoop, le Big Data est distribué en segments étalés sur une série de noeuds s'exécutant sur des périphériques de base. Au sein de cette structure, les données sont dupliquées dans différents endroits afin de récupérer l'intégralité des informations en cas de panne. Les données ne sont pas organisées par rangs ou par colonnes relationnelles, comme dans le cas de la gestion classique de la persistance, ce qui comporte une capacité à stocker du contenu structuré, semi-structuré et non structuré. Le noyau de Hadoop est constitué d'une partie de stockage : HDFS (Hadoop Distributed File System), et une partie de traitement appelée MapReduce, voir la Figure 5.2 [75].

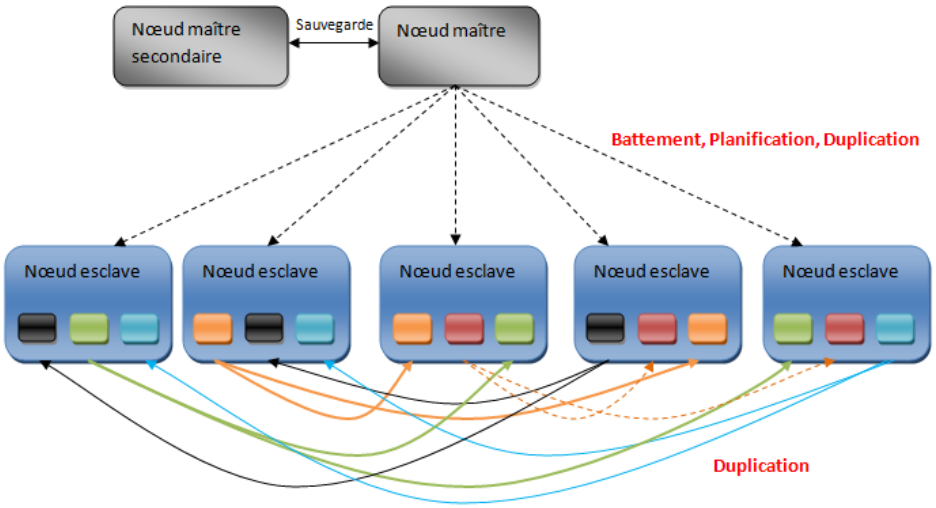

Figure 5.2 – Schéma de principe du HDFS

Hadoop fournit de nombreux sous-projets dont Hadoop Common, HDFS, MapReduce, Avro, Chukwa, HBase, Hive, Mahout, Spark, Pig et ZooKeeper, etc.

— *Apache Flume :* est un outil distribué utilisé pour collecter, agréger et déplacer l'énormes quantités de données en streaming en HDFS.

— *Apache HBase :* est une base de données distribuée open source, non relationnelle de la forme clé / valeur orientée colonne conçue pour fonctionner sur le système de fichiers distribué Hadoop (HDFS).

— *Apache Hive :* est un système d'entrepôt de données open source pour interroger et analyser de gros ensembles de données stockées dans des fichiers Hadoop.

— *Apache Oozie :* est un système de planification de workflow pour gérer les travaux de Hadoop

— *Apache Spark :* est un framework open source de calcul distribué capable de diffuser et de supporter SQL, l'apprentissage automatique et le traitement graphique. C'est un cadre applicatif de traitements big data pour effectuer des analyses complexes à grande échelle.

— *ZooKeeper :* Il s'agit d'un logiciel de gestion de configuration pour les systèmes distribués.

— *MapReduce :* est un patron d'architecture de développement informatique, inventé par Google. [76], dans lequel sont effectués des calculs parallélisés, et souvent distribués, de données potentiellement très volumineuses.

— *Mahout :* est une librairie visant à créer des implémentations d'algorithmes d'apprentissage automatique distribués.

5.2.2 MapReduce

MapReduce est un modèle de programmation conçu spécifiquement pour traiter et analyser des données massives [76]. MapReduce exécute deux fonctions essentielles : d'une part, il répartit les tâches sur plusieurs nœuds au sein du cluster (fonction Map) et, d'autre part, il organise et agrège les résultats de chacun des nœuds pour apporter une réponse à une requête, voir la Figure 5.3

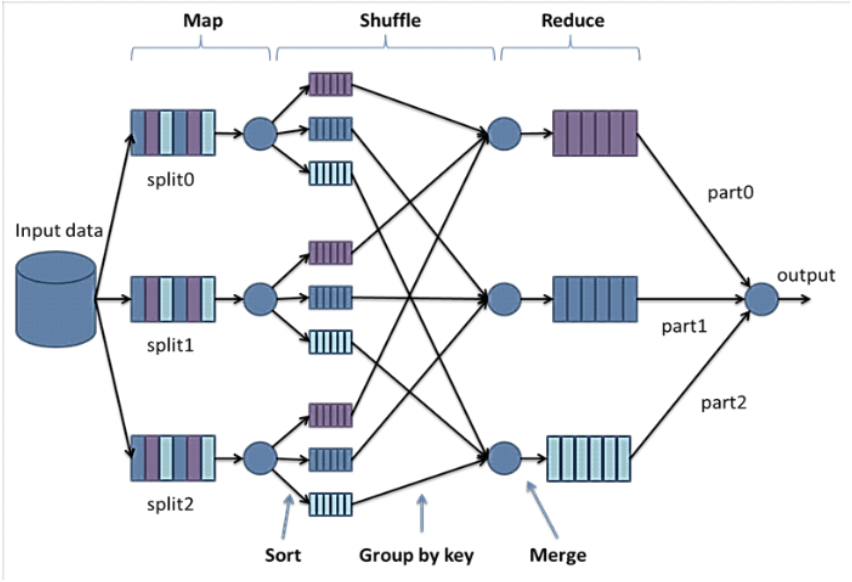

Figure 5.3 – Le Processus du MapReduce

La procédure Map effectue le filtrage et le tri. La procédure Reduce effectue une opération d'agrégation, une fois que les données d'entrée sont partitionnées en sous-données avec une taille appropriée (une taille de fichier idéale est de 64 Mo). La procédure Map prend une série des paires <clé / valeur> et génère des paires de clés / valeurs traitées, qui sont transmises à un réducteur particulier par certaines fonctions de partition. MapReduce est une excellente solution pour les calculs distribués, mais pas très efficace pour les cas d'utilisation nécessitant des calculs itératifs. Chaque étape d'un workflow de traitement étant constituée d'une phase de Map et d'une phase de Reduce, il est nécessaire d'exprimer tous les cas d'utilisation sous forme de patterns MapReduce pour tirer profit de cette solution. Les données en sortie de l'exécution de chaque étape doivent être stockées sur un système de fichier distribué avant que l'étape suivante commence. Par conséquent, cette approche est lente à cause de la réplication et de stockage sur le disque.

5.2.3 Apache Spark

Apache Spark [77] est un framework puissant pour le traitement de données à grande échelle, open source construit pour effectuer des calculs sophistiqués et conçus pour la rapidité et la facilité d'utilisation. Il a été développé à l'origine en 2009 dans AMPLab de l'Université de Californie, Berkeley, il est aujourd'hui un projet de la fondation Apache depuis 2010. Spark présente plusieurs avantages par rapport aux autres outils du big data

comme MapReduce et Storm. Apache Spark permet aux applications sur le clusters Hadoop de s'exécuter jusqu'à 100 fois plus rapide en mémoire, et 10 fois plus rapide sur le disque (Figure 5.4) par rapport à MapReduce, en plus des opérations de Map et Reduce, Spark supporte les requêtes SQL et le streaming de données et propose des librairies de machine learning et de traitements orientés graphe.

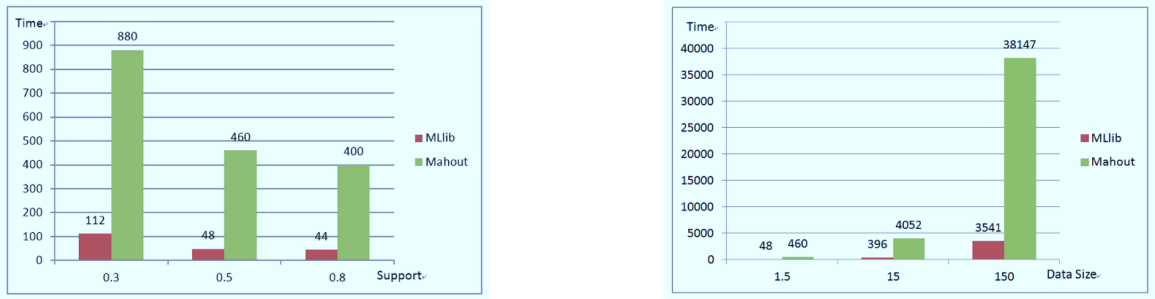

Figure 5.4 – MapReduce Vs Spark

L'architecture du Apache Spark

Spark utilise le système de fichiers HDFS pour le stockage de données. Il peut fonctionner avec n'importe quelle source de données compatible avec Hadoop, dont HDFS, HBase, Cassandra, etc. Spark peut être déployé comme un serveur autonome ou sur un framework de traitements distribués comme Mesos ou YARN. La Figure 5.5 illustre l'architecture du Spark.

L'architecture du Spark est basée sur deux abstractions principales

— Les données réparties résilientes (RDD)
— Graphe acyclique dirigé (DAG)

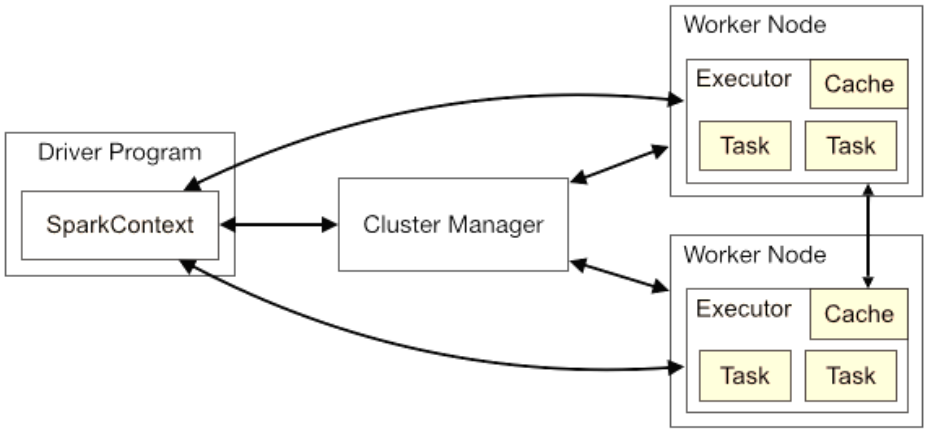

Figure 5.5 – Architecture du Spark

Écosystème du Spark

Spark est basé nativement sur le langage de programmation Scala mais des API en Python, R et en java existent. L'écosystème Spark comporte ainsi aujourd'hui plusieurs outils, voir la Figure 5.6 [78]. Chacun de ses outils apporte des fonctionnalités supplémentaires au framework, ce qui le rend attractif.

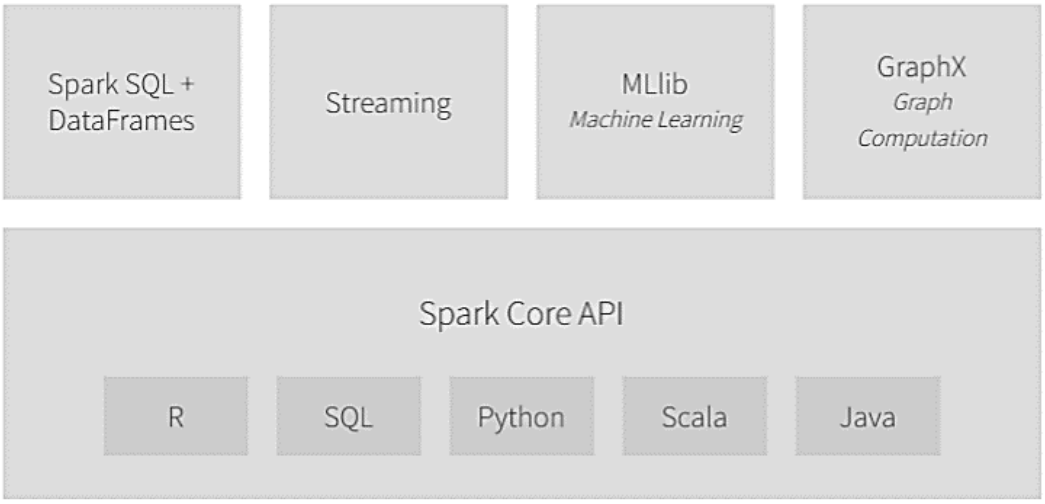

Figure 5.6 – Écosystème du Spark

— *Spark Streaming :* peut être utilisé pour le traitement en temps réel des données. Il s'appuie sur un mode de traitement en "micro batch" et utilise pour les données temps-réel DStream, c'est-à-dire une série de RDD (Resilient Distributed Dataset)

— *MLlib :* est une librairie de machine Learning qui contient un ensemble d'algorithmes et utilitaires d'apprentissage, à savoir la classification, la régression, le clustering, le filtrage collaboratif, la réduction de dimensions, etc.

— *GraphX :* est la nouvelle API pour les traitements de graphes, de plus, GraphX inclut une collection plus importante d'algorithmes pour simplifier les tâches d'analyse de graphes.

— *SparkSQL :* permet d'exposer les jeux de données Spark via API JDBC et d'exécuter des requêtes de type SQL en utilisant les outils BI et de visualisation traditionnelle. Spark SQL permet d'extraire, transformer et charger des données sous différentes formes (JSON, CSV, base de données) et les exposer pour des requêtes ad-hoc.

— *SparkR :* est un package permettant d'exploiter les fonctionnalités du Spark depuis R [79], le langage de programmation pour le traitement et l'analyse de données.

5.3 Approche proposée

5.3.1 Choix de la méthode d'analyse multicritère

Dans le contexte des bases de données massives, le nombre particulièrement important de règles générées par les algorithmes d'extraction des règles d'association ne permet pas aux décideurs de choisir les règles les plus pertinentes. La recherche des meilleures règles parmi le vaste ensemble de règles extraites, passe aussi par la recherche et l'utilisation des bonnes mesures. On se situe donc dans une problématique d'analyse multicritère. Pour faire face à cette problématique, l'intégration de l'analyse multicritères, en particulier la méthode de classement PROMETHEE, permet de classer les règles extraites par ordre des plus intéressantes aux moins intéressantes. La méthode PROMETHEE développée par Brans [52] a été appliquée dans plusieurs situations grâce à sa capacité à simplifier et à résoudre les problèmes complexes de classement.

Ainsi, nous proposons un processus de choix des règles d'association en deux étapes : Dans la première étape, nous proposons une évaluation d'un ensemble de mesures de qualités en fonction d'une liste de propriétés proposées dans [80] : en effet, certaines d'entre elles s'avèrent mal adaptées au contexte des règles d'association, tandis que d'autres s'avèrent inutiles. Dans la deuxième étape, nous suggérons d'utiliser une méthode d'analyse multicritère sur un ensemble des règles d'association extraites et l'ensemble de mesures de qualités précédemment extraites.

5.3.2 Mesures de qualité

Pour guider le décideur à identifier les règles pertinentes, de nombreuses mesures de qualités ont été proposées dans la littérature [80, 38], nous avons retenu les principales mesures de qualité des règles d'association les plus appropriées, voir le Tableau 5.1.

