UNIVERSITÉ FRANÇOIS RABELAIS

TOURS

École Doctorale SST

Laboratoire d'Informatique, Équipe BdTln

THÈSE présentée par :

Marie NDIAYE

soutenue le : 20 décembre 2010

pour obtenir le grade de : **Docteur de l'Université François Rabelais Tours**

Discipline / Spécialité : **Informatique**

EXPLORATION DE GRANDS ENSEMBLES DE MOTIFS

THÈSE dirigée par :

M. Arnaud Giacometti	Professeur des Universités, Université François Rabelais Tours
M. Cheikh Talibouya Diop	Maître de conférence, HDR, Université Gaston Berger de Saint Louis - Sénégal

RAPPORTEURS :

Mme Anne Laurent	Maître de Conférence, HDR, Université de Montpellier 2
M. Jean-Marc Petit	Professeur des Universités, INSA de Lyon

JURY :

M. Cheikh Talibouya Diop	Maître de conférence, HDR, Université Gaston Berger de Saint Louis - Sénégal
M. Arnaud Giacometti	Professeur des Universités, Université François Rabelais Tours
Mme Anne Laurent	Maître de Conférence, HDR, Université de Montpellier 2
M. Jean-Marc Petit	Professeur des Universités, INSA de Lyon
M. Arnaud Soulet	Maître de Conférence, Université François Rabelais Tours
Mme Karine Zeitouni	Professeur des Universités, Université de Versailles Saint-Quentin-en-Yvelines

Jàng yàggul, ñàkk jàng moo yàgg.

(Abdou Aziz Sy Dabakh)

À Mame Bigué Mbaye.

Remerciements

Je suis très reconnaissante envers Cheikh Talibouya Diop, Arnaud Giacometti, Patrick Marcel et Arnaud Soulet pour leurs conseils inlassables, leur patience et leurs encouragements. Ils ont su orienter mes travaux de recherche et leurs commentaires ont été très utiles pour améliorer la qualité de ma thèse. Je tiens également à remercier Mary Teuw Niane pour avoir accepté d'assurer la codirection officielle de cette thèse.

J'exprime ma profonde gratitude à Mme Anne Laurent et à M. Jean-Marc Petit qui m'ont fait l'honneur d'être les rapporteurs de mon mémoire. Leur lecture minutieuse et leurs remarques m'ont été très précieux. Je remercie également Mme Karine Zeitouni pour avoir accepté de faire partie de mon jury et pour ses remarques éclairées.

Ma thèse a été partiellement financée par le Service de Coopération et d'Action Culturelle de l'Ambassade de France au Sénégal. Je les remercie pour m'avoir mis dans d'excellentes conditions de travail.

Je remercie toute l'équipe de l'Antenne Universitaire de Blois, tant les enseignants-chercheurs que le personnel IATOS pour leur accueil chaleureux. Je remercie également les enseignants-chercheurs et le personnel administratif de l'UFR Sciences Appliquées et de Technologie de l'Université de Saint-Louis qui m'ont apportée un soutien inestimable.

Un grand merci à mes collègues de bureau pour leur soutien et leur enthousiasme à partager leur culture scientifique. Je pense aux anciens doctorants Adriana, Cheikh Ba, Ahmed Cheriat, Denio Duarte, Eynollah Khandjari, Hassina Mouloudi, Elsa Negre et Tonio Wandmacher. Je pense aussi aux futurs docteurs Julien Aligon, Lamine Baldé, Cheikh Niang, Damien Nouvel et Harinaina Ravelomanantsoa à qui je souhaite une très bonne fin de thèse.

Je voudrais remercier tous mes amis qui ont partagé les bons mais aussi les moins bons moments durant ces années. J'adresse un remerciement particulier à Philippe Marsault pour ses conseils, son soutien moral et son optimisme qu'il a su me transmettre. J'exprime ma profonde gratitude à Alioune Seck pour son soutien éprouvé et l'intérêt qu'il a porté à l'avancement de mes travaux. Il n'a jamais manqué de m'encourager dans les moments difficiles. J'ai une pensée particulière pour Michel Robert qui nous a quittés. Sa présence et nos échanges « philosophiques » me manquent beaucoup.

Je ne pourrai finir sans adresser un remerciement particulier à ma mère pour sa patience, sa compréhension et pour m'avoir offert l'opportunité d'effectuer de longues études. Ces remerciements vont aussi à l'endroit de mon père qui n'a jamais cesser de m'encourager. Je n'oublie pas mes frères et sœurs à qui je souhaite d'être passionné dans la quête du savoir. Enfin, je remercie tous les membres de ma famille au sens " africain " du terme. Je pense notamment à mon oncle Mamadou Ndiaye qui m'a toujours témoigné d'une grande affection, à ma tante Khady Ndiaye pour son soutien constant, à mon beau père Boubacar Diack pour ses encouragements, à mon oncle Famara Ibrahima Sagna pour son soutien et ses encouragements et à ma tante Fatou Ndiaye pour son accueil chaleureux lorsque je suis arrivée en France.

Résumé

L'Extraction de Connaissances à partir de Données (ECD) est une discipline dont l'objectif est de trouver de nouvelles connaissances, communément appelées motifs, à partir de bases de données. Elle repose sur des techniques issues de divers domaines tels que les bases de données, les statistiques ou encore l'intelligence artificielle. L'ECD est décrite comme un processus interactif qui consiste à préparer les données, extraire des connaissances à partir de ces données à l'aide d'algorithmes et interpréter les connaissances obtenues. L'interprétation des résultats d'extraction nécessite une exploration des connaissances. Dans ce mémoire, nous nous intéressons à cette étape d'exploration.

De nos jours, il existe beaucoup d'algorithmes d'extraction de connaissances. Ils produisent habituellement de grandes quantités de motifs. Pour faciliter l'exploration de ces motifs, deux approches sont souvent utilisées : la première approche consiste à résumer les ensembles de motifs extraits et la seconde approche repose sur la construction de représentations visuelles de ces motifs. Cependant, dans la plupart des travaux, les résumés ne sont pas structurés et ils sont proposés sans méthode d'exploration. D'autre part, les représentations visuelles n'offrent pas une vue globale des ensembles de motifs.

Notre première contribution est la définition d'un cadre générique qui permet de construire des résumés de grands ensembles de motifs à plusieurs niveaux de détail. Les résumés obtenus peuvent être structurés sous forme de cubes. Cette structuration permet d'explorer les ensembles de motifs via leurs résumés à l'aide d'opérateurs de navigation OLAP. Notre deuxième contribution est la proposition d'un algorithme qui fournit un premier résumé de taille inférieure à un seuil donné, pour initialiser l'exploration des motifs. Les résumés qu'il retourne sont obtenus en maximisant une mesure de qualité des résumés. Enfin, notre troisième contribution est l'instanciation de notre cadre avec les règles d'association. Dans ce contexte, nous proposons une mesure de qualité pour les résumés d'ensembles de règles d'association. Ensuite, nous testons notre algorithme sur des bases génériques de règles d'association en évaluant le temps d'exécution et la qualité des résumés qu'il produit.

Mots clés : Fouille de données, Cubes de données, Résumés d'ensembles de motifs, Règles d'association.

Abstract

Knowledge Discovery in Databases (KDD) is a discipline whose goal is to find from databases new knowledge commonly named patterns. It is based on techniques from various fields such as databases, statistics or artificial intelligence. KDD is described as an interactive process of preparing the data, extracting knowledge from data using algorithms and interpreting the obtained knowledge. The interpretation of extraction results requires to explore the knowledge. In this thesis, we focus at this step of exploration.

Nowadays, there are many algorithms for extracting knowledge. They usually produce large amounts of patterns. To facilitate the exploration of these patterns, two approaches are often used : the first approach is to summarize the sets of extracted patterns and the second approach relies on the construction of visual representations of these sets of patterns. However, in most work, the summaries are not structured and they are proposed without a method of exploration. Moreover, visual representations do not provide an overview of the sets of patterns.

Our first contribution is the definition of a generic framework for constructing summaries of large sets of patterns at different levels of detail. The obtained summaries can be structured in the form of cubes. This structure allows to explore the sets of patterns through their summaries with OLAP navigation operators.

Our second contribution is the proposal of an algorithm which generates a relevant cube based summary not exceeding a user-specified size, to initialize the exploration of a given rule set. The summaries generated by the algorithm are obtained by maximizing a quality measure of these. Finally, our third contribution is the instantiation of our framework with association rules. In this context, we propose a new quality measure for summaries of sets association rule sets. We test our algorithm on generic bases of association rules by evaluating the execution time and the quality of the summaries it produces.

Keywords : Data mining, Data cubes, Summarization of pattern sets, Association rules.

Table des matières

Liste des tableaux

Liste des figures

Introduction

Contexte

L'Extraction de Connaissances à partir de Données (ECD) est une discipline dont l'objectif est de trouver de nouvelles connaissances à partir de bases de données. Elle repose sur des techniques issues de divers domaines tels que les bases de données, les statistiques ou encore l'intelligence artificielle. L'ECD a commencé à se développer depuis le début des années 1990. Elle est aujourd'hui un allier incontournable dans plusieurs secteurs d'activités. Par exemple, elle est utilisée dans l'organisation des rayonnages dans les supermarchés pour regrouper des produits qui sont généralement achetés ensemble, les sites de vente en ligne pour proposer aux internautes des produits qui sont susceptibles de les intéresser, l'étude du génome humain pour identifier les molécules responsables de pathologies génétiques, etc.

L'ECD est décrite comme un processus interactif qui consiste à préparer les données, extraire des connaissances à partir de ces données à l'aide d'algorithmes et interpréter les connaissances obtenues [FPSS96]. L'interprétation des résultats d'extraction nécessite une exploration des connaissances. Dans ce mémoire, nous nous intéressons à cette étape d'exploration.

Problématique : exploration de grands ensembles de motifs

Les connaissances extraites, communément appelées motifs, s'expriment sous diverses formes. Par exemple, elles peuvent être représentées sous forme d'itemsets fréquents, de règles d'association, de séquences, de graphes, etc. De nos jours, il existe beaucoup d'algorithmes d'extraction de connaissances [AS94, Zak00, BPT+00]. Ces algorithmes produisent habituellement de grandes quantités de motifs. Pour faciliter l'exploration des motifs, deux approches sont souvent utilisées : la première approche consiste à résumer les ensembles de motifs extraits et la seconde approche repose sur la construction de représentations visuelles de ces motifs.

Des résumés qui ne permettent pas d'explorer les motifs Dans notre approche, nous définissons un résumé d'un ensemble de motifs comme étant une représentation synthétique de celui-ci, dont la taille peut être contrôlée directement ou indirectement. Le contrôle est direct si la taille exacte ou maximale du résumé peut être fixée. Il est indirect

si on peut agir sur des paramètres qui influent sur la taille du résumé. Plusieurs méthodes de construction de résumés ont été proposées dans la littérature [CK05, MM03]. Nous distinguons deux catégories de résumé : les résumés génératifs qui permettent de régénérer les motifs souvent avec leurs mesures d'intérêt et les résumés non génératifs qui ne peuvent pas permettre de déduire les motifs. Le principe des résumés génératifs est de faire un compromis entre la taille du résumé et la qualité de la régénération. Ainsi, ces résumés sont soit très synthétiques et l'erreur de régénération est plus importante, soit trop grands et ils sont dans ce cas difficiles à explorer. Par ailleurs, quelle que soit leur catégorie, les résumés ne sont pas organisés et ils sont proposés sans méthode d'exploration.

Des visualisations qui ne donnent pas une vue globale Beaucoup de travaux ont porté sur la construction de visualisations d'ensembles de motifs [BB04, LZBX06]. Cependant, les visualisations proposées sont limitées lorsque l'ensemble de motifs est grand. En effet, la représentation des motifs de tels ensembles aboutit à des visualisations qui sont trop grandes ou qui ne sont pas interprétables. Par conséquent, ces visualisations ne donnent pas une vue globale des ensembles de motifs.

Contribution

Nous constatons que les résumés proposés donnent une vue globale mais ils ne sont pas adaptés à l'exploration des motifs. D'autre part, les représentations visuelles n'offrent pas une vue globale des ensembles de motifs mais elles permettent de les explorer. Notre approche consiste à proposer des représentations qui regroupent les avantages de ces deux approches.

La première contribution de ce travail est la définition d'un cadre formel permettant de construire des résumés d'ensembles de motifs. La régénération n'étant pas de bonne qualité lorsque les résumés sont très synthétiques, nous orientons notre contribution sur les résumés non génératifs. Nos résumés sont basés sur un schéma qui permet de les représenter sous forme de cube. Cette représentation induit une structuration des résumés et permet d'explorer les motifs à l'aide d'opérateurs de navigation OLAP. D'autre part, notre cadre permet de résumer des ensembles de motifs à plusieurs niveaux de détail.

Notre seconde contribution est la proposition d'un algorithme qui fournit un premier résumé de taille inférieure à un seuil donné, pour initialiser l'exploration. Les résumés qu'il retourne sont obtenus en optimisant une mesure de qualité des résumés. Nous proposons d'utiliser des mesures de qualité monotones, i.e. des mesures telles qu'un résumé détaillé a une meilleure qualité qu'un résumé qui l'est moins. Nous considérons cette propriété car plus un résumé est détaillé, plus il fournit de l'information sur l'ensemble qu'il représente.

Enfin, notre troisième contribution est l'instanciation de notre cadre avec les règles d'association qui sont des motifs fréquemment utilisés pour l'aide à la décision. Dans ce contexte, nous proposons une mesure de qualité pour les résumés d'ensembles de règles d'association. Cette mesure repose sur l'entropie conditionnelle de Shannon. D'autre part, nous testons notre algorithme de construction de résumés sur des bases génériques de règles d'association en évaluant le temps d'exécution et l'homogénéité des résumés qu'il produit.

Organisation du mémoire

La première partie de ce mémoire dresse un état de l'art sur les méthodes de construction de résumés et les méthodes de visualisation. Plus précisément, nous présentons dans le chapitre 1 des définitions de base relatives aux motifs et aux résumés. Nous y positionnons également les résumés par rapport aux représentations condensées et à d'autres représentations synthétiques. Dans les chapitres 2 et 3, nous étudions respectivement des méthodes de construction de résumés et des méthodes de visualisation. Nous nous intéressons en particulier aux méthodes qui s'adressent aux itemsets fréquents et aux règles d'association. Nous soulignons dans chacun de ces chapitres, les avantages mais aussi les limites des méthodes proposées.