Tableau 5.1 – Mesures de qualité

Mesure	Formule
Support	$Supp(A \to B) = Supp(A \cup B) = P(A \cap B)$
Confiance	$Conf(A \to B) = P(B\|A) = \frac{A \cup B}{P(A)}$
Lift	$Lift(A \to B) = \frac{Conf(A \cup B)}{Supp(B)}$
Laplace	$Laplace(A \to B) = \frac{Supp(A \cap B)+1}{Supp(A)+2}$
Conviction	$Conv(A \to B) = \frac{1-Supp(B)}{1-conf(A \to B)}$
Leverage	$loevinger(A \to B) = \frac{p(A)\|p(B\|A)-p(B)}{p(\overline{B})}$
Jaccard	$Jaccard(A \to B) = \frac{Supp(A \cap B)}{Supp(A)+Supp(B)-Supp(A \cap B)}$
ϕ-coefficient	$\varphi(A \to B) = \frac{loevinger(A \to B)}{\sqrt{(Supp(A) \times Supp(B) \times (1-Supp(A)) \times (1-Supp(B))}}$

Dans le domaine de la fouille de données, les algorithmes d'extraction des règles d'association nécessitant un temps de réponse très élevé et l'espace mémoire, de plus ils

produisent un très grand nombre de règles d'association qui ne permettent pas aux décideurs de faire le choix des règles pertinentes. Pour résoudre ce problème, l'adaptation de l'algorithme FP-growth dans le contexte du big data peut réduire le temps de réponse et résoudre le problème de stockage. De plus, l'intégration de l'analyse multicritère, en particulier l'utilisation de la méthode PROMETHEE permet de classer les règles extraites des plus intéressantes aux moins intéressantes. L'algorithme proposé est constitué de deux étapes principales, la première est l'extraction des itemsets fréquents à l'aide de l'algorithme parallèle FP-growth (Figure 5.7), et la deuxième consiste à l'évaluation des règles d'associations extraites voir la Figure 5.8.

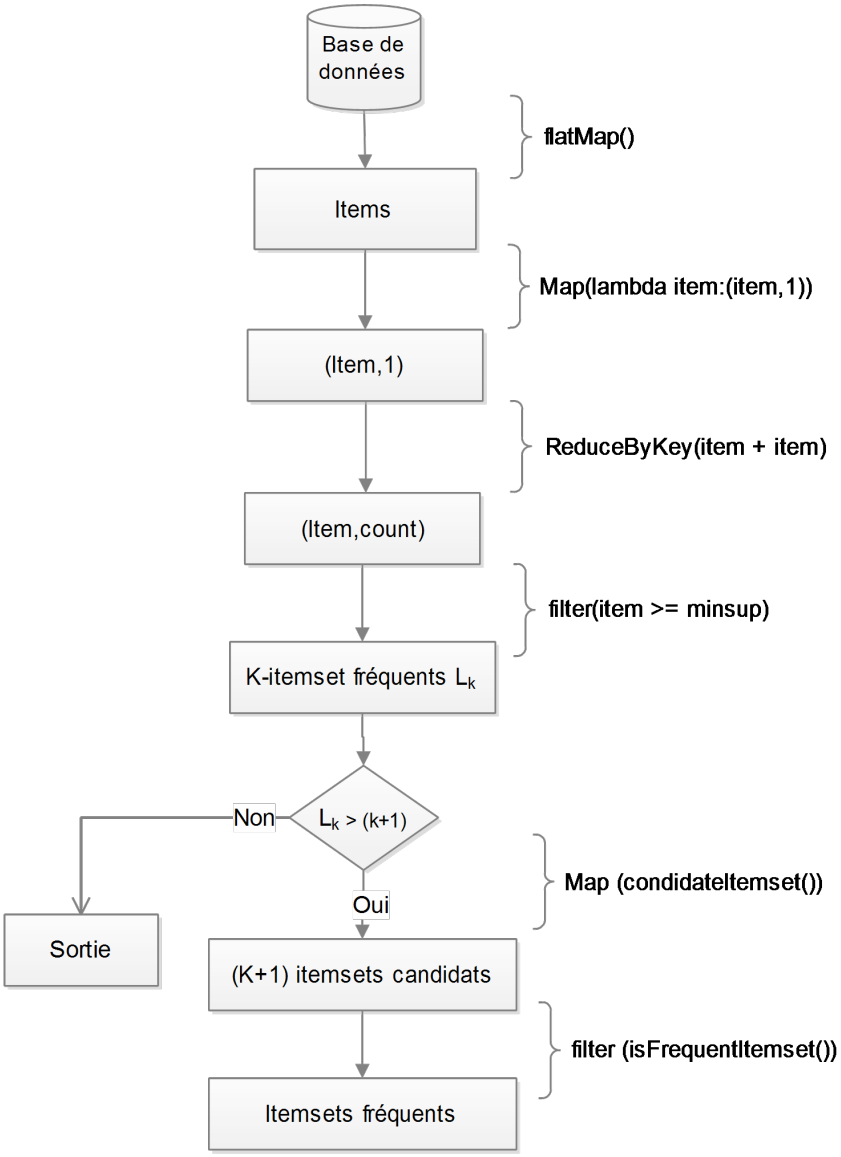

Figure 5.7 – Organigramme de l'algorithme PFP-growth

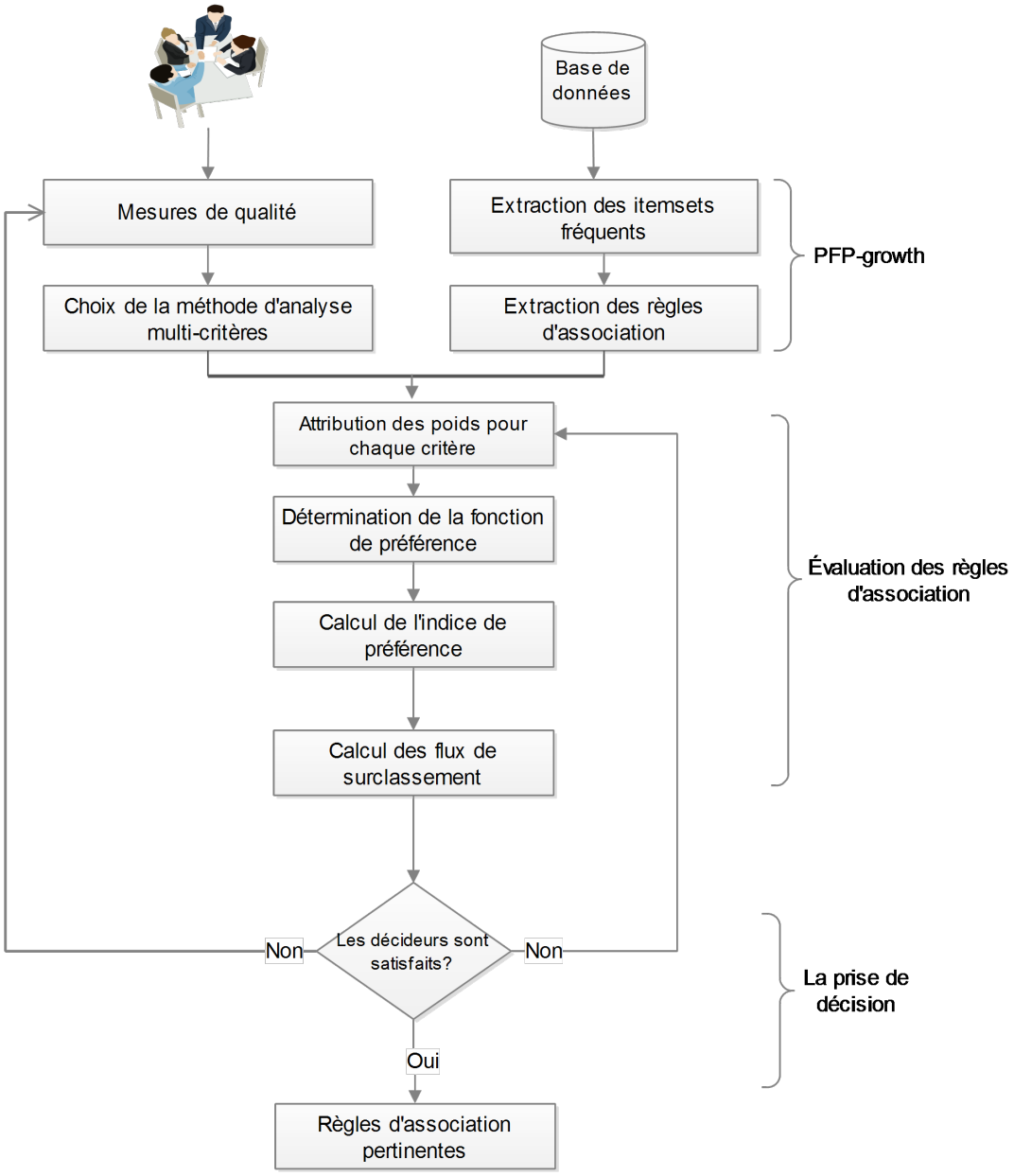

Figure 5.8 – Schéma général de l'évaluation des règles d'associations extraites

Les détails de l'approche proposée [81, 82] sont décrits par les étapes suivantes (Figure 5.9) :

Prétraitement

Dans cette étape, nous nous référons à un outil ETL (Extraction Transformation Loading) pour préparer et nettoyer les données.

Extraction distribuée des itemsets fréquents

C'est l'un des problèmes les plus intensément étudiés en termes de développement informatique et algorithmique, il constitue la technique principale pour extraire les règles d'association à partir des bases de données en introduisant le support minimum.

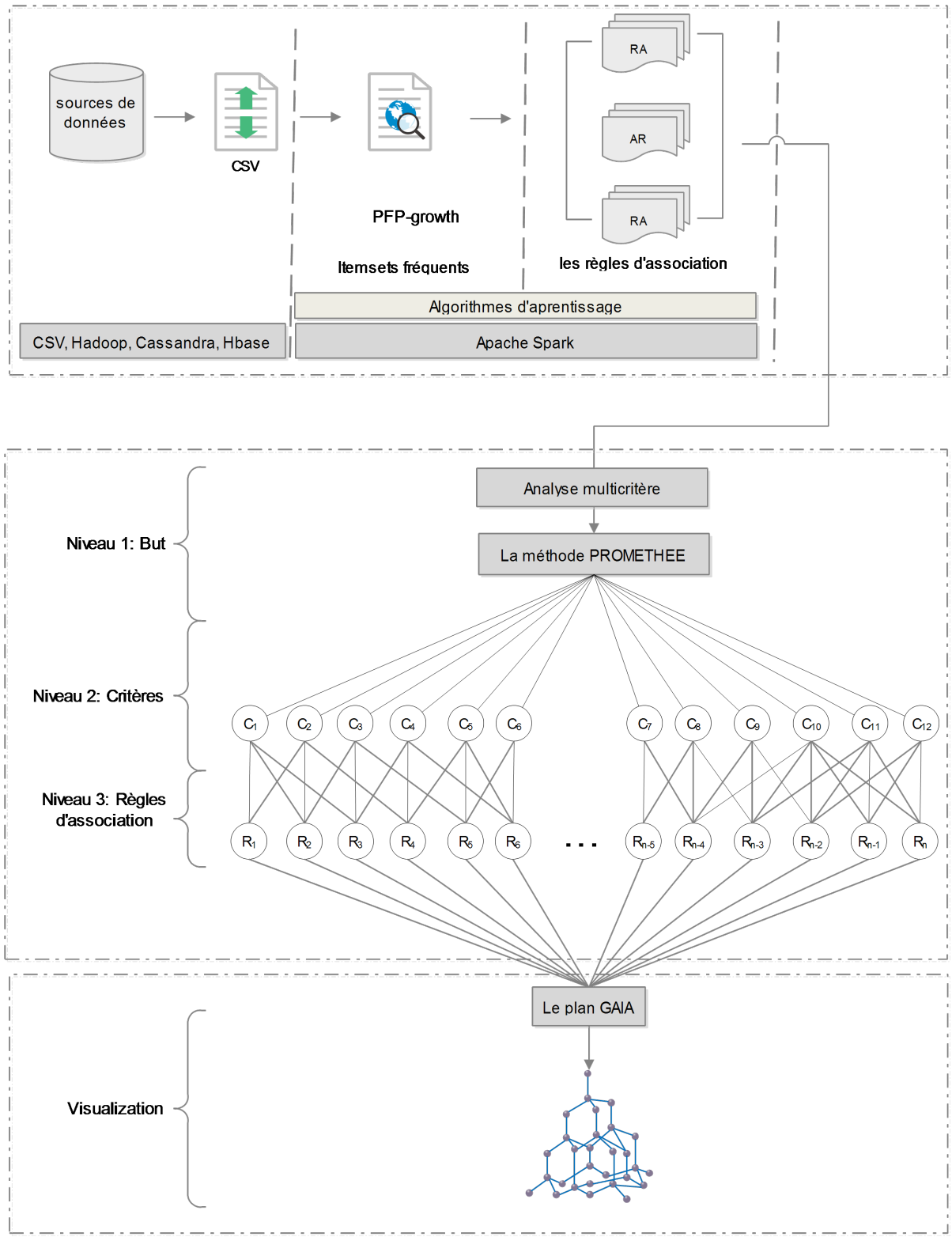

Figure 5.9 – L'approche proposée

Extraction distribuée des règles d'association

Nous avons utilisé la version distribuée de l'algorithme FP-growth dans le contexte du big data pour extraire les règles d'association pertinentes. La première étape de PFP-growth

est de calculer les fréquences des objets et d'identifier les éléments fréquents. La deuxième étape utilise une structure d'arbre du suffixe (FP-Tree) pour coder des transactions sans générer des ensembles de candidats explicitement. Ensuite, les itemsets fréquents peuvent être extraits du FP-tree en introduisant un support minimum.

Prioritisation des règles d'association extraites
Le processus d'extraction des règles d'association produit un très grand nombre de règles qui ne permettent pas aux décideurs de choisir les règles les plus intéressantes. Pour résoudre ce problème, nous avons utilisé la méthode PROMETHEE pour classer les règles extraites et prendre en compte les préférences des décideurs.

5.3.3 Résultats et discussion

Nous présentons dans ce qui suit, les résultats de l'approche proposée à travers deux étapes, la première est centrée sur l'extraction des règles d'association à partir des données relatives aux accidents routiers. La deuxième s'occupe de l'évaluation de ces règles à l'aide de la méthode PROMETHEE. Le tableau 3.1 présente les différents attributs liés aux accidents routiers.

Pour ce faire, le modèle de données est construit à partir des données et rapports annuels publiés par le Ministère de l'Équipement, du Transport et de la Logistique, il contient des attributs pertinents liés aux accidents de la route. Ce modèle peut prendre deux formats de représentations, la représentation de la forme (clé : valeur) ou la représentation binaire. Le modèle de données utilisé est donné dans la Figure 3.3.

Pour réduire le temps de réponse et l'espace mémoire lors de traitement de données très volumineuses, nous avons utilisé l'algorithme FP-growth dans un environnement du Big Data, en particulier, Apache Spark. Cet algorithme est composé de deux étapes principales la première est l'extraction distribuée des motifs fréquents, tandis que la deuxième consiste à l'extraction distribuée des règles d'association. L'environnement expérimental utilisé est donné dans le tableau 5.2.

Tableau 5.2 – Environnement expérimental

Outils	Caractéristiques
Apache Spark 2.0	Single node
Scala IDE 4.4.1	Memory : 12 Gb
Sbt 0.13	OS : Ubuntu 16.04 LTS, 64 bits
MLLIB	CPU : 2.7 GHz, i7

Extraction des itemsets fréquents
Pour extraire les règles d'association dans les bases de données massives, nous avons implémenté FP-growth (PFP-growth, voir la Figure A.2) dans Apache Spark version 2. L'algorithme PFP-growth prend un RDD de transactions, où chaque transaction est un

ensemble d'objets. Les itemsets fréquents extraits sont donnés dans le tableau 5.3.

Tableau 5.3 – Les itemsets fréquents

N°	Itemsets fréquents	Support
1	[Collapse road]	40
2	[Collapse road,Clear]	28
3	[Collapse road,Summer]	27
4	[Collapse road,M]	35
5	[Collapse road,M,Day]	28
6	[Collapse road,Day]	31
7	[[21-27],M]	27
8	[[21-27],Day]	32
9	[Clear]	54
10	[Clear,M]	37
11	[Clear,M,Day]	30
12	[Clear,Day]	45
13	[Horizontal,[21-27]]	27
14	[Horizontal,M]	27
15	[Summer]	52
...	...	...
30	[Car,Clear,Day]	30
31	[Car,M]	38
32	[Car,M,Day]	29
33	[Car,Day]	41
34	[Unpaved road]	37
35	[Unpaved road,Day]	27
36	[Fatal]	42
37	[Fatal,Clear]	33
38	[Fatal,Clear,M]	28
39	[Fatal,Clear,Day]	28
40	[Fatal,Summer]	29

Génération des règles d'association

Dans cette étape, nous avons utilisé l'algorithme PFP-growth pour extraire les règles d'association, voir le tableau 5.4. Une règle représente une association entre un ensemble des itemsets, cette association peut être quantifiée par un ensemble de mesures. Les deux mesures les plus classiques au niveau des règles d'association sont le support et la confiance. Dans cette étude, l'objectif est de mettre en évidence les différents attributs qui provoquent les accidents et transcrire la connaissance sous forme de règle d'association (si antécédent alors conséquent).

Tableau 5.4 – Les règles d'association extraites

N°	Antécédent	Conséquent	Confiance
1	Fatal	M	0.85
2	Fatal,Day	Clear	0.90
3	Clear,M	Day	0.81
4	Car,Clear	Day	0.21
5	Fatal,Summer	Clear	0.93
6	[12-18],Summer	Day	0.90
7	Summer,Clear	Day	0.89
8	Clear	Day	0.83
9	3	Car	1.00
10	2,Day	Clear	0.96
11	Collapse road,Day	M	0.90
12	Summer,Car	Day	0.96
13	Collapse road,M	Day	0.80
14	[6-12]	Day	0.93
15	[12-18],Day	Summer	0.83
16	<2, Clear	Day	0.96
17	[21-27]	Day	0.84
18	Summer,M	Day	0.88
19	Summer	Day	0.92
20	S	Day	0.84
21	[12-18]	Day	0.81
22	<2	Clear	0.90
23	<2	Car	0.93
24	<2	Day	0.90
25	Collapse road	M	0.87
26	Md	M	0.86
27	Fatal,Clear	M	0.84
28	Fatal,Clear	Day	0.84
29	Fatal,Clear	Summer	0.81

Cette tache d'extraction des règles d'association est très couteuse en terme du temps d'exécution. Dans ce cadre nous avons mené une étude comparative entre l'algorithme Apriori dans un contexte classique et PFP-growth dans un contexte du Big data (Spark) en utilisant trois tailles différentes de base de données (Tableau 5.5).

Tableau 5.5 – Temps d'exécution

Fichier	Apriori	PFP-growth
ACCIDENT.CSV (1MB)	8.80s	3.20s
ACCIDENT.CSV (20MB)	29mn	5.02s
ACCIDENT.CSV (100MB)	+3h(crashed)	9.12s

Évaluation des règles d'association extraites

Quel que soit l'algorithme utilisé, le nombre de règles qui vont être générées dépend fortement de la taille de données, du nombre d'attributs (colonnes) et du seuil de support minimum et confiance minimale. Une fois ces paramètres définis, on génère un ensemble de règles d'association. Ces règles doivent être analysées par le décideur afin qu'il puisse sélectionner les règles pertinentes.

Les algorithmes de la fouille de données fournissent une solution substantielle à l'extraction des règles d'association, la difficulté principale de cette phase d'analyse provient, d'une part, du grand nombre de règles d'association extraites et, d'autre part, de la qualité des règles extraites. Certaines approches proposées dans la littérature à ces difficultés restent limitées dans leur efficacité, pour faire face a ces problèmes, l'intégration d'analyse multicritère dans le processus d'extraction est pratiquement utile pour analyser la qualité des règles extraites. Dans cette approche, l'énorme quantité de règles extraites par l'algorithme PFP-growth nécessite l'utilisation d'une méthode de classement qui prend en compte un très grand nombre des alternatives (règles). Nous nous intéressons donc à une méthode existante appelée PROMITHEE en utilisant l'ensemble de règles précédemment extraites (Tableau 5.4) comme alternatives et les mesures de qualité (Tableau 5.1) comme critères.

La matrice de décision (Tableau d'évaluation) des règles extraites en fonction des mesures de qualité utilisée est donnée dans le tableau 5.6, ce tableau d'évaluation sert à décrire un problème d'analyse multicritères où chaque règle doit être évaluée selon un ensemble de critères, ces critères ont des poids d'importance, plus le poids est élevé, plus les critères sont importants, voir le Tableau 5.7.

Tableau 5.6 – la matrice de décision

Règle	Support	Lift	Laplace	Confiance	Conviction	Leverage	Jaccard	Phi-coeff
Rule1	36	85	97	2	87	30	50	19
Rule2	31	90	96	2	93	52	55	18
Rule3	37	81	97	2	83	66	65	78
Rule4	35	85	97	2	87	90	36	98
Rule5	29	93	96	3	96	90	45	69
Rule6	33	90	97	2	92	12	89	98
Rule7	39	89	97	2	91	93	45	89
Rule8	54	83	98	1	84	10	65	98
Rule9	29	10	96	3	2	59	43	85
Rule10	30	96	96	3	1	0	45	97
Rule11	31	90	96	2	93	20	16	58
Rule12	31	96	96	3	1	33	99	56
Rule13	35	80	97	2	82	99	89	45
Rule14	32	93	97	2	95	46	78	68
Rule15	36	83	97	2	85	85	98	86
Rule16	30	96	96	3	0	75	69	87
Rule17	38	84	97	2	86	56	68	84
Rule18	36	88	97	2	90	89	94	98
Rule19	52	92	98	1	93	58	59	56
Rule20	38	84	97	2	86	87	93	54
Rule21	44	81	97	1	83	69	98	36
Rule22	30	90	96	3	93	58	97	57
Rule23	30	93	96	3	96	69	94	58
Rule24	30	90	96	3	93	39	98	98
Rule25	40	87	94	2	92	93	93	97
Rule26	36	86	97	2	88	58	89	95
Rule27	33	84	97	2	86	79	97	97
Rule28	33	84	97	2	86	87	36	65
Rule29	33	81	97	2	83	89	91	95

Tableau 5.7 – Poids des critères

Règle	Support	Lift	Laplace	Confiance	Conviction	Leverage	Jaccard	Phi-coeff
Poids	1.00	1.00	0.05	0.07	0.90	0.50	0.20	0.60

L'étape suivante consiste à calculer la préférence entre les paires (Formule 2.7). Cette fonction exprime le degré de préférence entre $Rule_i$ et $Rule_j$ sur tous les critères. Ensuite, nous calculons le flux de surclassement partiel et global (Formule 2.8), voir le tableau 5.8, puis nous présentons le résultat final du classement des règles d'association en fonction de tous les critères utilisés par les décideurs, voir le Tableau 5.9.

Tableau 5.8 – Flux de préférences

Règle	Support	Lift	Laplace	Confiance	Conviction	Leverage	Jaccard	ϕ-coeff
Rule1	0.32	-0.25	0.32	-0.78	-0.25	-0.78	-0.57	-0.89
Rule2	-0.42	0.28	-0.60	0.39	0.35	-0.57	-0.50	-0.96
Rule3	0.50	-0.85	0.32	0.28	-0.85	-0.14	-0.32	-0.03
Rule4	0.10	-0.25	0.32	0.00	-0.25	0.60	-0.89	0.82
Rule5	-0.96	0.64	-0.60	0.85	0.67	0.60	-0.71	-0.10
Rule6	-0.10	0.28	0.32	0.21	0.10	-0.92	0.07	0.82
Rule7	0.71	0.07	0.32	-0.32	0.00	0.75	-0.71	0.32
Rule8	1.00	-0.67	0.96	-1.00	-0.71	-1.00	-0.32	0.82
Rule9	-0.96	1.00	-0.60	1.00	1.00	-0.21	1.00	0.10
Rule10	-0.71	0.85	-0.60	0.85	0.85	0.96	-0.71	0.57
Rule11	-0.42	0.28	-0.60	0.39	0.35	-0.85	-1.00	-0.35
Rule12	-0.42	0.85	-0.60	0.64	0.85	0.96	0.92	-0.57
Rule13	0.10	-1.00	0.32	-0.32	-1.00	0.85	0.07	-0.75
Rule14	-0.28	0.64	0.32	-0.17	0.57	-0.64	-0.07	-0.17
Rule15	0.32	-0.67	0.32	-0.17	-0.64	0.21	0.78	0.17
Rule16	-0.71	0.85	-0.60	0.85	0.85	0,07	-0.14	0.25
Rule17	0.60	-0.46	0.32	-0.53	-0.46	-0,50	-0.21	0.03
Rule18	0.32	0.00	0.32	0.07	-0.07	0.46	0.46	0.82
Rule19	0.92	0.50	0.96	-0.92	0.35	-0.35	-0.42	-0.57
Rule20	0.60	-0.46	0.32	-0.53	-0.46	0.32	0.32	-0.67
Rule21	0.85	-0.85	0.32	-0.85	-0.85	-0.03	0.78	-0.82
Rule22	-0.71	0.28	-0.60	0.53	0.35	-0.35	0.60	-0.46
Rule23	-0.71	0.64	-0.60	0.71	0.67	-0.03	0.46	-0.35
Rule24	-0.71	0.28	-0.60	0.53	0.35	-0.71	0.78	0.82
Rule25	0.78	-0.07	-1.00	-0.71	0.10	0.750	0.32	0.00
Rule26	0.32	-0.14	0.32	-0.07	-0.14	-0.35	0.07	0.42
Rule27	-0.10	-0.46	0.32	-0.53	-0.46	0.14	0.60	0.57
Rule28	-0.10	-0.46	0.32	-0.53	-0.46	0.32	-0.89	-0.25
Rule29	-0.10	-0.85	0.32	0.14	-0.85	0.46	0.21	0.42

Les résultats obtenus après l'intégration d'AMC en particulier la méthode PROMETHEE confirme que la règle 12 ($Summer, Car \longrightarrow Day$) a un flux de préférence très élevé, donc c'est une règle plus pertinente. Les règles intéressantes selon les préférences des décideurs sont présentées dans le tableau 5.9 de l'ordre 1 à l'ordre 16. Le succès de la mise en œuvre de la méthode PROMETHEE dans le processus décisionnel dépend de l'expérience et les préférences des décideurs.

Les règles extraites (Tableau 5.9) suggèrent qu'il existe une relation forte entre ces attributs (accidents mortels, conditions météorologiques, sexe du conducteur, état de la route et saison de l'année). Les résultats peuvent aider les décideurs à formuler de nouvelles politiques et stratégies pour améliorer et optimiser la sécurité routière.
Dans le chapitre 3, nous avons utilisé Apriori algorithme, le résultat confirme une association entre les comportements des conducteurs, les conditions météorologiques, les conditions de lumière et la gravité de l'accident. L'extension de cette contribution au contexte du Big Data confirme non seulement une association entre différentes variables liées aux accidents, mais aussi résoudre le problème de stockage en mémoire, l'amélioration de temps de réponse des algorithmes d'apprentissage et l'analyse de qualité des règles extraites. En résumé, l'intégration de AMC dans le processus d'extraction des règles d'association dans les bases de données massives favorise une signification très élevée et classe les règles différemment selon les préférences des décideurs.

Tableau 5.9 – Les règles associations pertinentes

Ordre	Règle	ϕ	$\phi+$	$\phi-$
1	Rule12	0.3571	0.6384	0.2813
2	Rule18	0.3170	0.6027	0.2857
3	Rule10	0.2768	0.5848	0.3080
4	Rule16	0.2054	0.5580	0.3527
5	Rule7	0.1518	0,5313	0,3795
6	Rule23	0.1250	0.5179	0.3929
7	Rule6	0.1161	0.4911	0.3750
8	Rule24	0.1116	0.4911	0.3795
9	Rule26	0.0714	0.4821	0.4107
10	Rule5	0.0670	0.4911	0.4241
11	Rule4	0.0670	0.4777	0.4107
12	Rule19	0.0670	0.5134	0.4464
13	Rule15	0.0536	0.4732	0.4196
14	Rule25	0.0402	0.4509	0.4107
15	Rule14	0.0402	0.4777	0.4375
16	Rule27	0.0268	0.4464	0.4196
17	Rule29	-0.0223	0.4330	0.4554
18	Rule22	-0.0268	0.4286	0.4554
19	Rule20	-0.0536	0.4107	0.4643
20	Rule9	-0.1071	0.4196	0.5268
21	Rule8	-0.1161	0.4241	0.5402
22	Rule17	-0.1339	0.3750	0.5089
23	Rule3	-0.1384	0.3839	0.5223
24	Rule21	-0.1741	0.3616	0.5357
25	Rule13	-0.2054	0.3527	0.5580
26	Rule2	-0.2455	0.3304	0.5759
27	Rule28	-0.2500	0.3080	0.5580
28	Rule11	-0.2679	0.3170	0.5848
29	Rule1	-0.3527	0.2768	0.6295

5.4 Conclusion

Dans ce chapitre, nous avons traité le problème d'extraction des règles d'association dans les bases de données massives (Big Data). Pour ce faire, nous avons utilisé Apache Spark et l'algorithme distribué PFP-growth pour extraire les itemsets fréquents et générer des règles d'association. De plus, l'intégration d'AMC dans le processus d'extraction des règles d'association a permet la résolution de problème de la pertinence et l'utilité des règles d'association. En fin, nous avons pu constater que Apache Spark, est un framwork puissant pour le traitement distribué et beaucoup plus rapide que Hadoop MapReduce, car il exploite les avantages des calculs en mémoire qui est particulièrement plus avantageux pour les calculs itératifs surtout dans le cas des algorithmes d'apprentissage automatique.

Chapitre 6

Implémentation et application dans le domaine de la sécurité routière

«All human things are subject to decay, and when fate summons, Monarchs must obey.»

Mac Flecknoe

Dans ce chapitre, nous allons décrire l'architecture de notre système, et nous présenterons les différentes phases de la préparation de données. Ensuite, nous implémentons les différents approches proposées dans cette thèse.

Sommaire

6.1 Introduction

L'objectif de ce chapitre est de proposer une implémentation logicielle permettant de tester concrètement l'apport des solutions proposées. Cette implémentation est constituée de trois interfaces. La première intitulée *interface ARM*. Cette interface est dédiée à l'extraction des règles d'association. La deuxième interface, intitulée *interface MCDA*, permet l'évaluation et l'extraction des règles d'association pertinentes. Quant à la dernière, intitulée *Time Series Forcasting*, elle est dédiée à la prédiction des accidents routiers, en particulier, le nombre des blessures et décès. Par ailleurs ces interfaces interactives et conviviales d'exploration de données ont été développées en utilisant le langage R et rshiny.

6.2 Système cible relatif à la sécurité routière

Selon les statistiques de l'Organisation Mondiale de la Santé (OMS) [11], chaque année, 1.3 million de personnes meurent dans les accidents de la route à travers le monde, et entre 20 et 50 millions de personnes subissent des traumatismes non mortels. Les accidents de la circulation sont une cause importante de décès chez les personnes âgées de 15 à 29 ans. Ces chiffres, spectaculaires, retiennent l'attention du public.