La deuxième partie présente nos contributions sur l'exploration de motifs. Dans le chapitre 4, nous décrivons notre cadre générique qui repose sur les cubes de données et notre algorithme de construction de résumés. Le chapitre 5 est consacré à l'instanciation de notre cadre avec les règles d'association. Nous décrivons également une nouvelle mesure de qualité pouvant être utilisé dans l'algorithme de résumé. Nous testons notre algorithme, en termes de temps d'exécution et de qualité des résumés, sur des bases de règles génériques en utilisant cette mesure de qualité.

Première partie

Exploration de motifs : état de l'art

Introduction

L'objectif de cette partie est de présenter un état de l'art sur les méthodes mises en œuvre pour une exploration efficace d'ensembles de motifs. Deux principales approches sont étudiées : la première approche consiste à résumer un ensemble de motifs afin d'obtenir une vue globale et synthétique des motifs tandis que la seconde approche est basée sur la présentation visuelle des motifs. Notons que ces deux approches peuvent être utilisées conjointement.

Le chapitre 1 introduit quelques notions de base sur les motifs et les résumés ainsi qu'un positionnement des résumés par rapport à d'autres modèles descriptifs.

Le chapitre 2 est consacré à l'étude des méthodes de construction de résumés. Nous distinguons deux grandes catégories de méthodes : celles qui produisent des résumés génératifs et celles qui produisent des résumés non génératifs. Les résumés génératifs permettent de régénérer les motifs et parfois leur mesure de qualité.

Le chapitre 3 décrit des méthodes de représentation visuelle d'ensembles de motifs. Ces méthodes utilisent divers supports. Nous distinguons globalement des représentations basées sur les tableaux, les matrices, les graphes et des caractéristiques géométriques.

Chapitre 1

Des motifs aux résumés

La fouille de données [1] est une étape du processus d'Extraction de Connaissances à partir de Données [2] (ECD) qui consiste à appliquer un algorithme sur un ensemble de données pour obtenir une connaissance [FPSS96]. Cette connaissance est fréquemment produite sous forme de motifs [3] qui sont ensuite utilisés pour l'aide à la prise de décision. Ce chapitre apporte quelques notions de base nécessaires à la compréhension de l'état de l'art sur les résumés et les représentations visuelles d'ensembles de motifs exposés dans les chapitres 2 et 3. Dans la section 1.1, nous rappelons quelques définitions concernant les données relationnelles qui sont très populaires dans le domaine de l'ECD. Ensuite, nous présentons dans la section 1.2 deux types de motifs souvent utilisés pour l'aide à la décision : les itemsets fréquents et les règles d'association. Puis, nous donnons des définitions relatives aux résumés dans la section 1.3. Enfin, dans la section 1.4, nous positionnons les résumés par rapport à d'autres modèles descriptifs qui sont proches de ces derniers.

1.1 Données relationnelles

Pour définir les données dans le cadre relationnel, nous nous appuyons sur le modèle relationnel [Cod70]. Une relation est caractérisée par un schéma $\mathcal{A}$ qui est un ensemble d'attributs et une instance qui est un ensemble de n-uplets. Chaque attribut $A \in \mathcal{A}$ prend ses valeurs dans un domaine noté $dom(A)$. Parmi ces attributs, il existe un attribut particulier que nous notons ID, dont les valeurs identifient les n-uplets de façon unique. Un n-uplet n est un élément du produit cartésien des domaines des attributs de $\mathcal{A}$, i.e. $n \in \times_{(A \in \mathcal{A})} dom(A)$. On note $dom(\mathcal{A})$ ce produit cartésien. Dans ce contexte, une donnée relationnelle définie sur $\mathcal{A}$ est un n-uplet de la relation de schéma A. Pour une meilleure lisibilité, nous utilisons dans ce manuscrit le terme "donnée relationnelle" ou simplement "donnée" pour désigner l'ensemble formé par les couples attribut-valeur d'une donnée relationnelle.

Le tableau 1.1 présente une relation concernant les modèles de téléphones portables proposés par un opérateur téléphonique à ses clients. Chaque ligne représente

1. Data Mining (DM) en anglais
2. Knowledge Discovery in Databases (KDD) en anglais
3. Patterns en anglais

une donnée et chaque colonne correspond à un attribut. Le schéma de la relation est : $\mathcal{A} = \{ID, Marque, Design, Connectivité, Écran, Autonomie, Appareil\ photo, PrixSansAbonnement\}$. La première colonne correspond à l'attribut ID et contient l'identifiant des données. Les autres attributs de $\mathcal{A}$ sont décrits ci-dessous.

- *Marque* : cet attribut décrit la marque des téléphones portables ;
- *Design* : les téléphones sont conçus en un seul bloc (monobloc) ou en mode coulissant ;
- *Connectivité* : les téléphones échangent des données avec d'autres médias via une connexion wifi (w), usb (u), bluetooth (b), etc ;
- *Écran* : les téléphones sont manipulés via leur écran (mode tactile) ou un clavier (mode non tactile) ;
- *Autonomie* : l'autonomie est le nombre d'heures durant lesquelles un portable peut fonctionner en mode appel.
- *Appareil photo* : cet attribut indique la taille des photos, en mégapixel (Mp), capturées par l'appareil photo des portables ;
- *PrixSansAbonnement* : le prix est celui des téléphones portables vendus sans abonnement.

Ainsi, nous avons pour ces attributs les domaines suivants :

- $dom(Marque) = \{Samsung, Nokia, Lg, Sony\ Ericsson\}$;
- $dom(Design) = \{monobloc, coulissant\}$;
- $dom(Connectivité) = \{b, ub, wub\}$;
- $dom(Écran) = \{tactile, non\ tactile\}$;
- $dom(Autonomie) = \{3h-5h, 6h-8h, 9h-11h\}$;
- $dom(Appareil\ photo) = \{2Mp-5Mp, 6Mp-9Mp, 10Mp-14Mp\}$;
- $dom(PrixSansAbonnement) = \{< 100\,€, 100\,€-200\,€, 200\,€-300\,€, > 200\,€\}$.

ID	*Marque*	*Design*	*Connectivité*	*Écran*	*Autonomie*	*Appareil photo*	*PrixSansAbonnement*
d_1	*Nokia*	*monobloc*	*wub*	*tactile*	$6h-8h$	$2Mp-5Mp$	$> 300\,€$
d_2	*Samsung Duo*	*monobloc*	*ub*	*tactile*	$3h-5h$	$2Mp-5Mp$	$100\,€-200\,€$
d_3	*Samsung*	*monobloc*	*wub*	*tactile*	$9h-11h$	$2Mp-5Mp$	$200\,€-300\,€$
d_4	*Sony Ericsson*	*monobloc*	*wub*	*tactile*	$9h-11h$	$10Mp-14Mp$	$> 300\,€$
d_5	*Sony Ericsson*	*monobloc*	*ub*	*tactile*	$3h-5h$	$6Mp-9Mp$	$> 300\,€$
d_6	*Samsung*	*coulissant*	*ub*	*non tactile*	$3h-5h$	$2Mp-5Mp$	$< 100\,€$
d_7	*Samsung*	*coulissant*	b	*non tactile*	$3h-5h$	$2Mp-5Mp$	$100\,€-200\,€$
d_8	*Lg*	*monobloc*	*ub*	*non tactile*	$3h-5h$	$2Mp-5Mp$	$< 100\,€$
d_9	*Lg*	*coulissant*	*ub*	*non tactile*	$3h-5h$	$2Mp-5Mp$	$200\,€-300\,€$
d_{10}	*Nokia*	*coulissant*	*ub*	*non tactile*	$3h-5h$	$2Mp-5Mp$	$100\,€-200\,€$
d_{11}	*Sony Ericsson*	*monobloc*	*wub*	*non tactile*	$9h-11h$	$2Mp-5Mp$	$100\,€-200\,€$

TABLE 1.1 – Modèles de téléphones portables

Les données de ce tableau seront utilisées tout au long de ce manuscrit pour illustrer les notions qui seront introduites.

1.2 Les motifs

Les motifs ont fait l'objet d'une multitude de travaux. Les sujets abordés concernent aussi bien leur extraction que leur traitement [BD03, HPYM04, Goe03]. Dans cette section,

nous présentons les itemsets fréquents qui sont les motifs les plus populaires dans le domaine de l'ECD ainsi que les règles d'association qui dérivent des itemsets fréquents.

1.2.1 Les itemsets fréquents

Les itemsets fréquents sont à l'origine générés pour la recherche de règles d'association [AIS93]. Ils sont utilisés de nos jours dans plusieurs autres tâches d'extraction de connaissances telles que la classification [LHM98], le clustering [WXL99], la recherche d'identification de corrélations [BMS97], etc.

Définition 1.2.1 (Item) *Étant donné un ensemble d'attributs $\mathcal{A}$, un item défini sur $\mathcal{A}$ est un couple attribut-valeur (A, a) tel que A est un élément de $\mathcal{A}$ et a appartient au domaine de A.*

Généralement, les items sont simplement notés par leur valeur si aucune ambiguïté n'est possible, i.e. si les domaines des attributs considérés sont deux à deux disjoints. Formellement, un itemset est défini de la manière suivante :

Définition 1.2.2 (Itemset) *Étant donné un ensemble d'attributs $\mathcal{A}$, un itemset X défini sur $\mathcal{A}$ est un ensemble d'items définis sur $\mathcal{A}$.*

Un itemset est aussi appelé un *k-itemset* s'il contient k items. Une donnée Y est couverte par un itemset X si et seulement si $X \subseteq Y$. Cette notion de couverture sera généralisée plus loin. Chaque itemset extrait à partir d'une collection de données est évalué par des mesures statistiques comme le *support* appelé aussi la *fréquence*.

Définition 1.2.3 (Support) *Le support d'un itemset X dans un ensemble de données $\mathcal{D}$, noté $sup(X, \mathcal{D})$, est le pourcentage de données de $\mathcal{D}$ qui sont couvertes par X.*

$$sup(X, \mathcal{D}) = \frac{|\{Y \mid Y \in \mathcal{D} \wedge X \subseteq Y\}|}{|\mathcal{D}|} \tag{1.1}$$

Un itemset X est dit fréquent dans $\mathcal{D}$ si son support dans $\mathcal{D}$ est plus grand qu'un seuil de support minimal *minsup* donné, i.e $sup(X, \mathcal{D}) \geq minsup$.

Exemple 1.2.1 *En fixant la valeur du support minimal à $minsup = 0.5$, on obtient à partir de l'ensemble de données du tableau 1.1, les itemsets fréquents suivants : $\{ub\}$, $\{non\ tactile\}$, $\{3h - 5h\}$, $\{2Mp - 5Mp\}$, $\{ub, 3h - 5h\}$, $\{non\ tactile, 2Mp - 5Mp\}$, $\{3h - 5h, 2Mp - 5Mp\}$. Le support des trois derniers itemsets est calculé comme suit :*

$$sup(\{ub, 3h - 5h\}, \mathcal{D}) = \frac{|\{d_2, d_5, d_6, d_8, d_9, d_{10}\}|}{|\mathcal{D}|} = \frac{6}{11} = 0.54$$

$$sup(\{non\ tactile, 2Mp - 5Mp\}, \mathcal{D}) = \frac{|\{d_6, d_7, d_8, d_9, d_{10}, d_{11}\}|}{|\mathcal{D}|} = \frac{6}{11} = 0.54$$

$$sup(\{3h - 5h, 2Mp - 5Mp\}, \mathcal{D}) = \frac{|\{d_2, d_6, d_7, d_8, d_9 d_{10}\}|}{|\mathcal{D}|} = \frac{6}{11} = 0.54$$

1.2.2 Les règles d'association

Les règles d'association sont des motifs introduits pour la première fois par [AIS93]. Elles sont traditionnellement utilisées pour *« l'analyse du panier de la ménagère »* dans le secteur de la distribution. L'objectif est de rechercher des associations entre produits sur les tickets de caisse [BMUT97]. De nos jours, leurs domaines d'application sont multiples. Nous pouvons citer entre autres, la détection de fraudes en recherchant des associations inhabituelles, la recherche de complications dues à des associations de médicaments, la gestion de la relation client, les services bancaires et de télécommunications. Intuitivement, une règle d'association est une implication du type *« si un téléphone portable a un écran tactile alors son prix est supérieur à 300 euros »* . Cette notion est formellement définie comme suit :

Définition 1.2.4 (Règle d'association) *Étant donné un ensemble d'attributs $\mathcal{A}$, une règle d'association définie sur $\mathcal{A}$ est une relation de la forme $X \Rightarrow Y$ où X et Y sont des itemsets définis sur $\mathcal{A}$ tels que $X \cap Y = \emptyset$.*

Une telle règle exprime la relation : si une donnée contient tous les items de X alors elle contient aussi tous les items de Y. X est appelé le corps ou l'antécédent et Y est appelé la tête ou la conséquence de la règle. Le support est aussi utilisé pour évaluer les règles d'association. Ainsi, le support de la règle $X \Rightarrow Y$ dans $\mathcal{D}$ correspond au support de l'union de sa tête et de son corps $X \cup Y$. Une règle est dite fréquente si son support est plus grand qu'un seuil de support *minsup* donné. En plus du support, les règles sont habituellement évaluées avec une seconde mesure d'intérêt appelée la *confiance*.

Définition 1.2.5 (Confiance) *La confiance d'une règle $X \Rightarrow Y$, notée $conf(X \Rightarrow Y, \mathcal{D})$, est la probabilité conditionnelle d'avoir Y dans une donnée sachant que cette donnée contient X.*

$$conf(X \Rightarrow Y, \mathcal{D}) = \frac{sup(X \cup Y, \mathcal{D})}{sup(X, \mathcal{D})} \tag{1.2}$$

Exemple 1.2.2 *La confiance de la règle $\{tactile\} \Rightarrow \{>300€\}$ est :*

$$conf(\{tactile\} \Rightarrow \{>300€\}, \mathcal{D}) = \frac{|\{d_1, d_4, d_5\}|}{|\{d_1, d_2, d_3, d_4, d_5\}|} = \frac{3}{5} = 0,6$$

Une règle est dite valide par rapport à la confiance si sa confiance est supérieure à un seuil de confiance *minconf* fixé. D'autre mesures telles que le lift ou la corrélation sont également utilisées pour mesurer la pertinence des règles d'association [TKS02, GH06a]. La notion de validité d'une règle est également employée par rapport à ces mesures de qualité.

L'extraction de motifs est de nos jours une tâche bien maîtrisée avec l 'existence d'une multitude d'algorithmes d'extraction. La problématique actuelle est l'analyse de ces motifs. Une des issues proposées dans la littérature consiste à résumer les ensembles de motifs qui sont générés.