En plus, près de 3500 personnes meurent chaque jour sur les routes. Des dizaines de millions de personnes sont blessées et victimes d'incapacités. Les enfants, piétons, cyclistes et personnes âgées sont parmi les usagers de la route les plus vulnérables. L'OMS travaille avec ses partenaires gouvernementaux dans le monde entier pour sensibiliser aux mesures permettant de prévenir les accidents de la route et promouvoir les bonnes pratiques telles que l'utilisation du casque ou de la ceinture de sécurité, la conduite à allure modérée, l'abstinence de consommation d'alcool et une bonne visibilité dans la circulation. Au Maroc, le Comité National de Prévention des Accidents de la Circulation (CNPAC) est un établissement d'utilité publique institué par le décret N° 2-72-275 du 27 Rajab 1397 (15 juillet 1977)[70] soumis au contrôle technique du Ministère de l'Équipement du Transport et de la Logistique [12], elle réunit des acteurs des secteurs publics et privés engageant de nombreux intervenants à débattre de la problématique des accidents de la circulation, prendre conjointement des décisions pour harmoniser les actions préventives et curatives et optimiser l'utilisation des moyens de lutte contre les accidents de la circulation. le bilan des accidents routiers dans les années 2015-2017 [70] est donné comme suit :

Les moyennes mensuelles des accidents en 2015 ont été respectivement de près de 6.500 accidents et 315 de décès par mois, voir la Figure 6.1.

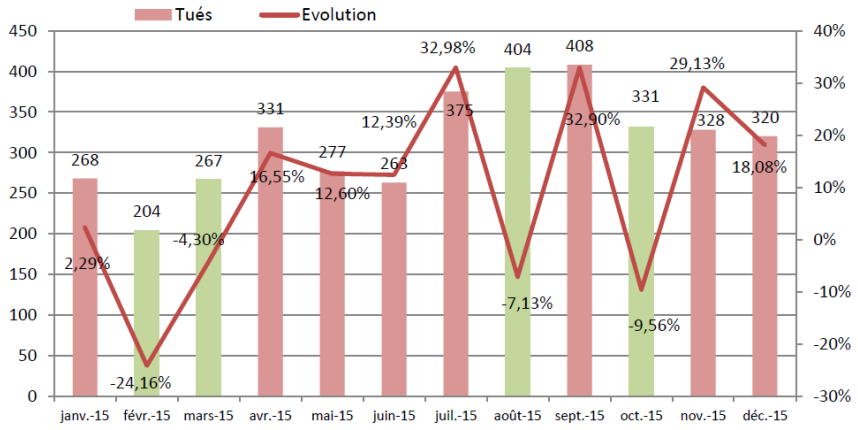

Figure 6.1 – Répartition des décès par mois en 2015

Les moyennes mensuelles des accidents et des décès en 2016 ont été respectivement de près de 6.723 accidents et 315 de décès par mois, voir la Figure 6.2.

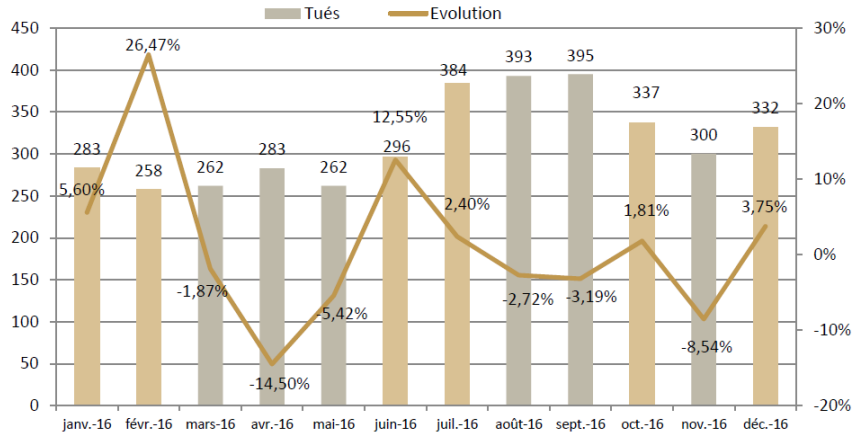

Figure 6.2 – Répartition des décès par mois en 2016

Dans le mois de mars 2017, une augmentation de l'ordre de 4.15% est enregistrée, la série des nombres des décès sur les 5 premiers mois de l'année 2017 a enregistré des diminutions continues, variant entre -3.25% en mois de janvier et -18.03% en mois de février 2017, voir la Figure 6.3.

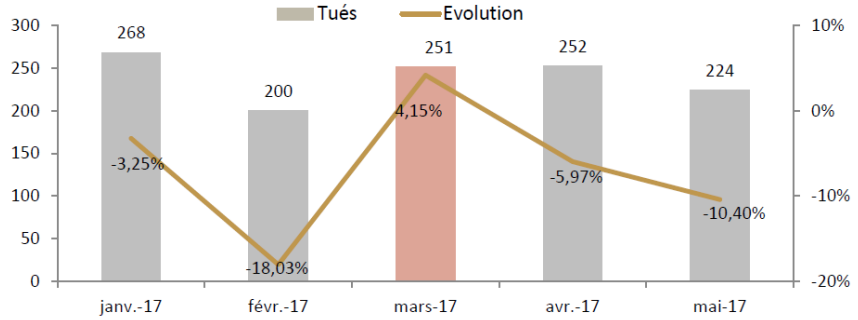

Figure 6.3 – Évolution mensuelle des décès au titre des 5 premiers mois de l'année 2017

6.3 Architecture du système

Pour répondre aux objectifs fixés dans cette thèse, nous avons conçu et développé un système d'aide à la décision basé sur les techniques descriptives et prédictives du Data Mining et l'analyse multicritère. L'architecture de ce système est constituée de quatre modules principaux (Figure 6.4) à savoir :

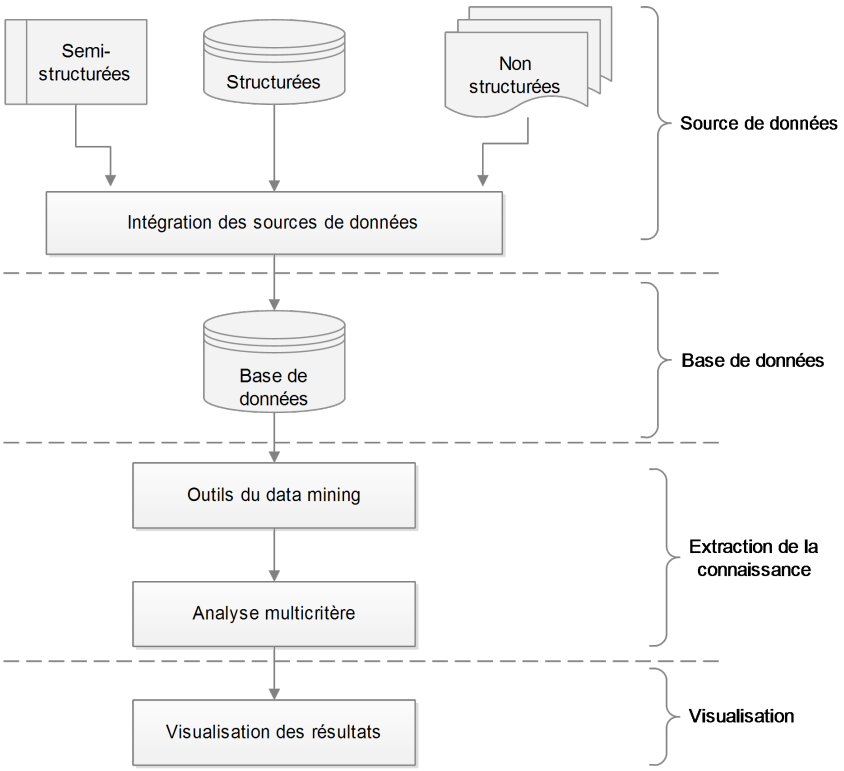

Figure 6.4 – Architecture globale du système

6.3.1 Module de sources de données

Pour chaque accident corporel, des saisies d'information décrivant l'accident sont effectuées par l'unité des forces de l'ordre (police, gendarmerie, etc.) qui sont intervenues sur le lieu de l'accident. Ces informations sont rassemblées dans une fiche intitulée bulletin d'analyse des accidents corporels. L'ensemble de ces fiches constitue le fichier national des accidents corporels de la circulation. Les bases de données, extraites du bulletin d'analyse des accidents corporels, répertorient l'intégralité des accidents corporels de la circulation intervenus durant une année précise au Maroc avec une description simplifiée. Cela comprend des informations de localisation de l'accident, telles que renseignées ainsi que des informations concernant les caractéristiques de l'accident et son lieu, les véhicules impliqués et leurs victimes, etc.

6.3.2 Module d'extraction de la connaissance

L'extraction de la connaissance connue aussi sous l'expression du Data Mining, forage de données, fouille de données, a pour objet d'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Il propose d'utiliser un ensemble d'algorithmes pour construire des modèles à partir de données, c'est-à-dire trouver des motifs intéressants selon des critères fixés au préalable, et d'en extraire un maximum de connaissances utiles et pertinentes.
Dans cette thèse nous avons utilisé les technique du Data Mining tel que les règles d'association, arbre de décision, la régression, et série temporelle (Time series). Ces techniques à pour but de traiter un ensemble de données complexe, elles sont utilisées dans divers domaines appartenant à la science et à la gestion. De plus, elles sont utilisées dans la détection de fraude et beaucoup plus dans le problème de la gravité et l'analyse des accidents de la route.

6.3.3 Module d'analyse multicritère

Les méthodes mathématiques d'analyse multicritère ont pour but de résoudre des problèmes d'analyse multicritère. Elles constituent une étape importante du processus de décision, qui suit celle d'identification et de définition du problème, et aboutissent au choix d'une ou plusieurs solutions parmi un ensemble discret de solutions, via une procédure de sélection. Elles permettent également de répondre aux problématiques de tri et de rangement, par l'intermédiaire d'une procédure d'affectation et de classement respectivement [83].

6.3.4 Module de visualisation

La visualisation des données est un ensemble de méthodes de représentation graphique, en deux ou trois dimensions. Les moyens informatiques ont permis de représenter des ensembles complexes de données, de manière plus simple, didactique et pédagogique. Dans le but d'illustrer les résultats obtenus nous avons utilisé un ensemble des techniques de visualisation à savoir, les arbres de décision, nuage de points (Scatter plot), coordonnées parallèles (Parallel Coordinates), les graphes (Graph-based visualization), etc.

6.4 Préparation des données

Les données proviennent de plusieurs sources peuvent contenir des anomalies ou des valeurs incorrectes qui compromettent la qualité du jeu de données. Les problèmes de qualité les plus fréquents sont les suivants :

— *Caractère incomplet :* des valeurs ou des attributs sont manquants ;
— *Bruit :* les données contiennent des enregistrements erronés ;

— *Incohérence :* les données contiennent des enregistrements en conflit, nommage, codage.

Le nettoyage de données est une étape très importante, mais malheureusement souvent négligée dans le processus du Data Mining. Le nettoyage de données s'applique particulièrement aux projets types du Data Mining où de grosses volumétries de données sont collectées de façon automatique. Pour obtenir des modèles prédictifs performants, il faut impérativement analyser les données, détecter les anomalies et déterminer les étapes de prétraitement et de nettoyage appropriées.

Le prétraitement est l'une des tâches importantes dans l'exploration de données (Figure 6.5). Il présente principalement l'élimination du bruit, la gestion des valeurs manquantes, la suppression des attributs non pertinents afin de rendre les données prêtes pour l'analyse. Les principales opérations à effectuer lors du prétraitement de données sont :

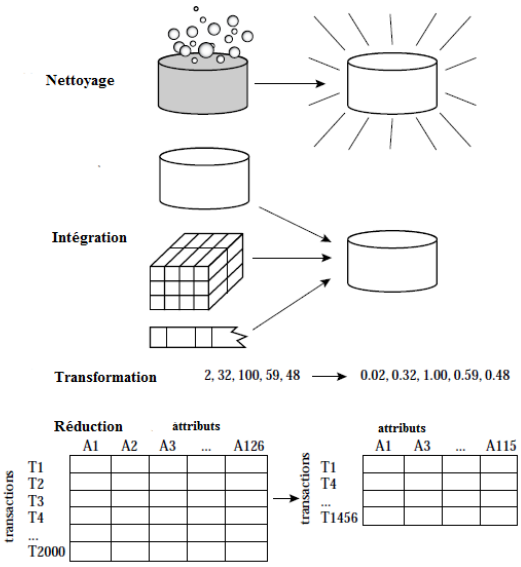

Figure 6.5 – Principales étapes dans le prétraitement de données

— *Nettoyage :* compléter les valeurs manquantes, détecter et supprimer les données bruyantes ;

— *Transformation :* normaliser les données pour réduire le volume et le bruit ;

— *Réduction :* échantillonner les enregistrements de données ou les attributs pour faciliter la manipulation ;

— *Discrétisation de données :* convertir des attributs continus en attributs catégoriels pour simplifier l'exploitation de données dans certains outils d'apprentissage automatique ;

— *Nettoyage du texte :* supprimer les caractères spéciaux qui peuvent perturber l'alignement de données, les nouvelles lignes qui peuvent interrompre des enregistrements.

6.4.1 Gérer les valeurs manquantes

Dans le cas où nous rencontrons des valeurs manquantes ou non disponibles, la première chose à faire est d'en identifier l'origine. Ces anomalies sont dues aux mauvais fonctionnements de l'équipement, considérés peu importantes au moment de la saisie, non saisies, car non ou mal comprises, etc. Les méthodes les plus courantes de traitement des valeurs manquantes sont les suivantes :

— *Suppression :* supprimer les enregistrements ayant des valeurs manquantes ;
— *Remplacement par une constante globale :* remplacer des valeurs manquantes par une valeur factice : par exemple, NA pour les valeurs catégorielles ou 0 pour les valeurs numériques ;
— *Remplacement par la moyenne :* si les données manquantes sont numériques, remplacez-les par la valeur moyenne, et si les données manquantes sont catégorielles, remplacer les par l'élément le plus fréquent ;
— *Utiliser la valeur la plus probable :* utiliser des formules Bayésiennes ou arbre de décision.

6.4.2 Gérer les données bruitées

Le bruit concerne l'erreur ou variance aléatoire d'une variable mesurée, produite par instrument de mesure défectueux, problème de saisie, limitation technologique, enregistrements dupliqués, et l'incohérence dans les conventions de nommage. Les méthodes les plus courantes de traitement des données bruitées sont les suivantes :

— *Partitionnement :* trier et partitionner les données ;
— *Clustering :* détecter et supprimer les exceptions ;
— *Régression :* lisser les données par des fonctions de régression.

6.4.3 Réduction des données

Les techniques de fouille de données peuvent être très lentes sur les données complètes. Pour faciliter la manipulation des données, plusieurs méthodes permettent de réduire la taille des données, les méthodes applicables sont les suivantes :

— *Échantillonnage des enregistrements :* obtenir une représentation réduite du jeu de données, plus petite en volume, mais qui produit les mêmes (ou presque) résultats analytiques ;
— *Échantillonnage des attributs :* sélectionner les attributs importants dans les données ;
— *Agrégation :* diviser les données en groupes et stocker les nombres de chaque groupe. Par exemple, le chiffre d'affaires quotidien d'une chaîne de transport sur les 20 dernières années peut être agrégé en un chiffre d'affaires mensuel pour réduire la taille des données.

6.5 Implémentation

L'implémentation de notre système, que nous avons proposé précédemment, utilise la plateforme Rstudio, le langage de programmation utilisé est le langage R version 3.4.1 ainsi que le Framework rshiny pour le développement des interfaces web interactives. La Figure 6.6 ci-dessous illustre l'architecture du système. Ce système possède une architecture client-serveur constituée de plusieurs composantes à savoir les packages R pour l'extraction des règles d'associations, analyse mulricritère et les techniques de visualisation.

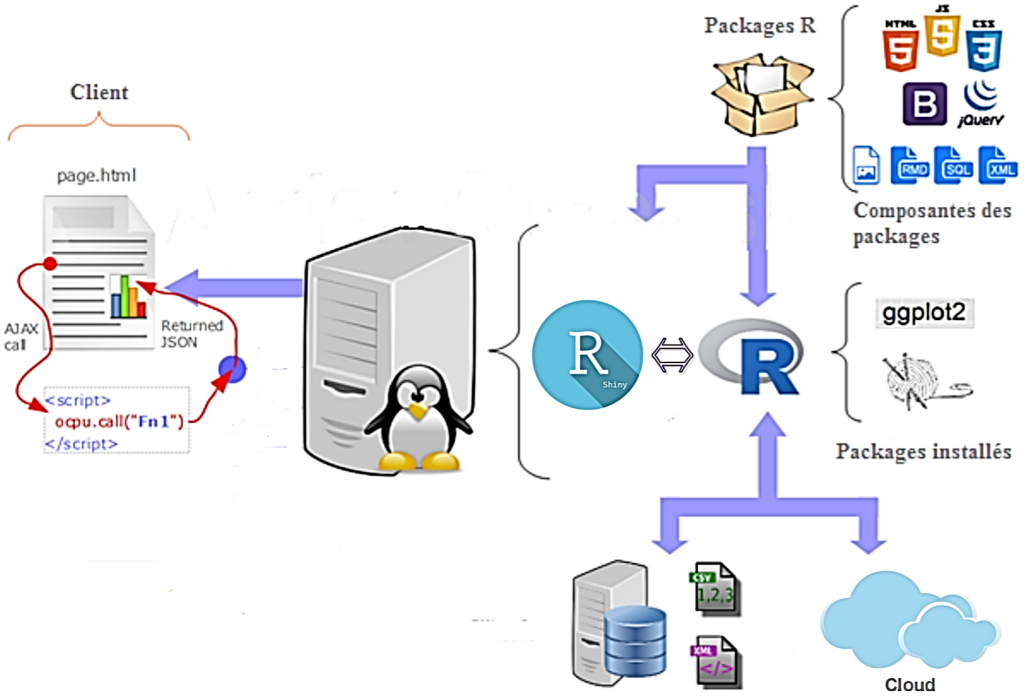

Figure 6.6 – L'architecture technique du système

Après le nettoyage et la transformation de différentes sources de données en un fichier CSV qui rassemblent les données relatives aux accidents routiers. Les packages R des algorithmes du Data Mining (Arule,ArulViz,Rpart,etc.) traite et analyse les données en entrée pour extraire les règles d'association, ensuite, ces dernières sont utilisées comme éléments d'entrée aux algorithmes d'analyse multicritère pour l'évaluation et le choix des règles d'associations les plus pertinentes. Le client émet une requête vers le serveur grâce à son adresse IP et le port, qui désigne un service particulier du serveur, le serveur reçoit la demande et répond à l'aide de l'adresse de la machine cliente et son port.

Le système développé est composé de trois parties dont la première est l'extraction des règles d'association (Figure 6.7), cette partie permet de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans la base de données. La deuxième partie présente l'utilisation de l'analyse multicritère pour l'analyse de qualité des règles extraites (Figure 6.8). Quant à la dernière est réservé à l'analyse de

série temporelle pour comprendre l'évolution d'une suite de valeurs numériques et pour en prévoir le comportement futur.

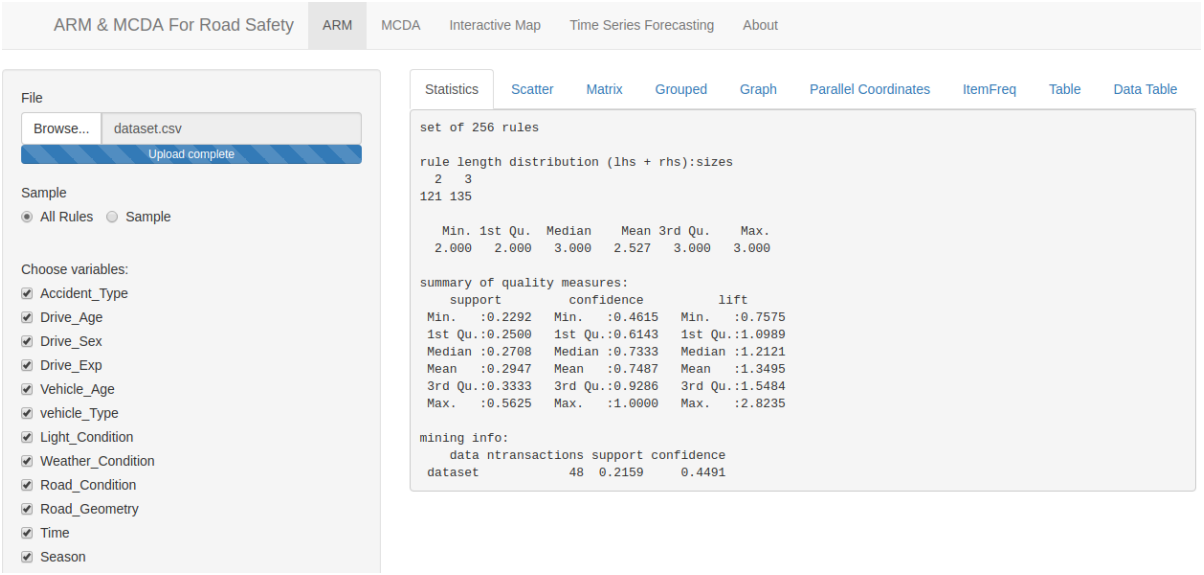

Figure 6.7 – L'extraction des règles d'association

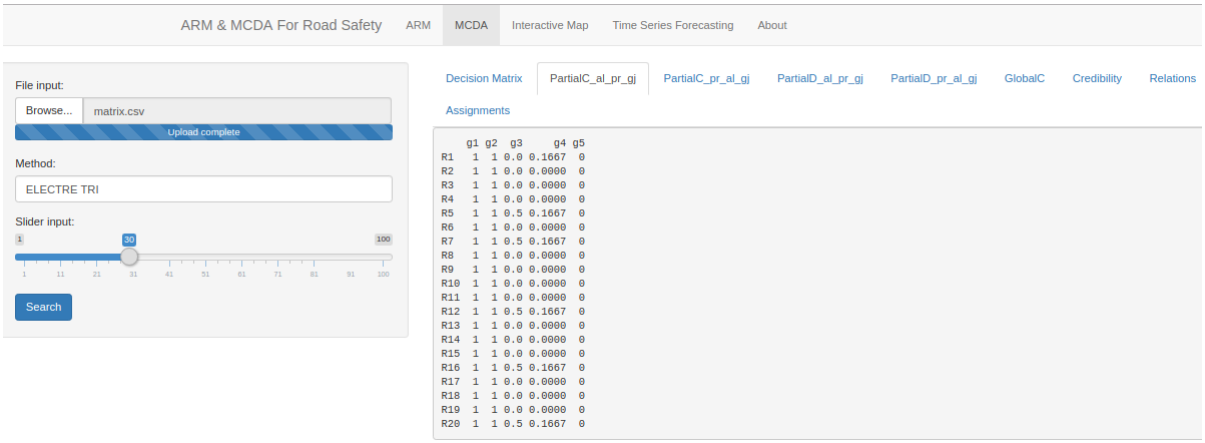

Figure 6.8 – Évaluation des règles d'association à l'aide d'analyse multicritère

Dans la première partie, pour l'extraction des règles d'association, nous avons implémenté l'algorithme Apriori (Figure 1.10) en deux étapes, la première est l'extraction des itemsets fréquents (Figure 6.9), la deuxième est la génération des règles d'association (Annexe A.3). Afin d'illustrer les résultats nous avons utilisé les techniques de visualisation nuage de points (Figure 6.10), Coordonnées parallèles (Figure 6.11), matrice groupée des règles d'association (Annexe A.4) et graphe de règles d'association (Annexe A.5).

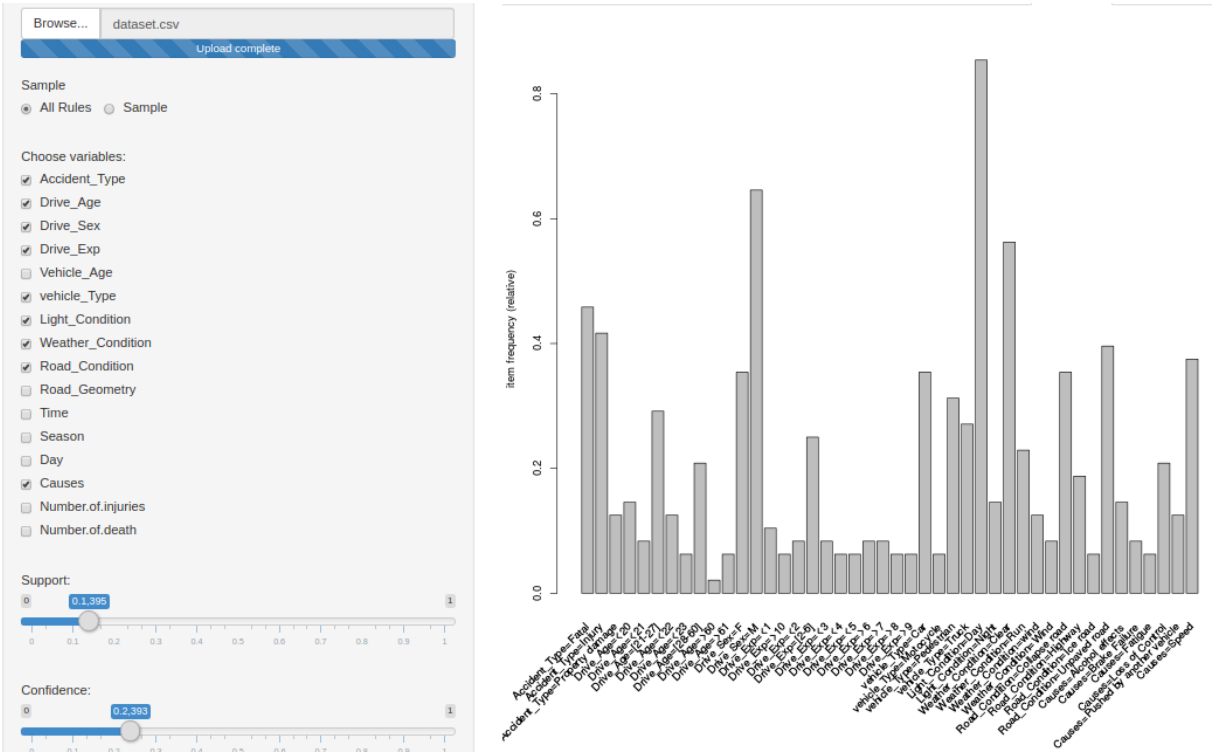

Figure 6.9 – Extraction des itemsets fréquents

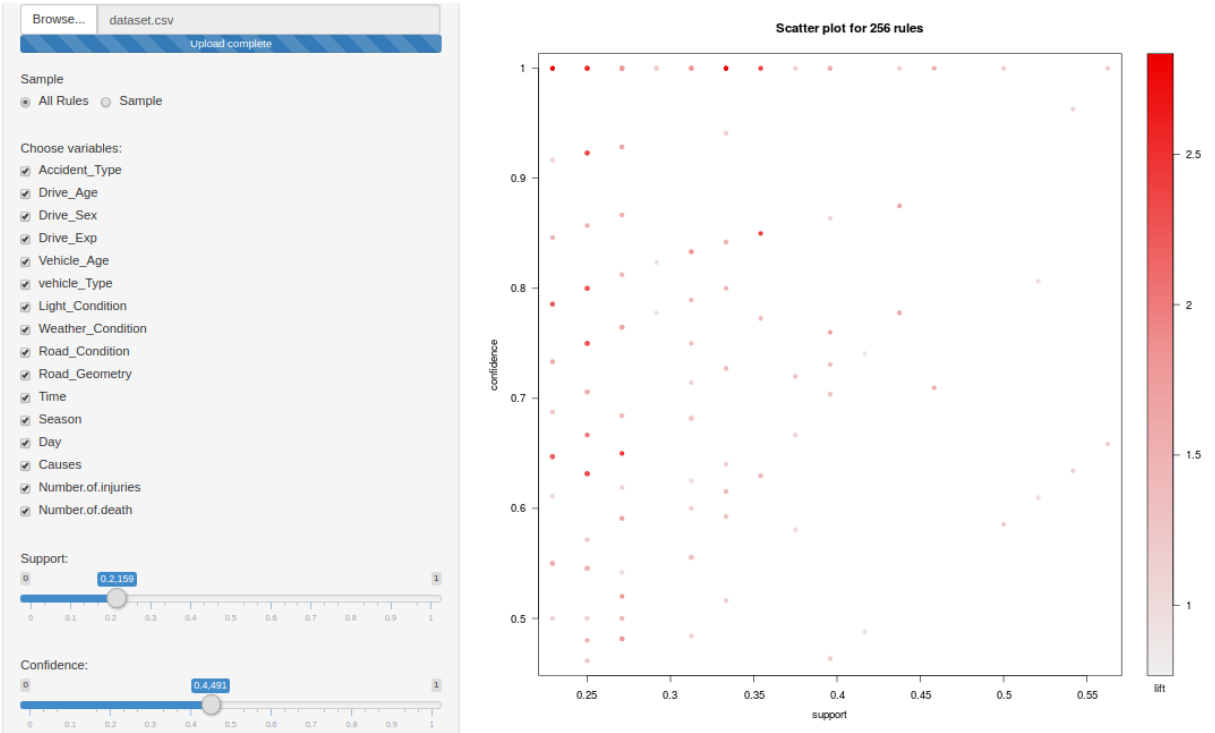

Figure 6.10 – Visualisation à l'aide de nuage de points

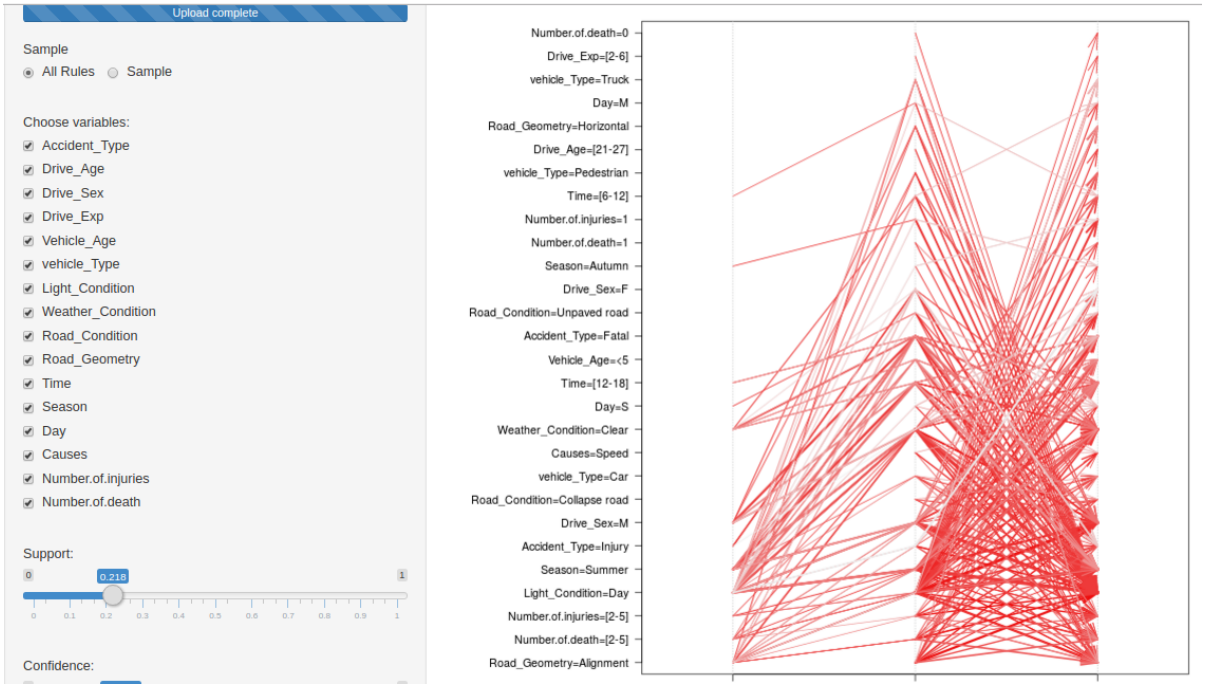

Figure 6.11 – Visualisation à l'aide de coordonnées parallèles

Pour l'évaluation des règles d'association extraites, nous avons implémenté la méthode ELECTRE TRI en calculons l'indice de concordance partielle (Annexe A.7), et l'indice de concordance globale (Annexe A.8). Le résultat est un tri d'affectation des règles d'association de plus pertinentes au moins pertinentes (Annexe A.9).