1.3 Résumés d'ensembles de motifs

Les motifs sont généralement produits en grande quantité par les algorithmes d'extraction. Pour faciliter leur analyse, plusieurs approches ont été proposées dans la littérature. Elles vont de la réduction du nombre de motifs extraits [SVA97, NLHP98, Zak00], à la visualisation de ceux-ci [KGLB00, CHYN07, LZBX06] en passant par les modèles descriptifs [LHM99, PTB+05, YCHX05]. Nos travaux s'inscrivent dans le cadre des modèles descriptifs. En particulier, nous nous intéressons aux résumés pouvant être utilisés pour explorer des ensembles de motifs. Un résumé peut être défini comme une représentation synthétique d'un ensemble de motifs. Nous illustrons les notions et les définitions introduites dans cette section sur un ensemble de requêtes fréquentes.

1.3.1 Exemple de motifs : requêtes de sélection, projection et jointure

Considérons une base de données relationnelles $\mathcal{D}$ de schéma $sch(\mathcal{D}) = \{R_1, \ldots, R_N\}$ où chaque R_n, $n \in \{1, \ldots N\}$, est une relation. Dans nos exemples, nous nous intéressons à des requêtes simples qui effectuent des projections, des sélections et des jointures [DGLS04, JLS08, CPL+09] sur une base de données relationnelles. Ces requêtes s'expriment en algèbre relationnelle sous la forme suivante :

$$\pi_{A_1,\ldots,A_I}(\sigma_{B_1=b_1,\ldots,B_J=b_J}(R_1 \bowtie \ldots \bowtie R_K))$$

où $\{R_1, \ldots, R_K\}$ est un sous-ensemble de $sch(\mathcal{D})$, les A_i et les B_j appartiennent au schéma d'au moins une des relations R_k et pour chaque égalité $B_j = b_j$, $b_j \in dom(B_j)$. Une telle requête exprime la projection suivant les attributs $A_1, \ldots, A_I$ des n-uplets qui contiennent les items $(B_1, b_1), \ldots, (B_J, b_J)$. Ces n-uplets sont sélectionnés à partir de la jointure des domaines des attributs appartenant aux relations $R_1, \ldots, R_K$. Le tableau 1.2 montre une collection de requêtes sur la base de données relationnelles $\mathcal{D}$ de schéma $\{Téléphone, Forfait, Tarif\}$ contenant des informations concernant les forfaits et les téléphones portables proposés par des opérateurs téléphoniques où :

- $sch(Téléphone) = \{SérieTéléphone, Marque, Design, Connectivité, Écran, Autonomie, Photo, PrixSansAbonnement\}$
- $sch(Forfait) = \{CodeForfait, Opérateur, Type, Libellé, Appels, Textos, Internet\}$
- $sch(Tarif) = \{SérieTéléphone, CodeForfait, PrixAvecAbonnement\}$

1.3.2 Relation de couverture

Une relation de couverture est une relation binaire qui lie des motifs pouvant appartenir à un même langage ou à des langages différents. Dans le cas où elle concerne des motifs d'un même langage, on parle de relation de spécialisation/généralisation ou simplement de relation de spécialisation. Une telle relation est formellement définie comme suit :

Définition 1.3.1 (Relation de spécialisation/généralisation) *Soit un langage de motifs $\mathcal{L}$, une relation de spécialisation/généralisation sur $\mathcal{L}$, généralement notée $\preceq$, est une relation d'ordre partiel sur $\mathcal{L}$.*

q_1	$\pi_{Opérateur,PrixAvecAbonnement}(\sigma_{SérieTéléphone="Player5",Type="bloqué"}(Téléphone \bowtie Forfait \bowtie Tarif))$
q_2	$\pi_{Appels,Textos,Internet}(\sigma_{Type="illimité"}(Forfait))$
q_3	$\pi_{Design,PrixAvecAbonnement}(\sigma_{Marque="Samsung"}(Téléphone \bowtie Tarif))$
q_4	$\pi_{Marque,SérieTéléphone,PrixAvecAbonnement}(\sigma_{Appels="1h",Type="bloqué"}(Téléphone \bowtie Forfait \bowtie Tarif))$
q_5	$\pi_{Internet,CodeForfait}(\sigma_{Type="illimité",Marque="Sony\ Ericsson"}(Téléphone \bowtie Forfait \bowtie Tarif))$
q_6	$\pi_{Appels,CodeForfait}(\sigma_{Type="bloqué"}(Forfait))$
q_7	$\pi_{SérieTéléphone}(\sigma_{Marque="Samsung",Type="illimité"}(Téléphone \bowtie Forfait \bowtie Tarif))$
q_8	$\pi_{Appels}(\sigma_{Type=bloqué,Opérateur=Orange}(Forfait))$
q_9	$\pi_{SérieTéléphone}(\sigma_{Type="illimité",Opérateur="SFR",Marque="Sony\ Ericsson"}(Téléphone \bowtie Forfait \bowtie Tarif))$
q_{10}	$\pi_{SérieTéléphone}(\sigma_{Marque="Samsung",Écran="tactile"}(Téléphone))$
q_{11}	$\pi_{SérieTéléphone}(\sigma_{Connectivité="wub",Marque="Samsung",Type="bloqué"}(Téléphone \bowtie Forfait \bowtie Tarif))$
q_{12}	$\pi_{SérieTéléphone,PrixAvecAbonnement}(\sigma_{Marque="Sony\ Ericsson",Opérateur="Orange",Écran="tactile"}\ (Téléphone \bowtie Forfait \bowtie Tarif))$

Table 1.2 – Ensemble de requêtes fréquentes

Étant donnés deux motifs p et p' d'un langage $\mathcal{L}$, la notation $p \preceq p'$ signifie que p *est plus général* que p' et que p' *est plus spécifique* que p. Dans l'exemple suivant, nous définissons une relation de spécialisation sur le langage de requêtes.

Exemple 1.3.1 *Considérons les requêtes fréquentes de la figure 1.2. Intuitivement, on pourrait dire que* $q_6 = \pi_{Appels,CodeForfait}(\sigma_{Type="bloqué"}(Forfait))$ *est plus générale que* $q_8 = \pi_{Appels}(\sigma_{Type=bloqué,Opérateur=Orange}(Forfait))$ *car les informations sélectionnées en appliquant* q_8 *sont aussi sélectionnées en appliquant* q_6*. En effet, la projection de* q_8 *porte sur le même attribut (Appels) que celle de* q_6*. De même, la table interrogée par* q_6 *est aussi interrogée dans la requête* q_8*. Il s'agit en l'occurrence de la table* $Forfait$*. Enfin, la valeur sélectionnée par* q_6 *apparaît dans la liste de sélection de* q_8*. Cette relation est généralisée comme suit :*

Soit $\mathcal{L}$ *le langage de requêtes sur une base de données relationnelles. La relation de spécialisation sur* $\mathcal{L}$ *est définie telle que, étant données deux requêtes* $q = \pi_{A_1,\ldots,A_I}(\sigma_{B_1=b_1,\ldots,B_J=b_J}(R_1 \bowtie \ldots \bowtie R_K))$ *et* $q' = \pi_{A'_1,\ldots,A'_{I'}}(\sigma_{B'_1=b'_1,\ldots,B'_{J'}=b'_{J'}}(R'_1 \bowtie \ldots \bowtie R'_{K'}))$ *de* $\mathcal{L}$*,* q' *est plus spécifique que* q *si :*

- *Les projections de* q *et* q' *sont effectuées sur les mêmes attributs, i.e.* $\{A_1, \ldots, A_I\} = \{A'_1, \ldots, A'_{I'}\}$.
- *Les tables interrogées avec* q *sont aussi interrogées avec* q'*, i.e.* $\{R_1, \ldots, R_K\} \subseteq \{R'_1, \ldots, R'_{K'}\}$.
- *Les valeurs sélectionnées dans* q *le sont aussi dans* q'*, i.e.* $\{(B_1, b_1), \ldots, (B_J, b_J)\} \subseteq \{(B'_1, b'_1), \ldots, (B'_{J'}, b'_{J'})\}$.

La relation de couverture repose sur les relations de spécialisation définies sur les langages des motifs concernés.

Définition 1.3.2 (Relation de couverture) *Étant donnés deux langages de motifs partiellement ordonnés* $(\mathcal{P}, \preceq_{\mathcal{P}})$ *et* $(\mathcal{S}, \preceq_{\mathcal{S}})$*, une relation de couverture* $\triangleleft$ *entre* $\mathcal{P}$ *et* $\mathcal{S}$ *est une relation binaire telle que :*

1. *pour tout* p, $p' \in \mathcal{P}$ *et* $s \in \mathcal{S}$*, si* $p \preceq_{\mathcal{P}} p'$ *et* $s \triangleleft p$ *alors* $s \triangleleft p'$ *;*
2. *pour tout* $p \in \mathcal{P}$ *et* s, $s' \in \mathcal{S}$*, si* $s' \preceq_{\mathcal{S}} s$ *et* $s \triangleleft p$ *alors* $s' \triangleleft p$.

La *couverture* d'un motif $s \in \mathcal{S}$ dans $\mathcal{P}$, notée $couverture(s, \mathcal{P})$, est l'ensemble des motifs de $\mathcal{P}$ couverts par s.

$$couverture(s, \mathcal{P}) = \{p \mid (p \in \mathcal{P}) \wedge (s \triangleleft p)\}$$

Exemple 1.3.2 *Soit le langage* $\mathcal{P}$ *des requêtes qui peuvent être effectuées sur la base de données* $\mathcal{D}$ *décrite dans la section 1.3.1, muni de la relation de spécialisation décrite dans l'exemple 1.3.1. Soit le langage* $\mathcal{S}$ *des itemsets définis sur l'ensemble des attributs des relations de* $\mathcal{D}$ *muni de l'inclusion. Soit la relation de couverture définie par : un itemset* $\{(A'_1, a'_1), \ldots, (A'_M, a'_M)\} \in \mathcal{S}$ *couvre une requête* $q = \pi_{A_1,\ldots,A_I}(\sigma_{B_1=b_1,\ldots,B_J=b_J}(R_1 \bowtie \ldots \bowtie R_K)) \in \mathcal{P}$ *si chacun de ses items apparaît dans la sélection de* q*, i.e.* $\{(A'_1, a'_1), \ldots, (A'_M, a'_M)\} \subseteq \{(B_1, b_1), \ldots (B_J, b_J)\}$*. Cette relation vérifie les conditions (i) et (ii) de la définition 1.3.2. En effet, tout itemset qui couvre une requête couvre aussi celles qui sont plus spécifiques. Réciproquement, toute requête couverte par un itemset est couverte par les généralisations de ce dernier.*

Notons que dans certaines approches, une relation de couverture est utilisée pour construire des résumés. Le principe de ces méthodes est de trouver un résumé qui couvre au mieux un ensemble de motifs. La couverture peut être totale (tous les motifs de l'ensemble sont couverts) ou partielle (il existe des motifs qui ne sont pas couverts).

1.3.3 Construction de résumés

La construction d'un résumé s'effectue en utilisant une méthode qui peut être décrite par un algorithme ou une fonction. Par exemple, dans le domaine de résumé de texte, une méthode possible consiste à sélectionner la phrase la plus représentative dans chaque paragraphe et à regrouper ces phrases pour former un résumé. Une autre méthode à pour principe la sélection des mots du texte les plus récurrents pour constituer le résumé.

D'autre part, une méthode de résumé ne serait pas intéressante si elle n'offrait pas la possibilité de contrôler la taille des résumés qui sont produits. En effet, s'il n'y a pas de contrôle de la taille, les résumés générés peuvent certes être plus petits que les ensembles qu'ils représentent mais ils peuvent être trop grands pour être explorés efficacement. Par exemple, si un ensemble de 3000 motifs est résumé avec un ensemble de 900 motifs, il y aura moins de motifs qu'au départ mais le résumé sera encore trop volumineux. Par conséquent, il serait difficile à interpréter.

Dans notre approche, nous considérons qu'une méthode de résumé est une fonction définie comme suit :

Définition 1.3.3 (Fonction de résumé) *Soient $\mathcal{P}$ et $\mathcal{S}$ deux langages de motifs. Une fonction de résumé est une fonction $\Psi_\alpha : 2^{\mathcal{P}} \rightarrow 2^{\mathcal{S}}$ qui associe à un ensemble de motifs $P \subseteq \mathcal{P}$, un ensemble de motifs $S \subseteq \mathcal{S}$ tel que :*

(i) $|S| \leq |P|$ *;*

(ii) Le paramètre α permet de contrôler la taille de S.

L'ensemble de motifs S est un résumé de P.

Le contrôle de la taille du résumé peut être direct ou indirect. Il est direct lorsque le paramètre α correspond à la taille du résumé désiré ou lorsqu'il majore celle-ci, i.e. pour tout $\alpha \in \mathbb{N}$ et pour tous $P \subseteq \mathcal{P}$, $|\Psi_\alpha(P)| \leq \alpha$. Ce contrôle est indirect quand α agit sur des caractéristiques du résumé qui influent de manière monotone sur sa taille. En d'autres termes, pour tous α_1 et α_2 appartenant à un langage muni d'une relation d'ordre $\preceq$, si $\alpha_1 \preceq \alpha_2$ alors $|\Psi_{\alpha_1}(P)| \leq |\Psi_{\alpha_1}(P)|$. Nous présenterons dans le chapitre 2, des méthodes de construction de résumés offrant ces deux moyens de contrôle. Notons qu'il existe des méthodes pour lesquelles la fonction de résumé s'applique directement aux données initiales. Cette définition s'adresse plus particulièrement aux méthodes de construction de résumés qui sont appliquées en post-traitement, c'est à dire après l'extraction des motifs. Elle peut toutefois être adaptée aux méthodes en pré-traitement.

1.3.4 Mesure de qualité de résumé

Des mesures sont généralement utilisées pour évaluer la qualité des résumés par rapport à des critères fixés au départ. Plusieurs mesures de qualité provenant de divers domaines

sont étudiées dans [GH06b, KH06]. De manière générale, une mesure de qualité peut être définie comme suit :

Définition 1.3.4 (Mesure de qualité de résumé) *Étant donné deux langages de motifs $\mathcal{P}$ et $\mathcal{S}$, une mesure de qualité de résumé est une fonction $\phi : 2^{\mathcal{P}} \times 2^{\mathcal{S}} \rightarrow \mathbb{R}$ qui à un ensemble de motifs P et un résumé S de cet ensemble associe une valeur réelle notée $\phi(P, S)$.*

Une mesure de qualité simple utilisée dans [CK05] est le taux de compression qui est le rapport entre la taille de l'ensemble de motifs initial P et la taille de son résumé S , i.e. $\phi(P, S) = \frac{|P|}{|S|}$.