La projection des résultats sur la carte vectorielle est donnée dans la Figure 6.12. Cette carte représente les détails des accidents en termes des blessures, décès et types de véhicules, etc. (Annexe A.6).

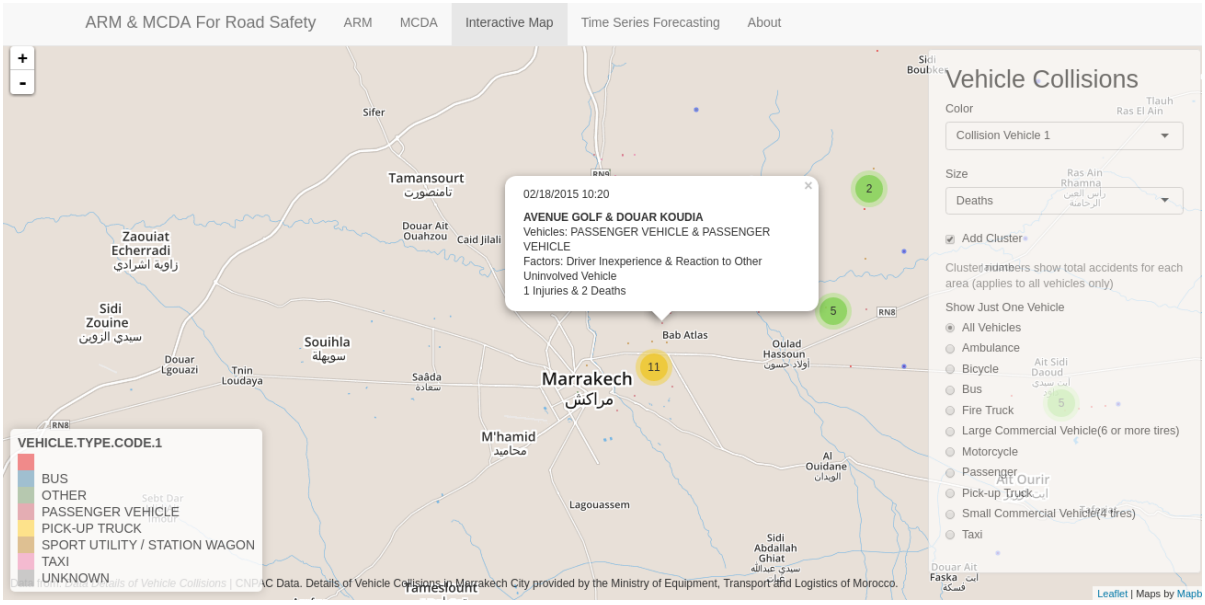

Figure 6.12 – Visualisation des accidents dans la ville de Marrakech

Une autre partie du système consiste à prédire le nombre des blessures et décès à partir des données relatives aux accidents routiers, pour cela nous avons utilisé les séries temporelles [84, 85].

Une série temporelle est une suite finie $(x_1, ..., x_n)$ de données indexées par le temps. L'indice temps peut être selon les cas, la minute, l'heure, le jour, l'année etc.

On peut voir une série temporelle comme une suite d'observations répétées d'un même phénomène à des dates différentes (par exemple le nombre d'accidents moyenne journalière en un lieu donné). On représente habituellement une série temporelle à l'aide d'un graphique en abscisse les dates et en ordonnée les valeurs observées. Les Figures 6.13 et 6.14 représentent deux séries temporelles qui correspondent aux prévisions des blessures et décès respectivement.

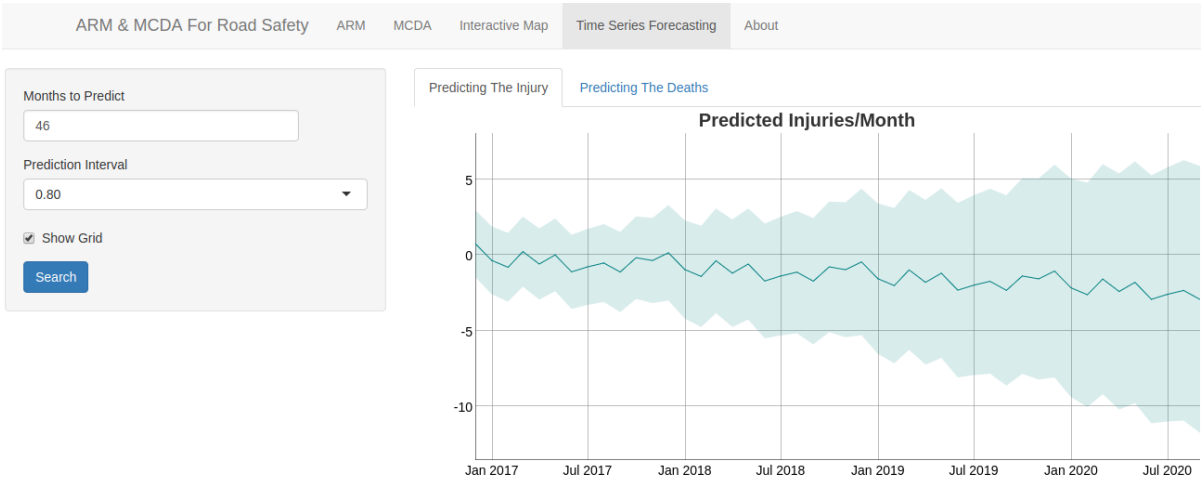

Figure 6.13 – Prédiction des blessures à l'aide des séries temporelles

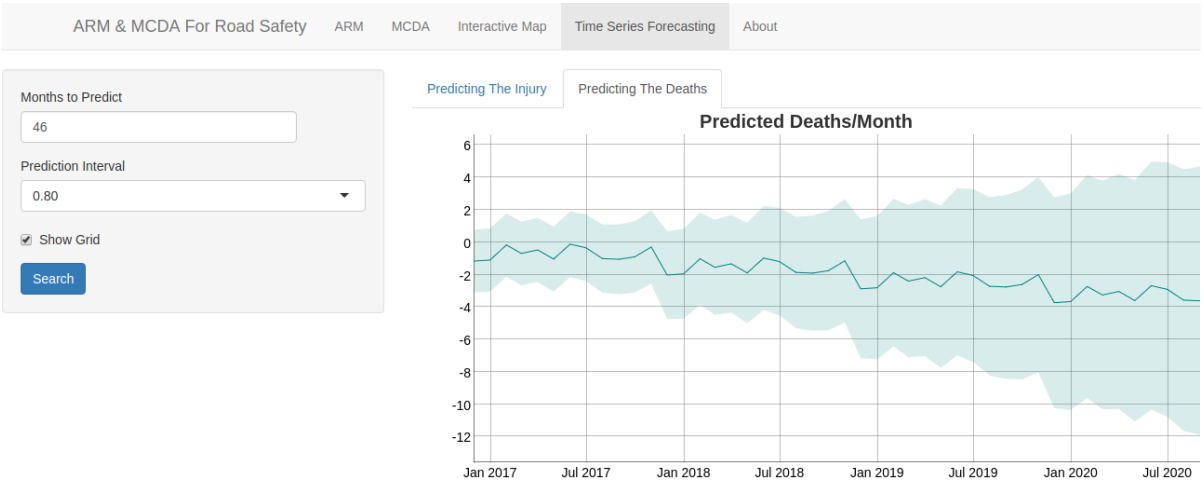

Figure 6.14 – Prédiction des décès à l'aide des séries temporelles

6.6 Conclusion

Dans ce chapitre, nous avons présenté une implémentation logicielle pour illustrer l'apport de nos différentes approches proposées en combinant la fouille de données, en particulier les règles d'association et l'analyse mulricritère. Les résultats qui en découlent permettent aux décideurs d'avoir une vision globale concernant le problème traité. Ce système commence par l'étape d'intégration et de transformation de données relatives aux accidents routiers, ensuite, l'utilisation de l'algorithme Apriori pour l'extraction des règles d'association entre les différentes variables de base de données. Ces dernières constituent les éléments d'entrée pour la méthode ELECTRE TRI afin d'évaluer et extraire les règles les plus pertinentes, en prenant en compte les préférences des décideurs et les poids d'importance des mesures de qualité.

Conclusion et Perspectives

«Not everything that can be counted counts and not everything that counts can be counted.»

Albert Einstein

6.7 Conclusion

Dans ce travail, nous avons proposé trois approches et une étude expérimentale. D'une part, nous avons proposé une démarche d'aide à la décision permettant de démontrer l'apport de l'intégration de l'analyse multicritère en vue de remédier à certains problèmes d'aide à la décision tels que ceux liés à la qualité des règles d'association issues d'un processus d'extraction. Nous avons démontré également que cette démarche méthodologique permet d'extraire les règles d'association pertinentes, cette première solution nous a permis de maitriser l'aspect qualité des règles extraites. Mais elle présente des limites au niveau des types de données, en fait les données peuvent être de type spatial. Dans ce contexte nous avons présenté les limites de certaines approches d'extraction des règles d'association spatiales existante ; à savoir les approches proposées par Koperski [8] , Salleb [7] et Marghoubi [9]. Ensuite, nous avons démontré l'apport de l'intégration de la logique floue pour l'extraction des règles d'association spatiales pertinentes.
Cette deuxième solution a permis l'extraction des règles d'association à référence spatiale en se basant sur la logique floue plus précisément, au niveau de l'estimation des distances entre les objets des couches thématiques considérées. Cette étude nous a amené également à proposer une approche décisionnelle globale combinant l'analyse multicritère, les outils du Big Data et les règles d'association en vue de proposer une solution analytique adaptée aux besoins des décideurs dans un contexte de données massives. Dans cette approche nous avons surmonté le problème de temps de réponse et d'espace mémoire, et nous avons démontré que les calculs distribués dans le contexte du Big Data constitue une solution efficace d'extraction des règles d'association pertinentes.

Notre contribution s'est déclinée en six chapitres : dans l'introduction générale, nous avons présenté le cadre général de la découverte de connaissances, ainsi que le contexte décisionnel qui a motivé les travaux de cette thèse. Dans le premier chapitre, nous avons

introduit l'extraction de la connaissance dans les bases de données, en particulier, l'extraction des itemsets fréquents et les règles d'association spatiales, et nous avons montré les limites de certains algorithmes et approches d'extraction en termes de qualité, de temps de réponse et d'espace mémoire. Ensuite, nous avons présenté dans le deuxième chapitre, les méthodes d'analyse multicritères les plus utilisées et la logique floue. Nous nous sommes intéressés dans le troisième chapitre à la proposition d'une démarche méthodologique basée sur l'analyse multicritère, en particulier la méthode ELECTRE TRI pour l'extraction et la réduction des règles d'association. Cette méthodologie permet d'évaluer les règles d'association issues d'un processus d'extraction, nous avons commencé ce chapitre par une introduction décrivant la problématique abordée qui porte sur l'extraction des règles d'association pertinentes pour l'analyse des accidents routiers. Ensuite nous avons présenté l'approche proposée tout au long de ce chapitre. Nous avons montré aussi les limites liées à cette solution, ces limites concernant l'extraction des règles d'association à référence spatiale, ainsi que le volume important de données qui évolue d'une façon très rapide.

Afin de surmonter la limite relative à l'intégration de la composante spatiale, nous avons proposé dans le quatrième chapitre, une approche décisionnelle basée sur la logique floue pour l'extraction des règles d'association spatiales, l'apport de la logique floue est de permettre la prise en compte de l'imprécision et l'incertitude des valeurs des distances entre les différents objets de couches thématiques utilisées. Ensuite, pour maitriser le volume massif de données, nous avons proposé dans le cinquième chapitre, une approche décisionnelle globale basée sur la méthode PROMETHEE et l'algorithme d'extraction FP-growth dans un contexte du Big Data, en particulier, Apache Spark pour l'extraction des règles d'association pertinentes. Dans cette approche nous avons présenté les limites de nos approches ERA-AMC et ERAS-LF en termes de temps de calcul et d'espace mémoire, puis en présentant la méthodologie suivie tout au long de ce chapitre.

Afin de mettre en œuvre la solution décisionnelle proposée, nous avons présenté dans le sixième chapitre un prototype logiciel permet de concrétiser l'apport des approches proposées. Enfin, pour l'illustration de ces approches, nous avons considéré comme étude de cas le problème de la sécurité routière.

6.8 Perspectives

La découverte de connaissances dans les bases de données, en particulier, le problème d'extraction des règles d'association reste un domaine de recherche très abordé dans le contexte du Data Mining, Data Mining spatial, Big Data et Text Mining. Dans ce cadre, les travaux de recherche que nous avons présenté ouvrent la voie aux perspectives suivantes :

1. Intégration des systèmes multi-agents pour le choix des préférences des décideurs afin d'automatiser le processus d'analyse de qualité des règles d'association à l'aide de l'analyse multicritère.

2. Couplage entre l'analyse multicritère et les outils du Big Data et plus particulièrement le framwork Apache Spark.
3. Le SOLAP comme outil de fouille de données visuelle. Dans le domaine d'accidentologie, le SOLAP peut offrir une issue rapide et facile d'analyse des comportements des chauffeurs. Les utilisateurs auront la possibilité de visualiser plusieurs indicateurs par croisement de plusieurs axes d'analyse (par exemple : type du véhicule, localisation relative). Le raisonnement des utilisateurs sur les différents croisements de données peut permettre la découverte visuelle de relations spatiales, non-spatiales et de motifs de mouvements.
4. La généralisation de notre méthodologie aux objets mobiles peut être une perspective intéressante. En effet, il y a aujourd'hui une énorme quantité de données décrivant des déplacements de mobiles (véhicule, personnes, etc.) qui sont stockées dans les bases de données volumineuses. L'une des problématiques de ces données est le développement de méthodes efficaces pour découvrir automatiquement des connaissances utiles.
5. Exploration de données spatiales réparties. En raison du volume élevé des données spatiales, des nouvelles approches peuvent être nécessaires pour le traitement et l'analyse spatiale distribuée.
6. Optimisation des performances de notre système.
7. Extension du prototype aux autres méthodes du Data Mining spatiales.
8. Extension du prototype au contexte du Big Data pour l'analyse des données massives relatives aux accidents routiers en implémentant une approche basée sur l'analyse de données en temps réel.