Nous distinguons deux types d'utilisation d'une mesure de qualité de résumé. La première utilisation consiste à s'en servir comme une heuristique pour construire les résumés. La seconde utilisation est l'évaluation de la représentativité des résumés.

1.4 Positionnement par rapport à d'autres modèles descriptifs

La construction de résumés d'ensembles de motifs s'inscrit dans un domaine plus large qui est celui de la construction de modèles descriptifs. Dans cette section, nous situons les résumés d'ensembles de motifs par rapport aux représentations condensées et à d'autres modèles descriptifs synthétiques.

1.4.1 Les représentations condensées

Comme nous l'avons souligné au début de la section 1.3 (cf. page 35), le résultat des algorithmes d'extraction est généralement très volumineux. Lors de l'exploration des motifs, le premier problème rencontré concerne le stockage des motifs. En effet, les motifs étant très nombreux, leur stockage et le calcul de leurs mesures de qualité nécessitent des ressources très importantes. Les représentations condensées sont utilisées pour réduire les motifs extraits afin de faciliter leur stockage, leur traitement ou encore leur interrogation. La notion de représentation condensée a été introduite pour la première fois dans [MT96]. Intuitivement, étant donné un langage de motifs $\mathcal{P}$, une représentation condensée d'un ensemble de motifs $P \subset \mathcal{P}$ est un ensemble de motifs $S \subset \mathcal{P}$ qui doit être aussi concis que possible et qui permet de répondre à une classe de requêtes sans accéder à P.

La plupart des représentations condensées proposées dans la littérature permettent de dériver des itemsets fréquents et leur support [CRB04]. Des approches visant à retrouver d'autres mesures de qualité des motifs sont aussi proposées dans [SC08]. Par ailleurs, certains travaux s'adressent à la construction de représentations condensées pour des ensembles de règles d'association [Kry98, BPT+00]. Ces représentations sont communément appelées des bases génériques de règles.

Les représentions condensées sont classées en deux catégories [Dio03] : les *représentations condensées exactes* et les *représentations condensées approximatives*.

- **Les représentations condensées exactes** Une représentation condensée exacte permet de retrouver les résultats exacts d'une classe de requêtes. Une des premières méthodes de représentation condensée exacte de motifs fréquents proposées repose sur les itemsets fréquents maximaux [BR98]. Les fréquents maximaux sont utilisés pour régénérer tous les itemsets fréquents mais ils ne permettent pas de retrouver le support exact de ceux qui ne sont pas maximaux. Des représentations permettant de conserver le support des motifs sont par la suite proposées. Les plus connues sont les itemsets fréquents fermés [PTB+05], les itemsets fréquents non dérivables [CG07], les itemsets libres appelés aussi itemsets clés [BTP+00] ou encore les itemsets libres disjonctifs [BR01] qui constituent une extension des itemsets libres. Cependant, la taille de ces représentations reste généralement importante. Ainsi, d'autres approches proposent des représentations plus petites qui donnent une valeur approximative du support des motifs.
- **Les représentations condensées approximatives** Une représentation condensée approximative fournit une approximation des résultats d'une classe de requêtes. Beaucoup de méthodes de représentation condensée approximative de motifs ont été proposées ces dix dernières années [BB00, BBR00, YCHX05, PG09]. Le principe général de ces méthodes est de faire un compromis entre la taille de la représentation et la précision sur le support des motifs. Ainsi, les représentations obtenues permettent de régénérer les motifs avec une approximation de leur support.

Nous distinguons trois principales différences entre les représentations condensées et les résumés :

- Les représentations condensées sont conçues dans le but de régénérer les motifs des ensembles qu'elles représentent et le plus souvent de retrouver aussi leurs mesures d'intérêt. Par contre, la régénération n'est pas nécessaire pour les résumés.
- Les méthodes proposées pour construire des représentations condensées exactes ne permettent pas de contrôler la taille des représentations tandis que les méthodes de résumé ont un paramètre destiné au contrôle de la taille des résumés. Notons que certaines méthodes de représentation condensée approximative offrent la possibilité de contrôler indirectement la taille des représentations. Nous considérons de telles représentations comme étant des résumés. Nous reviendrons sur ces représentations dans le chapitre 2.
- Les motifs des ensembles à représenter et les motifs des représentations condensées appartiennent au même langage. Par contre les motifs des résumés peuvent être pris dans un langage différent de celui des motifs des ensembles à résumer.

1.4.2 Autres modèles descriptifs synthétiques

Certaines méthodes de représentation synthétique d'ensembles de motifs sont considérées dans la littérature comme des méthodes de construction de résumés. Cependant, elles ne permettent pas de contrôler la taille de ces représentations. Nous exposons dans les sections 1.4.2.1 et 1.4.2.2 deux de ces approches. Pour chacune d'elles, nous présentons un exemple de représentation de l'ensemble de règles d'association qui est décrit dans le tableau 1.3. Les règles sont extraites à partir des données du tableau 1.1.

	Tête	*Corps*
r_1	$\{tactile\}$	$\{>300€\}$
r_2	$\{monobloc\}$	$\{>300€\}$
r_3	$\{Sony\ Ericson, tactile, monobloc\}$	$\{>300€\}$
r_4	$\{Samsung\}$	$\{3h-5h\}$
r_5	$\{Sony\ Ericson\}$	$\{9h-11h\}$
r_6	$\{Sony\ Ericson, wub\}$	$\{9h-11h\}$

Table 1.3 – Ensemble de règles d'association

1.4.2.1 Couverture d'un ensemble de règles d'association

L'approche proposée dans [OES06] s'adresse aux règles d'association. Elle consiste à représenter un ensemble de règles avec un sous-ensemble de celui-ci en utilisant une relation de couverture entre les règles ayant la même tête. La relation de couverture est définie comme suit : étant données deux règles $r : X \Rightarrow Y$ et $r' : X' \Rightarrow Y$, r couvre r' si $X' \subseteq X$. La représentation est construite telle que chaque règle de l'ensemble est couverte par au moins une règle de la représentation. Pour construire une représentation d'un ensemble de motifs P, les auteurs divisent P en groupes de règles ayant la même tête. Ensuite, un algorithme glouton est appliqué à chaque groupe pour trouver un sous-ensemble qui couvre la totalité des règles du groupe. L'union de ces sous-ensembles forme une représentation synthétique de P. L'exemple suivant montre une représentation d'un ensemble de règles.

Exemple 1.4.1 *Considérons l'ensemble P des règles du tableau 1.3. Il peut être résumé par $S = \{r_3, r_4, r_6\}$ sachant que r_1 et r_2 et r_3 sont couvertes par r_3 ; r_4 est couverte par elle-même ; r_5 et r_6 sont couvertes par r_6.*

La taille de la représentation ne peut pas être contrôlée avec cette méthode. La fonction de construction utilisée n'est donc pas une fonction de résumé au sens de la définition 1.3.3.

1.4.2.2 Rule cover

Dans [TKR+95], les auteurs proposent une méthode pour construire une représentation appelée "rule cover" [4], d'un ensemble de règles d'association. Cette méthode repose aussi sur une relation de couverture qui est définie entre un langage de données et un langage de règles d'association portant sur les mêmes attributs que les données. Dans leur approche, une règle $X \Rightarrow Y$ couvre une donnée d si $X \cup Y \subseteq d$. Intuitivement, le résumé d'un ensemble de règles P est un sous-ensemble S de celui-ci qui couvre les mêmes données que P :

$$\bigcup_{X \Rightarrow Y \in P} couverture(X \Rightarrow Y, \mathcal{D}) = \bigcup_{X' \Rightarrow Y \in S} couverture(X' \Rightarrow Y, \mathcal{D}) \tag{1.3}$$

où $\mathcal{D}$ est la collection de données à partir de laquelle les règles de P sont extraites. Dans ce contexte, trouver une "rule cover" d'un ensemble de règles P revient à résoudre le problème qui consiste à trouver le plus petit sous-ensemble de P qui couvre les mêmes

4. couverture de règles

données que P. Ce problème est une variante du problème classique de couverture d'ensemble [5] qui est NP-complet. Les auteurs proposent un algorithme glouton qui fournit une solution approchée. Il consiste à choisir itérativement une règle de P qui couvre le plus de données jusqu'à ce que toutes les données de $\mathcal{D}$ soient couvertes. L'ensemble constitué par les règles sélectionnées représente un "rule cover" de P. L'exemple ci-dessous montre une représentation de l'ensemble de règles du tableau 1.1.

Exemple 1.4.2 *En s'appuyant sur les couvertures des règles décrites dans le tableau 1.4, une représentation possible de l'ensemble* $\{r_1, \cdots, r_6\}$ *est* $\{r_1, r_4, r_5\}$ *car* r_1, r_4 *et* r_6 *couvrent toutes les données couvertes par les règles de cet ensemble.*

	Tête	*Corps*	*Couverture*
r_1	$\{tactile\}$	$\{> 300€\}$	$\{d_2, d_4, d_5\}$
r_2	$\{monobloc\}$	$\{> 300€\}$	$\{d_2, d_4, d_5\}$
r_3	$\{Sony\ Ericson, tactile, monobloc\}$	$\{> 300€\}$	$\{d_4, d_5\}$
r_4	$\{Samsung\}$	$\{3h - 5h\}$	$\{d_2, d_6, d_7\}$
r_5	$\{Sony\ Ericson\}$	$\{9h - 11h\}$	$\{d_4, d_{11}\}$
r_6	$\{Sony\ Ericson, wub\}$	$\{9h - 11h\}$	$\{d_4\}$

TABLE 1.4 – Données couvertes par les règles d'association du tableau 1.3

Les tests effectués sur des données réelles montrent que cette approche permet de produire des représentations de taille relativement réduite. Cependant, il est évident qu'elle ne permet pas de contrôler la taille de la représentation. Par conséquent, la fonction de construction correspondant à cette méthode n'est pas une fonction de résumé au sens de la définition 1.3.3.

Le principal problème de ces deux approches est que la taille des représentations n'est pas maîtrisée. Le fait de les appliquer aux règles de classification, i.e. les règles dont la tête est constituée d'un seul item, permet de limiter en amont la taille potentielle des résumés car le nombre de classes possibles est généralement réduit. Par conséquent, elles ne sont pas adaptées aux règles d'association classiques qui peuvent avoir plusieurs items dans la tête.

1.5 Conclusion

Dans ce chapitre, nous avons présenté les itemsets fréquents et les règles d'association qui sont utilisés dans plusieurs tâches d'extraction. Ces motifs sont généralement produits en grande quantité par les algorithmes d'extraction. Des modèles descriptifs tels que les résumés sont alors proposés pour faciliter leur analyse. Un résumé d'un ensemble de motifs est une représentation synthétique de cet ensemble dont la taille peut être contrôlée. Nous avons positionné les résumés par rapport aux représentations condensées et à d'autres modèles descriptifs synthétiques. Les résumés se distinguent de ces modèles par le fait que leur taille peut être contrôlée. Nous étudions dans le chapitre 2 différentes méthodes de

5. set covering problem

construction de résumés incluant des méthodes de représentation condensée approximative dont la fonction de construction permet de contrôler la taille des représentations.

Chapitre 2

Étude de diverses méthodes de construction de résumés

Le résumé est un moyen fréquemment utilisé pour représenter synthétiquement de grands ensembles de motifs issus des algorithmes d'extraction. De nombreuses méthodes de construction de résumés ont été proposées au cours de ces dernières années [MM03, AGM04, CK05, YCHX05, PG09, LHM99, JXL09]. Certaines produisent des résumés génératifs, d'autres donnent des résumés non génératifs. Les résumés génératifs permettent de régénérer les motifs souvent avec leur support tandis que les résumés non génératifs ne permettent pas de reconstituer les motifs. Nous décrivons dans ce chapitre quelques méthodes de construction de résumés. La section 2.1 détaille les caractéristiques des méthodes de construction de résumés sur lesquelles nous nous basons pour les décrire. Les sections 2.2 et 2.3 présentent des méthodes qui produisent respectivement des résumés génératif et des résumés non génératifs. Enfin, la section 2.4 fait une synthèse sur les méthodes de construction de résumés.

2.1 Caractéristiques des méthodes de construction de résumés

Le tableau 2.1 présente une catégorisation des résumés qui sont produits par les méthodes que nous étudions dans ce chapitre. Rappelons que les méthodes de construction de résumés sont celles qui permettent de contrôler la taille des représentations qui sont construits. Nous avons reporté dans la colonne grisée du tableau, les modèles descriptifs que nous avons exposés dans le chapitre précédent et qui ne sont pas des résumés car leur taille ne peut pas être contrôlée. Nous distinguons deux grandes catégories de résumés : les résumés *génératifs* et les résumés *non génératifs*. Les résumés génératifs permettent de régénérer les motifs des ensembles de départ avec leurs mesures d'intérêt le plus souvent. Les résumés qui entrent donc dans cette catégorie sont des représentations condensées approximatives. Les résumés non génératifs ne permettent de retrouver ni les motifs ni leurs mesures d'intérêt. Par ailleurs, les représentations condensées exactes permettent évidemment de régénérer les motifs tandis que les autres représentations synthétiques sont des

modèles non génératifs. Les méthodes de construction de résumés présentent des caracté-

		Contrôle de la taille du modèle		Pas de contrôle de la taille du modèle
		Direct	Indirect	
Modèle	Génératif	[AGM04] (section 2.2.1.1) [YCHX05] (section 2.2.1.2) [MM03] (section 2.2.1.3)	[PG09] (section 2.2.2.1)	[BR98] [PTB+05] [CG07] [BR01] (section 1.4.1)
	Non génératif	[Mac67] (section 2.3.1.1) [KR05] (section 2.3.1.1) [CK05] (section 2.3.1.2)	[LHM99] (section 2.3.2.1)	[OES06] (section 1.4.2.1) [TKR+95] (section 1.4.2.2)

TABLE 2.1 – Catégorisation des modèles descriptifs

ristiques qui peuvent servir d'appui pour les étudier. Dans notre étude, nous identifions cinq caractéristiques que nous détaillons ci-dessous.

- **Le langage des motifs à résumer et le langage des motifs des résumés** La construction d'un résumé fait intervenir deux langages : le langage des motifs de l'ensemble à résumer et celui des motifs du résumé. Il est donc essentiel de savoir à quel type de motif on peut appliquer les méthodes de construction de résumés. D'autre part, le résumé étant le produit final qui sera analysé par l'utilisateur, il est tout aussi important de connaitre le type de motifs qui le compose.
- **La couverture de l'ensemble de motifs à résumer** Certaines méthodes utilisent une relation de couverture pour construire les résumés. Dans ce contexte, il s'agit d'identifier si l'approche proposée permet d'avoir des résumés qui couvrent tous les motifs de l'ensemble de départ ou pas.
- **La régénération** Pour les méthodes qui produisent des résumés génératifs, nous décrivons la méthode de régénération proposée pour les motifs ou pour leur mesure d'intérêt si elle peut être estimée.
- **La mesure de qualité des résumés** Une mesure de qualité est souvent définie dans les approches étudiées. Elle est utilisée dans certaines méthodes comme un critère d'évaluation des résumés qui sont produits. D'autres méthodes l'utilisent plutôt comme un critère à optimiser lors de la construction des résumés.
- **Contrôle de la taille des résumés** La taille des résumés est un facteur important pour les méthodes de construction de résumés. En effet, un résumé très grand est difficilement exploitable alors qu'un résumé de taille réduite est plus facile à explorer mais il est généralement sujet à une perte d'information plus importante. Nous préciserons pour chaque méthode le moyen proposé pour contrôler la taille des résumés.

2.2 Les résumés génératifs

Les méthodes que nous décrivons dans cette section produisent des résumés génératifs. Plus précisément, les résumés obtenus sont des représentations condensées approximatives. Rappelons que ces représentations condensées ont pour principe de faire un compromis entre la taille des représentations et la qualité de la régénération des motifs. Nous présentons dans les sections 2.2.1 et 2.2.2 des méthodes qui permettent de contrôler directement et indirectement la taille des résumés.

2.2.1 Contrôle direct de la taille des résumés

Les méthodes que nous allons détailler à présent permettent de construire des représentations condensées en privilégiant leur taille.

2.2.1.1 Approximation d'une collection d'itemsets

Langages Dans [AGM04], les auteurs proposent une méthode de construction de résumés d'ensembles d'itemsets fréquents. Les résumés sont aussi constitués d'itemsets. L'objectif est d'obtenir un résumé qui permet de régénérer une approximation d'un ensemble d'itemsets.

Relation de couverture Pour construire un résumé, les auteurs utilisent une relation de couverture définie sur un langage d'itemsets. Cette relation de couverture correspond à l'inclusion. Dans leur approche, un itemset X couvre un autre itemset Y si Y est inclus dans X.

Régénération Ainsi, le résumé doit être construit tel que $P = \bigcup_{Y \in S} \mathcal{P}(Y)$ où $\mathcal{P}(Y)$ est l'ensemble des parties de Y. Chaque itemset du résumé permet de régénérer tous les itemsets qu'il couvre, i.e. tous ses sous-ensembles.

Exemple 2.2.1 *Soit la collection formée par les itemsets* $X_1 = \{Samsung, monobloc, wub\}$, $X_2 = \{Samsung, monobloc, tactile\}$, $X_3 = \{Samsung, wub, tactile\}$, $X_4 = \{Samsung, > 300€\}$ *et* $X_5 = \{monobloc, > 300€\}$ *et tous leurs sous-ensembles. Cette collection peut être représentée approximativement par* $\{Y_1, Y_2\}$ *avec* $Y_1 = \{Samsung, monobloc, wub, tactile\}$ *et* $Y_2 = \{Samsung, monobloc, > 300€\}$. *En effet,* X_1, X_2 *et* X_3 *et leurs sous-ensembles sont couverts par* Y_1. *De même* X_4, X_5 *et leurs sous-ensembles sont couverts par* Y_2.

Mesure de qualité Une mesure $\phi(P, S)$ appelée couverture est définie pour évaluer la qualité des résumés. Elle correspond à la taille de l'ensemble des itemsets de P qui sont couverts par le résumé :

$$\phi(P, S) = \left| P \cap \left(\bigcup_{Y \in S} \mathcal{P}(Y) \right) \right| \tag{2.1}$$

Cette mesure de qualité est utilisée comme heuristique pour construire les résumés.

Contrôle de la taille des résumés Les auteurs formulent le problème de construction d'un résumé d'un ensemble P comme étant un problème de recherche d'un ensemble S de K itemsets qui maximise la couverture $\phi(P, S)$. Notons que le contrôle direct de la taille du résumé se fait par le choix de K qui est le nombre d'itemsets que l'utilisateur souhaite avoir dans le résumé. Les auteurs proposent deux approches pour résoudre ce problème. Ces approches se distinguent par la collection des candidats à partir de laquelle les itemsets du résumé sont sélectionnés.

Dans la première approche, ces itemsets sont choisis parmi ceux de l'ensemble de motifs P. Par contre, dans la seconde approche, ils sont sélectionnés à partir des itemsets fréquents maximaux de P [1].

La seconde approche est moins coûteuse que la première car l'ensemble des candidats est plus réduit. Cependant, elle introduit des faux positifs lors de la régénération, i.e. des itemsets qui n'appartiennent pas à P.

2.2.1.2 Les profils

Langages Les profils ont été introduits dans [YCHX05] pour résumer des collections d'itemsets. Intuitivement, un profil est un itemset auquel on associe un support et un vecteur de probabilités qui permettent d'approximer le support des itemsets qui sont des sous-ensembles de l'itemset du profil. Il est formellement défini comme suit :

Définition 2.2.1 (Profil) *Étant donnés un ensemble de données $\mathcal{D}$, un ensemble P d'itemsets fréquents extraits à partir de $\mathcal{D}$ et $\mathcal{D}' = \cup_{p \in P} \mathcal{D}_p$ où $\mathcal{D}_p$ est l'ensemble des données de $\mathcal{D}$ contenant p, un profil M sur P est le triplet $\langle V, X, \sigma \rangle$ tel que :*

- *X, appelé motif maître, est l'union des itemsets de P, i.e. $X = \bigcup_{p \in P} p$.*
- *V est un vecteur qui est composé des supports V_a des items $a \in X$ dans $\mathcal{D}'$.*
- *$\sigma = \frac{|\mathcal{D}'|}{|\mathcal{D}|}$ est le support du profil.*

Exemple 2.2.2 *Considérons l'ensemble $\mathcal{D}$ de données décrits dans le tableau 1.1 et l'ensemble d'itemsets $P = \{\{Samsung\}, \{monobloc, wub, tactile\}, \{Samsung, tactile\}\}$ avec $\mathcal{D}' = \{d_1, d_2, d_3, d_4, d_6, d_7\}$. Les données de $\mathcal{D}'$ sont rappelées dans le tableau 2.2. Le profil sur P est constitué des composantes suivantes :*

- *$X = \{Samsung, monobloc, wub, tactile\}$*
- *$V = \langle V_{Samsung}, V_{monobloc}, V_{wub}, V_{tactile} \rangle$ avec $V_{Samsung} = \frac{4}{6}$, $V_{monobloc} = \frac{4}{6}$, $V_{wub} = \frac{3}{6}$ et $V_{tactile} = \frac{4}{6}$*
- *$\sigma = \frac{6}{11}$*

	Marque	Design	Connectivité	Écran	Autonomie	Appareil photo	Prix
d_1	Nokia	monobloc	wub	tactile	6h-8h	2Mp-5Mp	>300€
d_2	Samsung	monobloc	ub	tactile	3h-5h	2Mp-5Mp	100€ - 200€
d_3	Samsung	monobloc	wub	tactile	9h-11h	2Mp-5Mp	200€ - 300€
d_4	Sony Ericsson	monobloc	wub	tactile	9h-11h	10Mp-14Mp	>300€
d_6	Samsung	coulissant	ub	Non tactil	3h-5h	2Mp-5Mp	<100€
d_7	Samsung	coulissant	b	Non tactil	3h-5h	2Mp-5Mp	100€ - 200€

TABLE 2.2 – Transactions appartenant à l'ensemble de données $\mathcal{D}$'

Couverture L'inclusion est utilisée comme relation de couverture. Un profil couvre un itemset si son motif maître est un sur-ensemble de l'itemset. Ainsi, un résumé d'un ensemble d'itemsets P est défini comme un ensemble de profils S tel que chaque itemset de P est couvert par au moins un profil de S. Donc l'ensemble d'itemsets P doit être totalement couvert.

1. Un itemset X est maximal dans un ensemble d'itemsets P s'il n'existe pas d'itemset fréquent $X' \in P$ tel que $X \subset X'$.

Régénération Les profils permettent de régénérer les itemsets de l'ensemble de départ et d'approximer leur support. Tous les itemsets qui sont couverts par un profil peuvent être régénérés à partir de ce même profil. Le support d'un itemset p couvert par un profil $\langle V, X, \sigma \rangle$ est estimé par le support du profil qui est multiplié par les probabilités associées aux items de p :

$$\widehat{sup}(p, \mathcal{D}) = \sigma \times \prod_{a \in p} V_a \tag{2.2}$$

Si cet itemset est couvert par plusieurs profils, le support est estimé par le plus grand des supports calculés à partir de ces profils. Dans [JAAXR08], les auteurs proposent de l'approximer par la moyenne de ces supports.

Mesure de qualité Les auteurs proposent deux mesures pour évaluer la qualité des résumés. La première mesure détermine la qualité des profils pris individuellement. Cette qualité correspond à la probabilité de générer un motif maître à partir de son profil. Étant donné le profil $\langle X, V, \sigma \rangle$, la probabilité $f(Y)$ de générer son motif maître est calculée par la formule suivante :

$$f(X) = \prod_{a \in X} V(a) \tag{2.3}$$

Plus cette probabilité se rapproche de 1, meilleure est la qualité du profil.

La seconde mesure évalue la qualité totale du résumé. Elle calcule la moyenne des erreurs relatives entre le support exact et le support estimé des itemsets de l'ensemble de départ :

$$\phi(P, S) = \frac{1}{|P|} \sum_{p \in P} \frac{|sup(p, \mathcal{D}) - \widehat{sup}(p, \mathcal{D})|}{\widehat{sup}(p, \mathcal{D})} \tag{2.4}$$

Ces mesures ne sont pas directement utilisées dans la construction des résumés mais les auteurs montrent que $\phi(P, S)$ est optimisée dans les méthodes de construction proposées.

Contrôle de la taille des résumés Étant donné un ensemble d'itemsets P, la construction de résumés est considérée comme un problème de recherche d'un ensemble de K profils qui couvrent les itemsets de P. Les auteurs proposent deux méthodes pour résoudre ce problème. Ces méthodes reposent chacune sur un algorithme de clustering qui prend en entrée le nombre de clusters souhaité et qui utilise comme distance une mesure de similarité appelée la divergence de Kullback-Leibler [Kul59]. Le clustering avec cette distance correspond à l'optimisation de $\phi(P, S)$.

La première méthode débute par la construction d'un profil individuel pour chaque itemset. Le motif maître du profil est évidemment l'itemset lui-même. Ensuite, un algorithme de clustering hiérarchique ascendant est appliqué à l'ensemble de ces profils sachant qu'un profil correspond à un cluster. A chaque étape de l'algorithme, les deux profils les plus proches (selon la divergence de Kullback-Leibler) sont fusionnés en un profil. Le motif maître de ce nouveau profil correspond au profil dont le motif maître est l'union des motifs maîtres des deux profils qui sont regroupés. Son vecteur de probabilités ainsi que son support sont ensuite déterminés à partir des données.

La seconde méthode consiste à appliquer d'abord l'algorithme des K-means sur l'ensemble des itemsets P. Ensuite, un profil est construit pour chaque cluster. Au début de l'algorithme, K itemsets sont choisis au hasard comme des représentants de clusters. A chaque étape, chaque itemset de P est assigné au cluster dont le profil du représentant est plus proche de son profil (selon la divergence de Kullback-Leibler). Puis les représentants des clusters sont recalculés. Le représentant d'un cluster correspond à l'union des itemsets du cluster.

Nous remarquons que le contrôle de la taille du résumé se fait lors de la construction des clusters en fixant le nombre de clusters qui correspond au nombre de profils du résumé.

La première méthode est très couteuse car elle nécessite de calculer la distance entre le nouveau profil et les autres profils à chaque étape, ce qui implique plusieurs passages sur les données. Par ailleurs, les deux approches ne permettent pas de maîtriser l'erreur d'approximation sur le support des motifs. D'autre part, on peut régénérer à partir d'un profil, des itemsets qui n'appartiennent pas à l'ensemble d'itemsets sur lequel ce profil est défini. Par exemple, le profil sur l'ensemble P construit dans l'exemple 2.2.2 permet de régénérer l'itemset $\{Samsung, wub\}$ qui n'appartient pas à P. Pour résoudre ces problèmes, d'autres approches sont proposées dans [WP06, JAAXR08, PG09].

2.2.1.3 The pattern ordering problem [2]

Langages Dans [MM03], les auteurs s'intéressent en particulier aux représentations condensées, comme les itemsets fréquents fermés, qui permettent de retrouver le support des motifs. L'objectif est de trouver un compromis entre le nombre de motifs choisis et la précision sur l'approximation du support des motifs. Leur méthode consiste à ordonner les motifs d'un ensemble tel que chaque sous-ensemble des K premiers motifs permette d'approximer le support des motifs avec plus de précision que le sous-ensemble d'ordre $K-1$. Notons que l'approximation peut porter sur d'autres mesures d'intérêt telles que le lift. Nous décrivons cette méthode en nous basant sur le cas des itemsets fréquents fermés.

Couverture Les auteurs choisissent l'inclusion comme relation de couverture pour la régénération des itemsets. Ainsi, un itemset X est couvert par un itemset Y si $X \subseteq Y$.

Régénération Un itemset peut être régénéré à partir de n'importe quel itemset du résumé. Par contre, son support est estimé en considérant tous les itemsets qui le couvrent. Les auteurs proposent d'approximer le support d'un itemset $p \in P$ à partir du résumé S par le maximum des supports des sur-ensembles de p contenus dans S s'il en existe au moins un et par 0 sinon. Cette approximation, notée $\widehat{f}(p)$ est donnée par la formule suivante :

$$\widehat{f}(p) = \max\{f(s) \mid s \in S, p \subseteq s\} \cup \{0\} \tag{2.5}$$

Mesure de qualité La qualité des résumés est évaluée avec une mesure de perte d'information qui quantifie l'erreur générée en estimant le support des motifs de la collection à partir du résumé. Étant donnés un ensemble de motifs P et un résumé S de celui-ci, la

2. Le problème d'ordonnancement de motifs

formule 2.6 correspond à une mesure typique de perte d'information qui peut être utilisée pour ordonner les motifs.

$$\phi(P,S) = \sqrt{\sum_{p\in P}\left(f(p) - \widehat{f}(p)\right)^2} \tag{2.6}$$

où $f(p)$ est la valeur de la mesure d'intérêt du motif p est $\widehat{f}(p)$ est une approximation de cette valeur qui est calculée à partir du résumé S.