Annexe A

Annexe

#	Measure	Definition
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{A}\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A}\overline{B})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{A}\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A}\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i),-\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\big)$
9	Gini index (G)	$\max\big(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2] - P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2] - P(A)^2 - P(\overline{A})^2\big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\big)$
13	Conviction (V)	$\max\big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B) - P(A)P(B)$
17	Certainty factor (F)	$\max\big(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\big)$
18	Added Value (AV)	$\max(P(B\|A) - P(B), P(A\|B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{A}\overline{B})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{A}\overline{B})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$

Figure A.1 – Liste des mesures de qualités

```scala
package org.apache.spark.examples.mllib
import scopt.OptionParser
import java.io._
import scala.io.Source
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.mllib.fpm.FPGrowth
import org.apache.spark.mllib.fpm.AssociationRules
import org.apache.spark.mllib.fpm.FPGrowth.FreqItemset

/**
 * Example for mining frequent itemsets using FP-growth algorithm.
 * Mining frequent itemsets from road accident dataset.
*/
object FpgRules {

  case class Params(
    input: String = "/opt/big_data/eclipse/workspace/MachineLearningLibrary/input/dataset.csv",
    minSupport: Double = 0.3,
    numPartition: Int = 1) extends AbstractParams[Params]

  def main(args: Array[String]) {
    val defaultParams = Params()

    val parser = new OptionParser[Params]("FPGrowthExample") {
      head("FPGrowth: an example FP-growth app.")
      opt[Double]("minSupport").text(s"minimal support level, default: ${defaultParams.minSupport}").action((x, c) => c.copy(minSupport = x))
      opt[Int]("numPartition").text(s"number of partition, default: ${defaultParams.numPartition}").action((x, c) => c.copy(numPartition = x))
      opt[String]("input").text(s"input paths to input data set, default: ${defaultParams.input}").action((x, c) => c.copy(input = x))
    }
    parser.parse(args, defaultParams).map { params => run(params)}.getOrElse {sys.exit(1)}
  }

  def run(params: Params) {
    val conf = new SparkConf().setAppName(s"FPGrowthExample with $params").setMaster("local[*]");
    val sc = new SparkContext(conf)
    val transactions = sc.textFile(params.input).map(_.split(",")).cache()

    println(s"Number of transactions: ${transactions.count()}")

    /* generate frequent itemsets
     * store frequent itemsets in text file: FreqItemsets.txt
     */
    val file_freq = "/opt/big_data/eclipse/workspace/MachineLearningLibrary/output/FreqItemsets.txt"
    val writer_freq = new BufferedWriter(new OutputStreamWriter(new FileOutputStream(file_freq)))
    val model = new FPGrowth()
      .setMinSupport(params.minSupport)
      .setNumPartitions(params.numPartition)
      .run(transactions)
    println(s"Number of frequent itemsets: ${model.freqItemsets.count()}")
    model.freqItemsets.collect().foreach { itemset => println(itemset.items.mkString("[", ",", "]") + ", " + itemset.freq)
    writer_freq.write(itemset.items.mkString("[", ",", "]") + ", " + itemset.freq + "\n")

    }
    writer_freq.close()

    /* generate association rules
     * store extracted association rule in text file: AssociationRules.txt
     */

    val ar = new AssociationRules().setMinConfidence(0.8)
    val file_arule = "/opt/big_data/eclipse/workspace/MachineLearningLibrary/output/AssociationRules.txt"
    val writer_ar = new BufferedWriter(new OutputStreamWriter(new FileOutputStream(file_arule)))
    val results = ar.run(model.freqItemsets)

    results.collect().foreach { rule =>
      println("[" + rule.antecedent.mkString(",") + "=>" + rule.consequent.mkString(",") + "]," + rule.confidence)
      writer_ar.write("[" + rule.antecedent.mkString(",") + "=>" + rule.consequent.mkString(",") + "]," + rule.confidence + "\n")
    }

    writer_ar.close()
    sc.stop()
  }
}
```

Figure A.2 – Implémentation de l'algorithme FP-growth dans Apache Spark

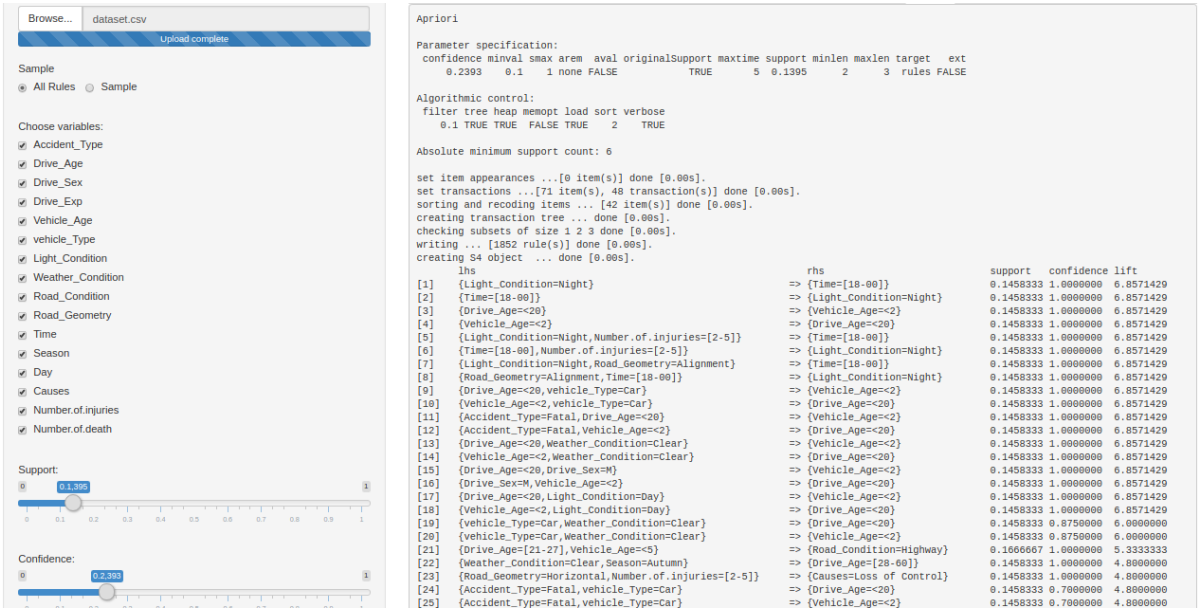

Figure A.3 – Visualisation des règles d'association extraites

Figure A.4 – Visualisation à l'aide de matrice groupée

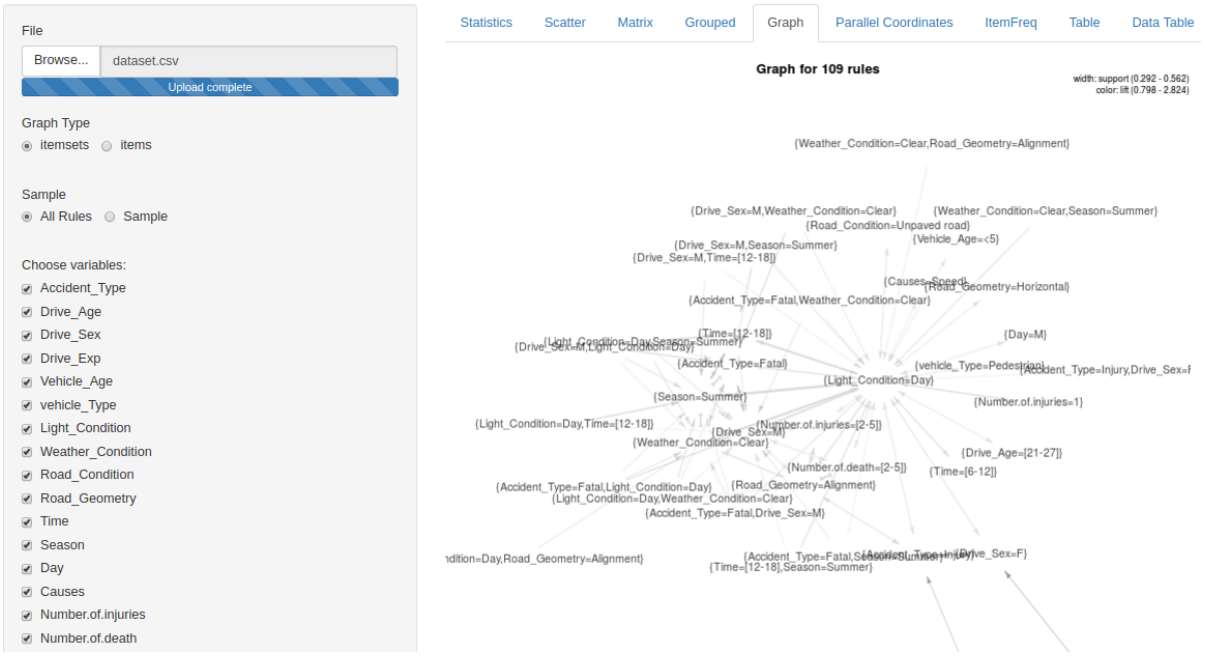

Figure A.5 – Visualisation graphique des règles d'association extraites

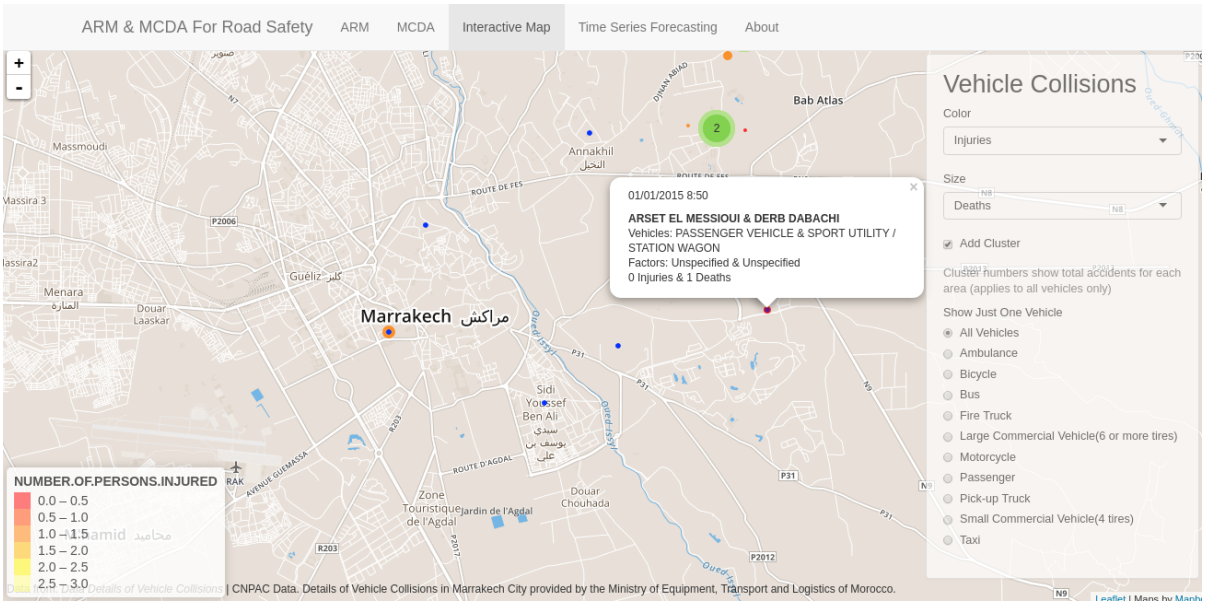

Figure A.6 – Gravité des accidents et zones dangereuses dans la ville de Marrakech

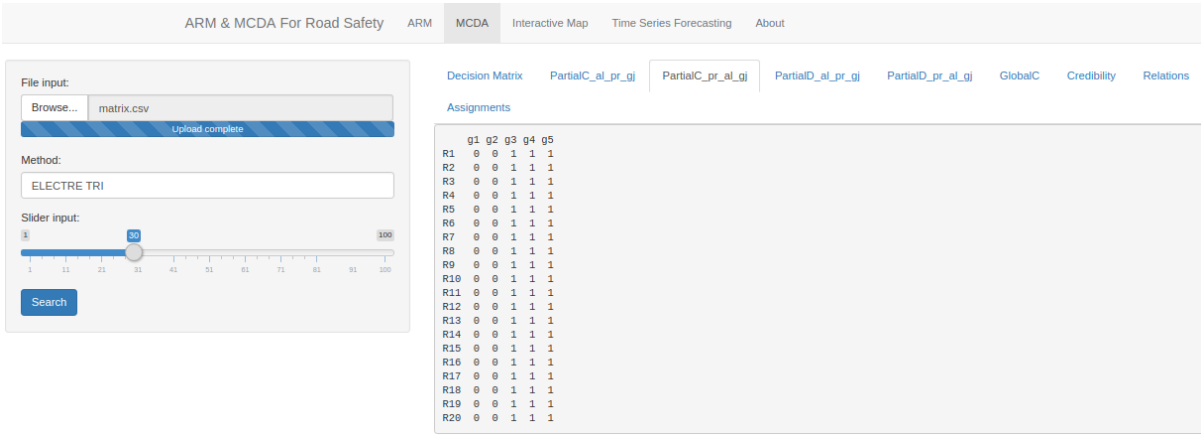

Figure A.7 – Indice de concordance partielle

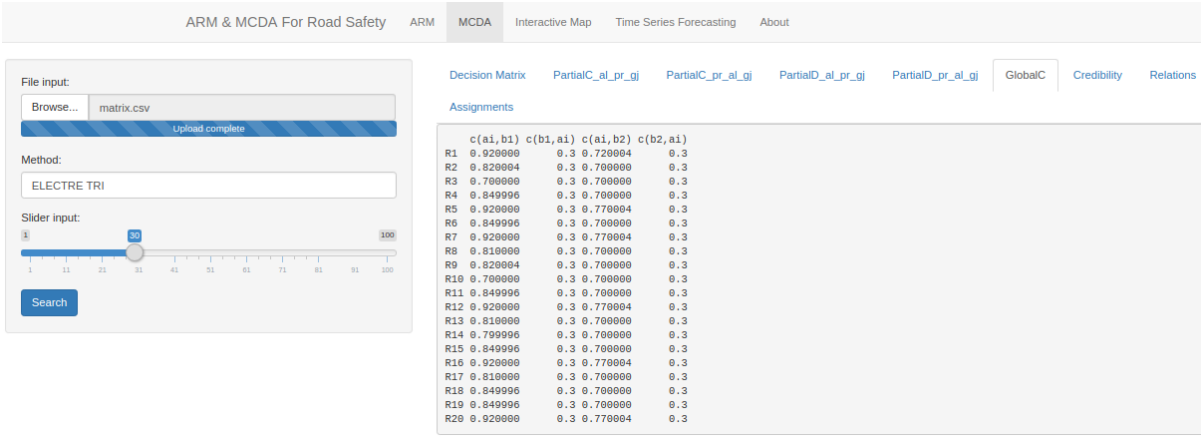

Figure A.8 – Indice de concordance globale

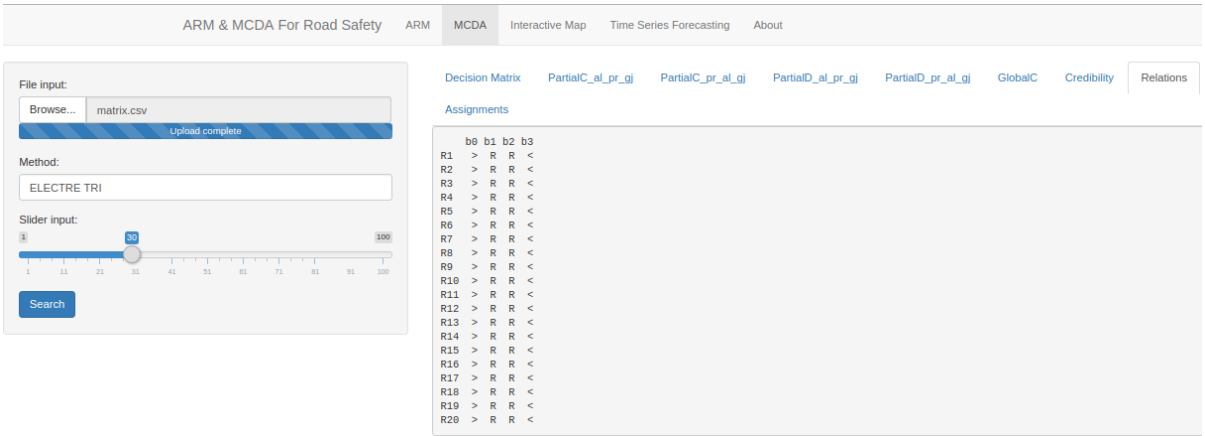

Figure A.9 – Affectation des règles aux catégories

Bibliographie

[1] P. S. Usama M. Fayyad, Gregory Piatetsky-Shapiro and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA, USA : ACM, 1996.