Exemple 2.2.3 *Soit un ensemble d'itemsets* $P = \{p_1, p_2, p_3, p_4\}$ *avec* $p_1 = \{monobloc, wub\}$, $p_2 = \{monobloc, tactile\}$, $p_3 = \{wub, tactile\}$ *et* $p_4 = \{monobloc, wub, tactile\}$ *et le support de chacun de ces itemsets :* $\sigma_1 = \frac{4}{11}$, $\sigma_2 = \frac{5}{11}$, $\sigma_3 = \frac{3}{11}$ *et* $\sigma_4 = \frac{3}{11}$. *Supposons que nous disposons d'un premier résumé* $S_0 = \{\}$. *Nous souhaitons ajouter un itemset choisi parmi les itemsets de* p *pour obtenir* S_1. *La mesure de qualité utilisée est le support. Rappelons que le support d'un itemset* p *est estimé par le maximum des supports des itemsets du résumé qui sont des sur-ensembles de* p *s'il en existe au moins un et par* 0 *sinon. Pour déterminer le motif à ajouter, nous calculons la perte d'information pour chaque combinaison. Ainsi, nous obtenons :*

- $\phi(P,\{p_1\}) = \sqrt{\left(\frac{4}{11} - \frac{4}{11}\right)^2 + \left(\frac{5}{11} - 0\right)^2 + \left(\frac{3}{11} - 0\right)^2 + \left(\frac{3}{11} - 0\right)^2} = 0.35$
- $\phi(P,\{p_2\}) = \sqrt{\left(\frac{4}{11} - 0\right)^2 + \left(\frac{5}{11} - \frac{5}{11}\right)^2 + \left(\frac{3}{11} - 0\right)^2 + \left(\frac{3}{11} - 0\right)^2} = 0.28$
- $\phi(P,\{p_3\}) = \sqrt{\left(\frac{4}{11} - 0\right)^2 + \left(\frac{5}{11} - 0\right)^2 + \left(\frac{3}{11} - \frac{3}{11}\right)^2 + \left(\frac{3}{11} - 0\right)^2} = 0.41$
- $\phi(P,\{p_4\}) = \sqrt{\left(\frac{4}{11} - \frac{3}{11}\right)^2 + \left(\frac{5}{11} - \frac{3}{11}\right)^2 + \left(\frac{3}{11} - \frac{3}{11}\right)^2 + \left(\frac{3}{11} - \frac{3}{11}\right)^2} = 0.04$

Donc, le résumé d'ordre 1 *est* $S_1 = \{p_4\}$.

Contrôle de la taille des résumés Les auteurs formulent la construction d'un résumé comme un problème de minimisation de perte d'information. Ils montrent que ce problème d'ordonnancement est NP-complet et proposent un algorithme glouton pour approximer la solution optimale. Dans ce contexte, pour tout K donné, $K \leq |P|$, on peut obtenir un résumé de taille K.

Le problème principal de cette méthode est que plus K est petit, plus l'erreur d'approximation est grande, cette erreur est accentuée par le fait que le support des motifs qui n'ont pas de sur-ensemble dans le résumé est estimé automatiquement par 0. Ce défaut est d'ailleurs commun à toutes les méthodes décrites précédemment qui produisent des résumés génératifs avec un contrôle direct de la taille des résumés. Nous présentons dans la section suivante une méthode qui permet de contrôler la taille des résumés en agissant sur l'erreur d'approximation.

2.2.2 Contrôle indirect de la taille des résumés

Dans cette section, nous nous intéressons à une méthode qui privilégie la qualité de régénération des motifs.

2.2.2.1 Les c-profils

Langages Dans l'approche utilisant des profils que nous avons exposée dans la section 2.2.1.2, un itemset donné peut être couvert par plusieurs profils. Dans ce cas de figure, les auteurs proposent d'approximer le support d'un tel itemset avec la moyenne des supports calculés à partir de ces profils, ce qui entraine une erreur d'approximation élevée. Poernomo et Gopalkrishnan introduisent dans [PG09] une extension des profils afin d'approximer avec plus de précision le support des itemsets et d'éviter de couvrir des faux positifs. Ces profils étendus sont appelés des c-profils. Intuitivement, un c-profil est un profil auquel est associé un itemset appelé base qui indique les itemsets dont le support peut être régénéré à partir de ce profil. La définition suivante donne la formalisation d'un c-profil.

Définition 2.2.2 *Étant donné un ensemble de motifs P, un c-profil est un motif de la forme $Y[C]sup(Y,\mathcal{D})|\langle V,X,\sigma\rangle$ où :*
- *Y est un itemset appelé base.*
- *C est l'ensemble d'items tel que pour tout $c \in C$, $Y \cup \{c\}$ a le même support que Y.*
- *$sup(Y,\mathcal{D})$ est le support de Y.*
- *$\langle V,X,\sigma\rangle$ est un profil.*

Couverture Un c-profil $Y[C]sup(Y,\mathcal{D})|\langle V,X,\sigma\rangle$ couvre les sur-ensembles de Y qui contiennent un sous-ensemble de $Y \cup C \cup X$, i.e. $\{X' \mid Y \subseteq X' \subseteq Y \cup C \cup X\}$.

Exemple 2.2.4 *$\{Samsung\}[\{2Mp-5Mp\}]\frac{4}{11}|\langle\langle 1,1\rangle,\{monobloc,tactile\},\frac{5}{11}\rangle$ est un c-profil qui couvre tous les itemsets $\{X \mid \{Samsung\} \subseteq X \subseteq \{Samsung\} \cup \{2Mp-5Mp\} \cup \{monobloc,tactile\}\}$.*

Un résumé, appelé cp-summary, correspond à un ensemble de c-profils. Les résumés proposés sont tels que chaque itemset de l'ensemble de départ est couvert par un seul c-profil du résumé.

Régénération Le support d'un itemset p couvert par un c-profil $Y[C]sup(Y,\mathcal{D})|\langle X,V,\sigma,\rangle$ est estimé par $\widehat{sup}(p,\mathcal{D})$ dont la valeur est donnée par la formule suivante :

$$\widehat{sup}(p,\mathcal{D}) = \begin{cases} sup(Y,\mathcal{D}) & si\ Y \cap X = \emptyset \\ sup(Y,\mathcal{D}) \times \sigma \times \left(\prod_{a \in p \setminus (Y \cup C)} V(a) \right) & sinon \end{cases} \tag{2.7}$$

Mesure de qualité Les auteurs ne proposent pas une mesure de qualité particulière pour évaluer les résumés mais ils garantissent que le support des itemsets est estimé avec précision. En d'autres termes, si un itemset p n'est pas fréquent alors $\widehat{sup}(p,\mathcal{D})$ est inférieur au seuil de support utilisé lors de l'extraction. Par contre, si cet itemset est fréquent alors son support est non seulement plus grand que ce seuil, mais il est en plus estimé avec une

erreur $err(p)$ inférieure à un seuil de tolérance ϵ fixé. Cette erreur d'approximation est déterminée par :

$$err(p) = \begin{cases} 1 - \frac{\widehat{sup}(p,\mathcal{D})}{sup(p,\mathcal{D})} & si\ \widehat{sup}(p,\mathcal{D}) < sup(p,\mathcal{D}) \\ 1 - \frac{sup(p,\mathcal{D})}{\widehat{sup}(p,\mathcal{D})} & sinon \end{cases} \tag{2.8}$$

Contrôle de la taille des résumés Un algorithme de construction de résumés est proposé par les auteurs. Il fournit un résumé qui permet de régénérer les itemsets d'un ensemble donné avec une erreur d'approximation de leur support qui ne dépasse pas un seuil fixé. Contrairement à la méthode de construction proposée pour les profils, la taille des résumés ne peut pas être fixée directement mais elle peut être modifiée en faisant varier le seuil d'erreur. Cependant, les expérimentations décrites par les auteurs montrent qu'une variation du seuil d'erreur entraine une variation très faible de la taille des résumés. Cette approche n'est donc pas intéressante en terme de contrôle de la taille des résumés.

2.3 Les résumés non génératifs

Comme dans la section précédente, nous commençons par décrire des méthodes avec un contrôle direct de la taille des résumés et nous finissons avec les méthodes permettant un contrôle indirect.

2.3.1 Contrôle direct de la taille des résumés

Ces méthodes reposent sur la notion de représentant. Chaque motif de l'ensemble à résumer est représenté par au moins un motif du résumé. Le contrôle direct porte sur le choix du nombre de représentants qui constituent le résumé.

2.3.1.1 Méthodes de clustering par partitionnement

Le clustering est une approche descriptive visant à trouver des groupes d'objets similaires appelés clusters ou classes. Les méthodes de clustering par partitionnement consistent à partitionner un ensemble d'objets en K clusters disjoints. Elles sont basées sur les notions de représentant de cluster et de distance entre objets. Chaque motif de l'ensemble partitionné est associé au cluster dont le représentant est plus proche en terme de distance. L'ensemble formé par les représentants des clusters constitue une description synthétique de l'ensemble de motifs qui est partitionné. Nous décrivons dans cette section deux méthodes de clustering de ce type : les K-means [Mac67] et les K-medoïdes [KR05].

Les K-means Pour cette méthode, le représentant d'un cluster, appelé centroïde, est le point médian des objets de ce cluster. Notons que le centroïde ne correspond presque jamais à un objet du cluster. La méthode des K-means est décrite par l'algoritme suivant :

1. Sélectionner K objets comme les centroïdes initiaux ;
2. Affecter chaque objet de l'ensemble au cluster dont le centroïde est le plus proche ;
3. Recalculer le centroïde de chaque cluster ;

4. Répéter les étapes 2 et 3 jusqu'à ce que les centroïdes ne changent plus.

La proximité d'un cluster avec un centroïde est déterminée avec une mesure de distance. Les centroïdes sont calculés en fonction des objets du cluster. Les centroïdes ne changent plus lorsqu'un minimum local est atteint, i.e. une solution qui minimise la somme des distances entre les objets et le centroïde du cluster auquel ils sont rattachés.

Les K-médoïdes À la différence des K-means, le représentant d'un cluster pour les K-médoïdes, appelé médoïde, correspond à un objet du cluster. Cet objet est le plus représentatif du cluster suivant une mesure de distance donnée. L'algorithme suivant décrit la méthode des K-médoïdes.

1. Sélectionner K objets comme les medoïds initiaux ;
2. Calculer le coût de toutes les configurations possibles obtenues en changeant un médoïde par un objet non médoïde ;
3. Sélectionner la configuration de plus faible coût ;
4. Si cette configuration a un coût plus faible que celle de la configuration courante alors revenir à l'étape 2 ;
5. Associer chaque objet non médoïde au médoïde le plus proche.

Conceptuellement, le calcul du coût d'une configuration se fait comme suit : la distance entre chaque objet non médoïde et son médoïde le plus proche est calculée. La somme de ces distances représente le coût de la configuration. Cette méthode est évidemment coûteuse puisqu'il faut calculer le coût de toutes les configurations possibles à chaque étape.

Langages Les méthodes de clustering par partitionnement sont applicables lorsqu'une mesure de distance peut être définie entres les objets du langage utilisé.

Couverture Chaque représentant de cluster couvre les objets de son cluster. Par conséquent, l'ensemble qui est partitionné est totalement couvert. Notons que cette notion de couverture est différente de la définition 1.3.2 de couverture que nous avons données au chapitre 1. En effet, la relation de couverture employée est basée sur la distance entre les objets alors que celle que nous avons définie repose sur la relation de spécialisation/généralisation entre les motifs.

Mesure de qualité La somme des distances entre les objets et leur représentant est la mesure de qualité utilisée comme heuristique pour construire les partitions.

Contrôle de la taille des résumés Les méthodes de partitionnement sont des fonctions de résumé avec un contrôle direct de la taille des résumés. En effet, fixer le nombre de clusters revient à fixer la taille du résumé puisque chaque cluster est représenté par un motif.

Ces méthodes sont mal adaptées lorsqu'on veut résumer des ensembles qui contiennent des motifs aberrants car ces motifs sont obligatoirement affectés à un cluster, ce qui dégrade la description de leur cluster, i.e. le représentant de celui-ci.

2.3.1.2 Résumés individuels

Langages L'approche développée dans [CK05] a pour objectif de résumer une collection de transactions[3] $\mathcal{D}$ par une collection plus réduite $S = \{Y_1, \ldots, Y_I\}$ d'itemsets appelés résumés individuels.

Couverture Un résumé individuel couvre une transaction s'il est inclus dans cette transaction. Chaque transaction de $\mathcal{D}$ est couverte par au moins un résumé individuel et chaque résumé individuel couvre un sous-ensemble de $\mathcal{D}$.

Mesure de qualité Pour évaluer la qualité d'un résumé S d'un ensemble de transactions $\mathcal{D}$, les auteurs proposent deux mesures : le gain de compacité et la perte d'information. Le gain de compacité est la réduction obtenue lorsqu'on passe de la collection de transactions au résumé. Il est déterminé par le ratio entre le nombre de transactions et le nombre de résumés individuels comme le montre la formule 2.9.

$$\phi_1(\mathcal{D}, S) = \frac{|\mathcal{D}|}{|S|} \tag{2.9}$$

La perte d'information est définie comme la quantité totale d'information contenue dans la collection de transactions et qui est absente du résumé. Étant donnée une transaction d couverte par un résumé individuel Y, l'information contenue dans d qui est perdue étant donné Y est évaluée par $q(d, Y) = \sum_{(A,a)\in d} W(A) \times |\{(A, a)\} \cap Y|$ où $W(A)$ est un poids associé à l'attribut A et $|\{(A, a)\} \cap Y|$ vaut 1 si (A, a) appartient à Y et 0 sinon. La perte d'information totale est déterminée en effectuant la somme des pertes d'information des transactions étant donné leur meilleur résumé individuel. Le meilleur résumé individuel d'une transaction est celui qui fournit la plus petite perte d'information. La perte d'information totale est définie par la formule 2.10.

$$\phi_2(\mathcal{D}, S) = \sum_{d\in\mathcal{D}} q(d, Y_d^*) \tag{2.10}$$

où $Y_d^* = \arg\min\{q(d, Y) \mid Y \in S\}$ est le meilleur résumé individuel de d contenu dans S.

Dans l'exemple suivant, nous évaluons le gain de compacité et la perte d'information d'un résumé.