[2] S. Sayad, "An introduction to data mining." Accessed : 2017-06-30.

[3] G. P.-S. William J. Frawley and C. J. Matheus, "Knowledge discovery in databases an overview," *AI Magazine*, vol. 13, no. 10, pp. 57–70, 1992.

[4] T. I. R. Agrawal and A. Swami, "Mining association rules between sets of items in large databases," in *International Conference on Management of Data*, (Washington,U.S.A.), pp. 207–216, ACM, 2013.

[5] M. Zhang and C. He, *Survey on Association Rules Mining Algorithms*, pp. 111–118. Berlin, Heidelberg : Springer Berlin Heidelberg, 2010.

[6] L. Geng and H. J. Hamilton, "Interestingness measures for data mining : A survey," *AI Magazine*, vol. 38, no. 3, 2006.

[7] E. Alsolami, *An examination of keystroke dynamics for continuous user authentication*. PhD dissertation, Queensland University of Technology, 2012.

[8] K. Koperski, J. Adhikary, and J. Han, "Spatial data mining : Progress and challenges - survey paper," in *SIGMOD Workshop on Research Issues on data Mining and Knowledge Discovery (DMKD*, pp. 1–10, 1996.

[9] R. Marghoubi, A. Boulmakoul, and K. Zeitouni, "The use of the galois lattice for the extraction and the visualization of the spatial association rules," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 606–611, Aug 2006.

[10] K. Koperski and J. Han, "Discovery of spatial association rules in geographic information databases," in *Proceedings of the 4th International Symposium on Advances in Spatial Databases*, SSD '95, (London, UK, UK), pp. 47–66, Springer-Verlag, 1995.

[11] WHO, "World health organization." Accessed : 2017-06-30.

[12] Ministère, "Ministère de l'équipement, des transports et de la logistique." Accessed : 2017-06-30.

[13] E. O. Ashoka Savasere and S. B. Navathe., "An Efficient Algorithm for Mining Association Rules in Large Databases.," in *the 21th International Conference on Very Large Data Bases (VLDB '95)*, (San Francisco, CA, USA), pp. 432–444, 1995.

[14] D.-I. Lin and Z. M. Kedem, *Pincer-search : A new algorithm for discovering the maximum frequent set*, pp. 103–119. Berlin, Heidelberg : Springer Berlin Heidelberg, 1998.

[15] S. O. M. Zaki, M. J. ; Parthasarathy and W. Li, "New Algorithms for Fast Discovery of Association Rules," in *the Third Int l Conf. on Knowledge Discovery in Databases and Data Mining*, pp. 283–286, 1997.

[16] R. J. Bayardo, Jr., "Efficiently mining long patterns from databases," *SIGMOD Rec.*, vol. 27, pp. 85–93, June 1998.

[17] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient mining of association rules using closed itemset lattices," *Information Systems*, vol. 24, no. 1, pp. 25 – 46, 1999.

[18] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, (London, UK, UK), pp. 398–416, Springer-Verlag, 1999.

[19] N. Pernelle, M.-C. Rousset, H. Soldano, and V. Ventos, "Zoom : a nested galois lattices-based system for conceptual clustering," *J. Exp. Theor. Artif. Intell.*, vol. 14, pp. 157–187, 2002.

[20] M. J. Zaki and C.-J. Hsiao, "Charm : An efficient algorithm for closed association rule mining. technical report 99-10," tech. rep., Computer Science Dept., Rensselaer Polytechnic, October 1999.

[21] R. Kuo and C. Shih, "Association rule mining through the ant colony system for national health insurance research database in taiwan," *Computers and Mathematics with Applications*, vol. 54, no. 11, pp. 1303 – 1318, 2007.

[22] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, (San Francisco, CA, USA), pp. 487–499, Morgan Kaufmann Publishers Inc., 1994.

[23] J. D. U. Sergey Brin, Rajeev Motwani and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *the international conference on Management of data (SIGMOD '97)*, (New York, NY, USA), pp. 255–264, 1997.

[24] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *SIGMOD Rec.*, vol. 29, pp. 1–12, May 2000.

[25] E. L. G. Escovar, M. Biajiz, and M. T. P. Vieira, "Ssdm : A semantically similar data mining algorithm," in *SBBD*, 2005.

[26] J. B. et Alice Denoyel, "Cancer environnement." Accessed : 2017-06-30.

[27] W. Tobler, "The hyperelliptical and other new pseudo cylindrical equal area map projections," vol. 78, pp. 1753–1759, 04 1973.

[28] elokunnas T., "Object-oriented approaches applied to gis development," in *Acta Polytechnica Scandinavica, Mathematics and computing in engineering series No. 75*, 1995.

[29] O. F. Kraak M.J., "Cartography : Visualization of spatial data, 3rd edition," *The Canadian Geographer / Le Géographe canadien*, vol. 59, no. 1, pp. e7–e7, 2015.

[30] T. Armitage, *Getting started with oracle spatial.* Oracle Corporation, 2006.

[31] D. U., *Exploring geographical metadata by automatic and visual data mining.* PhD dissertation, oyal Institute of Technology, Stockholm, 2004.

[32] M. Ester, H.-P. Kriegel, and J. Sander, *Spatial data mining : A database approach*, pp. 47–66. Berlin, Heidelberg : Springer Berlin Heidelberg, 1997.

[33] S. Shekhar, P. Zhang, Y. Huang, and R. R. Vatsavai, "Trends in spatial data mining," 2004.

[34] M. Ester, H.-P. Kriegel, and J. Sander, "Knowledge discovery in spatial databases," in *Proceedings of the 23rd Annual German Conference on Artificial Intelligence : Advances in Artificial Intelligence*, KI '99, (London, UK, UK), pp. 61–74, Springer-Verlag, 1999.

[35] K. Koperski, J. Han, and N. Stefanovic, "An efficient two-step method for classification of spatial data," pp. 45–54, 1999.

[36] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets : Generalizing association rules to correlations," *SIGMOD Rec.*, vol. 26, pp. 265–276, June 1997.

[37] S. Panda and F. Somenzi, "Who are the variables in your neighborhood," in *Proceedings of the 1995 IEEE/ACM International Conference on Computer-aided Design*, ICCAD '95, (Washington, DC, USA), pp. 74–77, IEEE Computer Society, 1995.

[38] P. Lenca, P. MEYER, P. Picouet, B. Vaillant, and S. Lallich, "Critères d'évaluation des mesures de qualité des règles d'association," *Revue des Nouvelles Technologies de l'Information*, vol. RNTI-1, pp. 123–134, 2003.

[39] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, "On selecting interestingness measures for association rules : User oriented description and multiple criteria decision aid," *European Journal of Operational Research*, vol. 184, no. 2, pp. 610 – 626, 2008.

[40] J. Pearl, *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference.* San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1988.

[41] G. R. R. H. Lerman, I. C., "Élaboration et évaluation d'un indice d'implication pour des données binaires. i," *Mathématiques et Sciences Humaines*, vol. 74, pp. 5–35, 1981.

[42] J. Loevinger, *A systematic approach to the construction of tests of ability.* PhD dissertation, University of California, 1944.

[43] G. Piatetsky-Shapiro, "Discovery, analysis and presentation of strong rules," in *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro and W. J. Frawley, eds.), pp. 229–248, AAAI Press, 1991.

[44] N. Lavrač, P. Flach, and B. Zupan, *Rule Evaluation Measures : A Unifying View*, pp. 174–185. Berlin, Heidelberg : Springer Berlin Heidelberg, 1999.

[45] I. J. Good, *The estimation of probabilities : An essay on modern Bayesian methods.* Cambridge, MA : MIT Press, 1965.

[46] B. Roy, "ELECTRE III : Un algorithme de classements fondé sur une représentation floue des préférences en présence de critères multiples," *Cahiers du CERO*, vol. 20, no. 1, pp. 3–24, 1978.

[47] H. A. Simon, *The New Science of Management Decision.* Upper Saddle River, NJ, USA : Prentice Hall PTR, 1977.

[48] B. Roy, "Decision science or decision-aid science ?," *European Journal of Operational Research*, vol. 66, pp. 184–203, 1993.

[49] B. Roy, "À propos de la signification des dépendances entre critères : quelle place et quels modes de prise en compte pour l'aide à la décision ?," *RAIRO - Operations Research*, vol. 43, no. 3, p. 255–275, 2009.

[50] B. Roy and D. Bouyssou, *Aide Multicritère à la Décision : Méthodes et Cas.* Paris : Economica, 1993.

[51] B. Roy, "The outranking approach and the foundations of electre methods," *Theory and Decision*, vol. 31, pp. 49–73, Jul 1991.

[52] J. Brans, P. Vincke, and B. Mareschal, "How to select and how to rank projects : The promethee method," *European Journal of Operational Research*, vol. 24, no. 2, pp. 228 – 238, 1986. Mathematical Programming Multiple Criteria Decision Making.

[53] B. Mareschal and J. P. Brans, "The promethee-gaia decision support system for multicriteria investigations," ulb institutional repository, ULB – Universite Libre de Bruxelles, 1994.

[54] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338 – 353, 1965.

[55] P. Lenca, B. Vaillant, and S. Lallich, "On the robustness of association rules," in *2006 IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1–6, June 2006.

[56] Y. Le Bras, P. Meyer, P. Lenca, and S. Lallich, "A robustness measure of association rules," in *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases : Part II*, ECML PKDD'10, (Berlin, Heidelberg), pp. 227–242, Springer-Verlag, 2010.

[57] A. M. Addi, A. Tarik, G. Fatima, and D. Badi, "A choice of relevant association rules based on multi-criteria analysis approach," in *2015 5th International Conference on Information Communication Technology and Accessibility (ICTA)*, pp. 1–6, Dec 2015.

[58] A. Ait-Mlouk, F. Gharnati, and T. Agouti, "Multi-agent-based modeling for extracting relevant association rules using a multi-criteria analysis approach," *Vietnam Journal of Computer Science*, vol. 3, pp. 235–245, Nov 2016.

[59] V. Mousseau, J. Figueira, and J.-P. Naux, "Using assignment examples to infer weights for electre tri method : Some experimental results," *European Journal of Operational Research*, vol. 130, no. 2, pp. 263 – 275, 2001.

[60] A. Ait-Mlouk, F. Gharnati, and T. Agouti, "Multi-criteria decisional approach for extracting relevant association rules," *Int. J. Computational Science and Engineering*, vol. 3, pp. 235–245, Nov 2016.

[61] A. Ait-Mlouk, F. Gharnati, and T. Agouti, "An improved approach for association rule mining using a multi-criteria decision support system : a case study in road safety," *European Transport Research Review*, vol. 9, p. 40, Jul 2017.

[62] M. Hahsler and S. Chelluboina, "Visualizing association rules : Introduction to the r-extension package arulesviz," 2014.

[63] L. C. Dias and V. Mousseau, "Iris : a dss for multiple criteria sorting problems," *Journal of Multi-Criteria Decision Analysis*, vol. 12, no. 4-5, pp. 285–298, 2003.

[64] C.-H. Chen, L. P. Khoo, Y. T. Chong, and X. F. Yin, "Knowledge discovery using genetic algorithm for maritime situational awareness," *Expert Syst. Appl.*, vol. 41, pp. 2742–2753, May 2014.

[65] F. Shahzad, S. Asghar, and K. Usmani, "A fuzzy based scheme for sanitizing sensitive sequential patterns," *Int. Arab J. Inf. Technol.*, vol. 12, pp. 60–68, 2015.

[66] A. Kumar, A. Kakkar, R. Majumdar, and A. S. Baghel, "Spatial data mining : Recent trends and techniques," in *2015 International Conference on Computer and Computational Sciences (ICCCS)*, pp. 39–43, Jan 2015.

[67] R. Marghoubi, A. Boulmakoul, and K. Zeitouni, "The use of the galois lattice for the extraction and the visualization of the spatial association rules," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 606–611, Aug 2006.

[68] F. T. Chan and N. Kumar, "Global supplier development considering risk factors using fuzzy extended ahp-based approach," *Omega*, vol. 35, no. 4, pp. 417 – 431, 2007.

[69] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Information Sciences*, vol. 8, no. 3, pp. 199 – 249, 1975.

[70] CNPAC, "Cnpac : Comité national de prévention des accidents de la circulation." Accessed : 2017-06-30.

[71] A. Ait-Mlouk, F. Gharnati, and T. Agouti, "Intelligent transport system for road safety based data mining approach," *International Journal of Control and Automation*, vol. 10, no. 8, pp. 13–22, 2017.

[72] A. M. Addi, A. Tarik, and G. Fatima, "Comparative survey of association rule mining algorithms based on multiple-criteria decision analysis approach," in *2015 3rd International Conference on Control, Engineering Information Technology (CEIT)*, pp. 1–6, May 2015.

[73] B. Roy, "Classement et choix en présence de points de vue multiples," *RAIRO - Operations Research - Recherche Opérationnelle*, vol. 2, no. V1, pp. 57–75, 1968.

[74] Apache, "le projet hadoop." Accessed : 2017-06-30.

[75] Hadoop, "Hadoop." Accessed : 2017-06-30.

[76] J. Dean and S. Ghemawat, "System and method for efficient large-scale data processing," Jan. 19 2010. US Patent 7,650,331.

[77] Apache, "Spark." Accessed : 2017-06-30.

[78] Databriks, "Spark." Accessed : 2017-06-30.

[79] P. R, "R." Accessed : 2017-06-30.

[80] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, (New York, NY, USA), pp. 32–41, ACM, 2002.

[81] A. Ait-Mlouk, T. Agouti, and F. Gharnati, "Mining and prioritization of association rules for big data : multi-criteria decision analysis approach," *Journal of Big Data*, vol. 4, no. 1, p. 42, 2017.

[82] A. Ait-Mlouk, F. Gharnati, and T. Agouti, "pplication of big data analysis with decision tree for road accident," *Indian Journal of Science and Technology*, vol. 10, no. 29, pp. 1–10, 2017.

[83] wikipedia, "Méthode mathématique d'analyse multicritère." Accessed : 2017-06-30.

[84] G. M. J. G. E. P. Box, "Time series analysis forecasting and control," 1976.

[85] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Forecasting*, pp. 137–191. John Wiley and Sons, Inc., 2008.