Exemple 2.3.1 *Soit* $S = \{Y_1, Y_2, Y_3\}$ *un résumé de l'ensemble* $\mathcal{D}$ *du tableau 1.1 avec* $Y_1 = \{monobloc, tactile\}$, $Y_2 = \{Samsung, coulissant, non\ tactile, 3h - 5h, 2Mp - 5Mp\}$ *et* $Y_3 = \{monobloc, wub, 9h - 11h, 2Mp - 5Mp\}$. *Le gain de compacité du résumé vaut* $\phi_1(\mathcal{D}, S) = \frac{11}{3}$. *Maintenant, nous souhaitons calculer la perte d'information totale du résumé* S *en considérant que les poids associés aux attributs sont tous égaux à* 1. *Le tableau 2.3 affiche pour chaque transaction la perte d'information étant donnés les résumés individuels contenus dans* $\mathcal{D}$. *La dernière ligne du tableau contient la perte d'information de chaque transaction étant donné son meilleur résumé individuel. Ainsi, la perte d'information totale de* S *est* $\phi_2(\mathcal{D}, S) = 12$.

3. Une transaction est un ensemble d'items.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
Y_1	3	2	2	3	3	0	0	1	0	0	1
Y_2	1	3	2	0	1	5	5	3	4	4	2
Y_3	3	2	4	3	1	1	1	2	1	1	4
$q(d, Y_d^*)$	1	2	2	0	1	1	1	1	1	1	1

TABLE 2.3 – Perte d'information des transactions

Contrôle de la taille des résumés La construction d'un résumé pour un ensemble de transactions est considérée comme un problème de double optimisation du gain de compacité et de la perte d'information. Les auteurs proposent plusieurs méthodes qui permettent de contrôler directement la taille des résumés. Nous décrivons ci-dessous deux de ces méthodes. La première méthode est basée sur un algorithme de clustering et la seconde méthode utilise une représentation condensée pour construire un résumé.

- La première méthode consiste à partitionner l'ensemble de transactions en appliquant un algorithme de clustering classique. La distance utilisée pour former les clusters est basée sur les poids associés aux attributs des items. Après la construction des clusters, un résumé individuel est déterminé pour chaque cluster. Le résumé individuel d'un cluster correspond à l'intersection de ses transactions. Notons que le nombre de clusters est précisé au départ. Par conséquent, le nombre de résumés individuels est contrôlé.
- La seconde méthode consiste à considérer l'ensemble des transactions comme le premier résumé. Puis, un meilleur candidat est choisi, de manière itérative, dans l'union de l'ensemble des transactions et l'ensemble des itemsets fréquents fermés générés à partir de cet ensemble de transactions. Le meilleur candidat est celui qui minimise la perte d'information et qui maximise le nombre de transactions couvertes. A chaque étape, le meilleur candidat est ajouté au résumé tandis que tous les itemsets pour lesquels ce candidat est le meilleur résumé individuel sont supprimés de l'ensemble des candidats. Cette opération est itérée jusqu'à l'obtention d'un résumé de taille inférieure ou égale à un seuil fixé.

La première méthode est peu efficace s'il y a des transactions aberrantes. En effet, si un cluster a une transaction qui est différente des autres, cela dégrade sa description par un résumé individuel. Dans la pratique, le gain de compacité n'est pas optimisé avec ces méthodes car la taille du résumé est fixée au départ. Notons que la seconde méthode peut produire des résumés de taille plus petite que le seuil fixé.

2.3.2 Contrôle indirect de la taille des résumés

Les méthodes avec un contrôle indirect de la taille des résumés reposent sur la manipulation d'autres paramètres qui influent sur la taille des résumés. La méthode que nous décrivons dans cette section est basée sur le test du χ^2.

2.3.2.1 Direction Setting rules

Langages Dans [LHM99], les auteurs s'intéressent à la construction de résumés d'ensembles de règles d'association. Leur objectif est de trouver pour un ensemble de règles

donné, un sous-ensemble qui contient les relations essentielles dans les données.

Couverture Les auteurs n'utilisent pas de couverture pour construire les résumés. Leur approche repose plutôt sur le test du χ^2 qui permet de déterminer l'indépendance entre le corps et la tête d'une règle $X \Rightarrow Y$. S'il existe une dépendance entre le corps et la tête d'une règle alors cette règle est un candidat pour le résumé.

Mesure de qualité Les auteurs ne proposent pas de mesure de qualité pour les résumés qu'ils construisent. Ils utilisent le test du χ^2 pour évaluer individuellement les règles d'association. Les règles dont χ^2 est supérieur à un seuil donné sont des candidats pour former le résumé. Notons que le χ^2 n'est pas une mesure de qualité telle que nous l'avons définie au chapitre précédent car il n'évalue pas le résumé dans sa globalité.

Liu et al. définissent un résumé comme étant un ensemble de règles qu'ils nomment les direction setting (DS) rules, i.e. les règles qui donnent l'orientation. La notion de DS repose sur le test du χ^2. La valeur du χ^2 est calculée avec la formule suivante :

$$\chi^2 = |\mathcal{D}| \times \sum_{A \in \{X, \neg X\}, B \in \{Y, \neg Y\}} \frac{(sup(A \Rightarrow B, \mathcal{D}) - \widehat{sup}(A \Rightarrow B, \mathcal{D}))^2}{\widehat{sup}(A \Rightarrow B, \mathcal{D})} \tag{2.11}$$

où $\mathcal{D}$ est l'ensemble des transactions à partir desquelles les règles sont extraites, $sup(A \Rightarrow B, \mathcal{D})$ est le support de $A \Rightarrow B$ dans $\mathcal{D}$ (la valeur observée) et $\widehat{sup}(A \Rightarrow B, \mathcal{D})$ est le support que devrait avoir $A \Rightarrow B$ (la valeur attendue). La valeur attendue $\widehat{sup}(A \Rightarrow B, \mathcal{D})$ correspond à $sup(A, \mathcal{D}) \times sup(B, \mathcal{D})$. Si la valeur du χ^2 est plus grande que 0 alors la tête est statistiquement dépendante du corps. Étant donnés un seuil de support σ et un niveau de signification δ, la tête et le corps d'une règle $X \Rightarrow Y$ sont dits (σ, δ)-corrélés si $sup(X \Rightarrow Y, \mathcal{D}) > \sigma$ et la valeur renvoyée par le test du χ^2 est plus grande que δ. Ils sont positivement corrélés si en plus d'être corrélés, la valeur observée est plus grande que la valeur attendue, i.e. $\frac{sup(X \Rightarrow Y, \mathcal{D})}{\widehat{sup}(X \Rightarrow Y, \mathcal{D})} > 1$. Dans ce cas, on dit que la direction de $X \Rightarrow Y$ est 1. Par contre, si $\frac{sup(X \Rightarrow Y, \mathcal{D})}{\widehat{sup}(X \Rightarrow Y, \mathcal{D})} < 1$, on dit que la direction de $X \Rightarrow Y$ est -1. Si $sup(X \Rightarrow Y, \mathcal{D}) > \sigma$ et la valeur du χ^2 est inférieure à δ alors X et Y sont indépendants, on dit que la direction de $X \Rightarrow Y$ est 0.

Exemple 2.3.2 *Soit la règle* $\{tactile\} \Rightarrow \{monobloc\}$ *extraite à partir des transactions du tableau 1.1 (cf. page 32). La table de contingence 2.4 contient les données qui permettent d'effectuer le test du* χ^2 *pour cette règle. Chaque cellule montre le support de la règle dont*

	$\{monobloc\}$	$\{\neg monobloc\}$	Support ligne
$\{tactile\}$	0.45%	0	0.45%
$\{\neg tactile\}$	0.18%	0.37%	0.55%
Support colonne	0.63%	0.37%	100%

Table 2.4 – Table de contingence de la règle $\{tactile\} \Rightarrow \{monobloc\}$

la tête est en colonne et le corps est en ligne. Nous constatons que 0.45% *des portables sont tactiles et* 0.55% *ne sont pas tactiles. Si* $\{monobloc\}$ *est indépendante de* $\{tactile\}$ *alors*

$(0.45\times0.63)\%$ *des portables avec un design monobloc devraient être tactiles et* $(0.55\times0.63)\%$ *de ces portables devraient être non tactiles. De même pour les portables avec un design qui n'est pas monobloc. Ces valeurs sont celles qui sont attendues. Les valeurs observées sont celles qui sont affichées dans le tableau. Ainsi, nous avons :*

$$\chi^2 = |\mathcal{D}| \times \left[\frac{(0.45-0.29)^2}{0.29} + \frac{(0.55-0.35)^2}{0.35} + \frac{(0.45-0.16)^2}{0.16} + \frac{(0.55-0.2)^2}{0.2}\right] = 14,69$$

Si 14.69 *est supérieure au seuil fixé, on peut dire que* $\{monobloc\}$ *est corrélée à* $\{tactile\}$ *et que la direction de* $\{tactile\} \Rightarrow \{monobloc\}$ *est* 1 *car* $\frac{sup(\{tactile\}\Rightarrow\{monobloc\},\mathcal{D})}{\widehat{sup}(\{tactile\}\Rightarrow\{monobloc\},\mathcal{D})} = \frac{0.45}{0.29}$ *est plus supérieure à* 1.

Ainsi, une règle DS est formellement définie comme suit :

Définition 2.3.1 (Règle DS) *Étant donnée une règle d'association* $r : X \Rightarrow Y$, *r est une règle DS si elle vérifie les conditions suivantes :*

1. *r est de direction* 1.
2. *Pour tout couple de règles* $(X_1 \Rightarrow Y, X_2 \Rightarrow Y)$ *telle que* $X = X_1 \cup X_2$, $X_1 \Rightarrow Y$ *ou* $X_2 \Rightarrow Y$ *est de direction* -1.

Plus précisément, la condition 2 signifie que la règle r ne doit pas être le résultat d'une combinaison de deux règles telles que l'une est de direction 1 et l'autre est de direction 0 ou telle qu'elles sont toutes de direction 1. En effet, si tel est le cas, r est prévisible lorsque ces règles sont connues.

Contrôle de la taille des résumés Les auteurs proposent un algorithme exhaustif qui fonctionne par niveau. Cet algorithme prend en entrée un ensemble de règles d'association et un seuil pour les tests de χ^2 et il fournit un résumé de cet ensemble qui est constitué des règles DS suivant le seuil donné. Ce seuil influe sur la taille des résumés, en effet, plus il est petit, plus la taille du résumé augmente.

En définitive, les méthodes qui produisent des résumés non génératifs reposent sur le clustering par partitionnement pour permettre un contrôle direct de la taille des résumés. Le contrôle indirect est basé sur la manipulation d'autres paramètres tels que le seuil du χ^2.

2.4 Synthèse sur les méthodes de construction de résumés

Nous proposons dans cette section une synthèse sur les méthodes de construction de résumés exposées dans ce chapitre. Le tableau 2.5 présente un récapitulatif sur ces méthodes. Nous y reprenons les caractéristiques que nous avons énoncées dans la section 2.1.

– **Langage des motifs à résumer et langage des motifs des résumés** Les méthodes de construction de résumés sont presque toutes dépendantes des motifs à résumer. Elles s'appliquent aux transactions [CK05], aux itemsets fréquents [AGM04, YCHX05, PG09] avec ou sans leur support ou aux règles d'association [LHM99].

Seule la méthode proposée dans [MM03] et les méthodes de clustering par partitionnement [Mac67, KR05] sont indépendantes de la nature des motifs. Notons que la méthode proposée dans [LHM99] s'applique seulement aux règles de classification. Par ailleurs, les motifs des résumés sont de même nature que les motifs de l'ensemble de départ. Certaines méthodes produisent des résumés avec des motifs légèrement différents. Cette différence ne concerne pas la structure des motifs, mais elle est relative à l'information associée à ces motifs. En effet, dans [YCHX05, CK05], les résumés sont composés de profils qui sont des itemsets accompagnés d'informations supplémentaires permettant d'approximer le support des itemsets qu'ils couvrent.

– **Couverture de l'ensemble de motifs à résumer** Pratiquement toutes les méthodes décrites utilisent une relation de couverture. Cependant, pour les méthodes de clustering, la relation de couverture est différente de celle que nous avons définie dans le chapitre précédant. Elle est plutôt basée sur la distance entre les motifs et leurs représentants. Pour les résumés génératifs, la relation de couverture permet d'identifier les motifs du résumé à partir desquels un motif ou son support est régénéré. La relation de couverture utilisée Les résumés proposés dans [YCHX05, PG09, CK05, Mac67, KR05] couvrent la totalité des motifs de l'ensemble qui est résumé. Par contre, la couverture est partielle pour les résumés proposés dans [AGM04, MM03].
– **Régénération des motifs** Les résumés génératifs sont des représentations condensées approximatives. Leur principe est de faire un compromis entre la taille des résumés et la qualité de la régénération des motifs. Plus précisément, il y a deux approches qui sont utilisées.
  La première approche consiste à construire les résumés avec un contrôle direct de leur taille. Dans ce cas de figure, la qualité de la régénération dépend de la taille : plus la taille est petite, moins la régénération est précise.
  Dans le deuxième approche, l'erreur de régénération est contrôlée et la taille varie de manière anti-monotone en fonction de cette erreur.
  Quelle que soit l'approche utilisée, un paramètre est privilégié par rapport à un autre. Ainsi, soit le résumé est synthétique et les motifs sont mal régénérés, soit le résumé est grand et la régénération est meilleure.
  Une couverture incomplète des ensembles de motifs a pour effet une mauvaise régénération de certains motifs. Plus précisément, dans [MM03], le support des motifs non couverts est estimé à 0. Dans [AGM04], les motifs de l'ensemble de départ ne peuvent pas tous être régénérés.
  D'autre part, la couverture de faux positifs (des motifs qui ne sont pas dans l'ensemble de départ) entraîne une perte de précision lors de la régénération [AGM04].
  Enfin, la redondance sur la couverture de motifs présente un problème lorsqu'il faut régénérer les motifs. En particulier, dans [YCHX05] , un profil du résumé peut couvrir plus d'un itemset. Dans ce cas de figure, les auteurs proposent d'estimer son support par la moyenne des supports calculée à partir de ces profils, ce qui entraine une mauvaise régénération.
– **Mesure de qualité des résumés** Presque toutes les méthodes utilisent comme heuristique une mesure de qualité des résumés. Ces mesures dépendent des objectifs visés. En effet, pour les méthodes qui produisent des résumés génératifs, la mesure

de qualité utilisée porte sur l'erreur d'approximation du support des motifs [MM03, YCHX05, PG09], ou sur le nombre de motifs couverts [AGM04]. Par contre, pour les autres méthodes, la mesure de qualité concerne la perte d'information [CK05]. La méthode décrite dans [LHM99] utilise le χ^2 qui s'applique à un motif du résumé et non au résumé dans sa globalité. Cette mesure n'évalue donc pas la qualité des résumés. La taille des résumés [MM03, LHM99] ou le gain de compacité [CK05] est évaluée dans les approches permettant un contrôle indirect de la taille des résumés.

		Résumés génératifs				Résumés non génératifs	
		Contrôle direct			Contrôle indirect	Contrôle direct	Contrôle indirect
Référence		[AGM04]	[YCHX05]	[MM03]	[PG09]	[CK05]	[LHM99]
Page		47 Approximation d'une collection d'itemsets	48 Les profils	50 The pattern ordering problem	52 Les c-profils	55 Résumés individuels	56 Direction Setting rules
Motifs à résumer		Itemsets	Itemsets	Motifs fréquents	Itemsets	Transactions	Règles d'association
Motifs des résumés		Itemsets	profils	Motifs fréquents	c-profils	Itemsets	Règles d'association
Couverture		Partielle	Totale	Partielle	Totale	Totale	Pas de relation de couverture
Régénération		Motifs	Motifs et supports des motifs	Motifs et mesure d'intérêt des motifs	Motifs et supports	Non	Non
Mesure de qualité	Heuristique	Nombre de motifs couverts	Erreur d'approximation des supports	Erreur d'approximation des supports	Erreur d'approximation des supports	Perte d'information	-
	Evaluation après construction	-	-	Taille du résumé	-	Gain de compacité	Taille du résumé

Table 2.5 – Récapitulatif sur les méthodes de construction de résumés

2.5 Conclusion

Nous avons étudié tout au long de ce chapitre des travaux sur les méthodes de construction de résumés d'ensembles de motifs. Ces méthodes sont classées en deux grandes catégories : les méthodes qui produisent des résumés génératifs et ceux qui produisent des résumés non génératifs. Les résumés génératifs sont des représentations condensées approximatives. Les méthodes qui produisent ces résumés privilégient soit la taille des résumés soit la qualité de régénération des motifs. Ainsi, ces résumés sont soit très synthétiques et l'erreur de régénération est plus importante, soit trop grands et ils sont dans ce cas difficiles à explorer. Les méthodes qui produisent des résumés non génératifs s'affranchissent de la régénération des motifs pour mieux représenter les ensembles de motifs.

Par ailleurs, nous constatons qu'aucune approche ne propose une méthode d'exploration des ensembles de motifs via leurs résumés. Nous nous intéressons donc dans le prochain chapitre aux méthodes de visualisation d'ensembles de motifs.

Chapitre 3

Représentation visuelle d'ensembles de motifs

La visualisation de motifs est un sujet qui a fait l'objet de nombreux travaux en Data Mining [KK96a, WB97a, BKK97, Kei02, LZBX06]. Dans ce chapitre, nous nous intéressons aux méthodes qui s'appliquent aux règles d'association et aux itemsets fréquents car notre contribution porte sur ces types de motifs. Ces méthodes s'appuient sur divers supports tels que les tableaux, les graphes, les caractéristiques géométriques, les matrices et les cubes. Nous décrivons dans la section 3.1 des critères pour étudier ces méthodes. Ensuite, nous présentons dans les sections 3.2 à 3.6 quelques méthodes pour chaque type de support. Une synthèse sur ces méthodes de visualisation est présentée dans la section 3.7. Les tableaux 3.1 et 3.2 montrent respectivement un ensemble d'itemsets fréquents fermés et un ensemble de règles d'association que nous utilisons pour illustrer les méthodes présentées. Les itemsets fréquents fermés sont générés à partir des données du tableau 1.1 (cf. page 32) avec un support minimal ($minsup$) de 0.45. Les règles d'associations sont obtenues à partir de ces itemsets fréquents fermés avec une confiance minimale ($minconf$) fixée à 0.8.

	Itemset	Sup
X_1	{2Mp-5Mp}	$\frac{9}{11}$
X_2	{3h-5h}	$\frac{7}{11}$
X_3	{monobloc}	$\frac{7}{11}$
X_4	{3h-5h,ub}	$\frac{6}{11}$
X_5	{2Mp-5Mp,3h-5h}	$\frac{6}{11}$
X_6	{2Mp-5Mp,non tactile}	$\frac{6}{11}$
X_7	{monobloc,tactile}	$\frac{5}{11}$
X_8	{2Mp-5Mp,monobloc}	$\frac{5}{11}$
X_9	{2Mp-5Mp,3h-5h,ub}	$\frac{5}{11}$
X_{10}	{2Mp-5Mp,3h-5h,non tactile}	$\frac{5}{11}$

TABLE 3.1 – Ensemble d'itemsets fermés

	Corps	Tête	Sup	Conf
r_1	{ub}	{3h-5h}	$\frac{6}{11}$	1
r_2	{3h-5h}	{ub}	$\frac{6}{11}$	$\frac{6}{7}$
r_3	{tactile}	{monobloc}	$\frac{5}{11}$	1
r_4	{non tactile}	{2Mp-5Mp}	$\frac{6}{11}$	1
r_5	{3h-5h}	{2Mp-5Mp}	$\frac{6}{11}$	$\frac{6}{7}$
r_6	{ub}	{2Mp-5Mp}	$\frac{5}{11}$	$\frac{5}{6}$
r_7	{3h-5h,2Mp-5Mp}	{ub}	$\frac{5}{11}$	$\frac{5}{6}$
r_8	{non tactile}	{3h-5h}	$\frac{5}{11}$	$\frac{5}{6}$
r_9	{3h-5h,2Mp-5Mp}	{non tactile}	$\frac{5}{11}$	$\frac{5}{6}$

TABLE 3.2 – Ensemble de règles d'association

3.1 Les critères d'étude

Pour étudier les techniques de visualisation, nous observons des critères que nous avons définis en considérant notre problématique initiale, à savoir que l'objectif est d'arriver à explorer efficacement de grands ensembles de motifs qui peuvent être des règles d'association ou des itemsets fréquents. Il existe des travaux qui se sont intéressés à l'étude de techniques de visualisation. Cependant, dans la plupart des études proposées, aucun critère n'est défini [WB97b, Kei02]. Les critères décrits dans les travaux qui en proposent ne sont pas adaptés à notre problématique car ils sont soit trop orientés sur l'aspect visuel et interactif [BCL01], soit spécifiques aux techniques de visualisation de données multidimentionnelles [KK96b]. Nous détaillons ci-dessous les critères que nous proposons.

- **Les motifs** Comme nous l'avons précisé dans l'introduction, nous nous intéressons aux méthodes qui s'appliquent aux itemsets fréquents ou aux règles d'association. Nous précisons pour chaque méthode étudiée, le type des motifs qui sont visualisés.
- **La richesse de la représentation** Nous identifions pour chaque méthode, les informations qui sont représentées dans la visualisation. Plus précisément, nous observons la capacité des visualisations à montrer des liens entre les motifs, le nombre de mesures d'intérêt des motifs qu'elles peuvent présenter et leur capacité à donner une vue globale des ensembles de motifs.
- **La lisibilité** Nous examinons si les représentations peuvent supporter un nombre important de motifs. En d'autres termes, il s'agit d'observer leur comportement lorsqu'on passe à l'échelle.
- **La navigation** Nous distinguons deux types de navigations : les navigations classiques de filtrage et de sélection et les navigations qui permettent d'agréger ou de détailler les visualisations.
- **L'interprétabilité** Dans notre contexte, une visualisation est interprétable si elle est non redondante et consistante. Elle est non redondante si les motifs ne sont pas représentés plusieurs fois dans la visualisation. Elle est consistante si elle n'introduit pas de faux positifs, i.e. des motifs qui n'existent pas dans l'ensemble qui est représenté.
- **Contrôle de la taille des visualisations** Il s'agit d'identifier si les méthodes proposées permettent de contrôler la taille des visualisations ou pas.

Nous décrivons des techniques de visualisation de la section 3.2 à la section 3.6 en nous basant sur le support utilisé pour représenter les motifs.

3.2 Les tableaux

Dans [CZ03], les auteurs visualisent des ensembles de règles d'association sous forme de tableaux semblables au tableau 3.2. Chaque ligne correspond à une règle et chaque colonne représente une caractéristique des règles. Ils proposent une interface qui permet de filtrer les règles en spécifiant les items que l'utilisateur désire avoir dans la règle (la tête ou le corps) ou en fixant un seuil de support ou de confiance. Elle permet aussi de trier les règles par ordre croissant ou décroissant de support ou de confiance. La taille des visualisations peut être contrôlée directement en fixant le nombre de règles à visualiser simultanément.

Notons que d'autres types de motifs tels que les itemsets peuvent être visualisés avec cette méthode. De plus, plusieurs mesures d'intérêt peuvent être affichées en même temps. Cependant, cette approche ne permet pas d'avoir une vue globale des ensembles de motifs. D'autre part, elle n'est pas adaptée pour visualiser de grands ensembles de motifs car la longueur des tableaux augmente en fonction du nombre de règles visualisées. Enfin, elle ne permet pas de mettre en évidence les relations entre les motifs.

La section suivante présente des méthodes qui visualisent des ensembles de motifs sous forme de graphe. Cette structure permet de mettre en évidence des liens entre les motifs.

3.3 Les graphes

Les graphes sont habituellement utilisés pour représenter des relations entre les motifs. Ils sont donc bien adaptés pour visualiser les règles d'association en reliant leur tête et leur corps par un arc. Les arcs peuvent être orientés (habituellement du corps vers la tête) [KMR+94] ou non orientés [BB04]. Dans ce dernier cas, d'autres indicateurs sont utilisés pour différencier la tête du corps des règles.

3.3.1 Hypergraphe orienté cyclique

Dans [KMR+94], Klemettinen et al. utilisent des hypergraphes pour visualiser des ensembles de règles de classification. Dans leur approche, chaque nœud correspond à un item. Une règle d'association est représentée par un arc entre un groupe d'items qui constituent le corps et un item qui est la tête de la règle. L'épaisseur de l'arc représente le support de la règle. La couleur de l'arc est utilisée pour visualiser d'autres mesures d'intérêt telles que la confiance. Une règle est détaillée à la demande de l'utilisateur de façon textuelle avec ses mesures d'intérêt qui sont visualisées sous forme d'un diagramme en barre. La

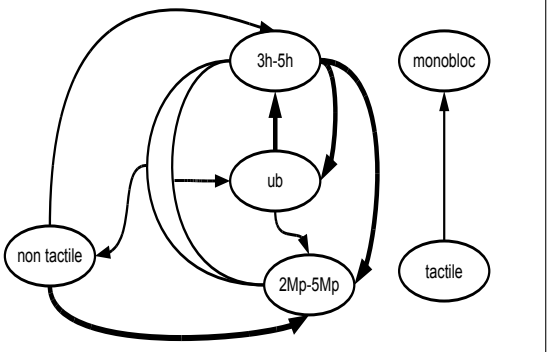


FIGURE 3.1 – Hypergraphe représentant un ensemble de règles

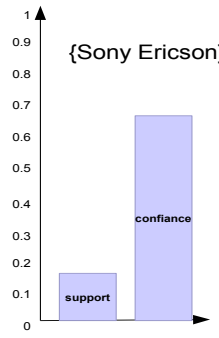


FIGURE 3.2 – Détail d'une règle

figure 3.1 montre un exemple de représentation des règles du tableau 3.2 ayant un item dans la tête qui sont décrites dans le tableau 3.2. La figure 3.2 montre le détail de la règle

r_5. Les auteurs proposent une interface permettant d'effectuer de la sélection et du filtrage. La taille de la visualisation peut être contrôlée de manière indirecte en fixant le nombre de règles à visualiser.

3.3.2 Graphe non orienté

Des graphes non orientés sont utilisés dans [BB04] pour représenter des ensembles de règles d'association. Les nœuds correspondent à des itemsets et chaque arc relie la tête et le corps d'une règle. Notons que la tête est un itemset qui est inclus dans le coprs. Deux couleurs sont utilisées pour distinguer le corps et la tête des règles. Si un itemset est en même temps le corps d'une règle et la tête d'une autre règle, alors son nœud porte les deux couleurs. Trois autres couleurs sont utilisées pour exprimer le support des règles sur les arcs : une couleur claire pour les supports faibles, une couleur foncée pour les supports élevés et une couleur noire pour les supports de 1. La confiance d'une règle est représentée par la longueur de l'arc qui lie son corps et sa tête. Plus la confiance est élevée, plus l'arc est long. La figure 3.3 montre une représentation des règles du tableau 3.2 avec cette méthode.

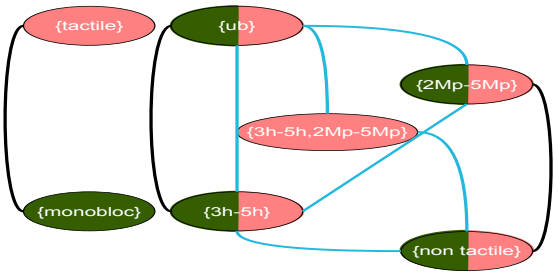


FIGURE 3.3 – Graphe non orienté

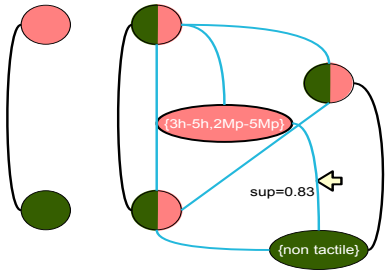


FIGURE 3.4 – Réduction des nœuds et détail d'une règle

S'il y a beaucoup de règles, les nœuds sont transformés en petits cercles comme le montre la figure 3.4. Dans ce contexte, une règle est détaillée si on pose la souris sur l'arc qui relie sa tête et son corps. Le graphe peut aussi être déplacé (tourné) afin de mettre en évidence une partie du graphe. Ces deux fonctionnalités lui donnent son aspect interactif. La taille de la visualisation ne peut pas être contrôlée. Pour une exploration plus efficace des règles, Bruzzese et Buono proposent de sélectionner un sous-ensemble des règles qui partagent la même tête et de les représenter avec les coordonnées parallèles. Nous détaillons ce mode de représentation dans la section 3.4.3.

Globalement les visualisations basées sur les graphes montrent vite leurs limites quand le nombre de règles est important : les arcs s'entrecroisent très souvent et l'abondance des nœuds et des arcs entraîne une occlusion. La figure 3.5 illustre ces défauts. D'autre part, les visualisations ne donnent pas une vue globale des ensembles de motifs car toutes les faces du graphe n'apparaissent pas simultanément.

Avec les méthodes basées sur les graphes, les relations entre les motifs ainsi que leurs mesures d'intérêt sont représentées avec des arcs. D'autres méthodes utilisent des formes ou

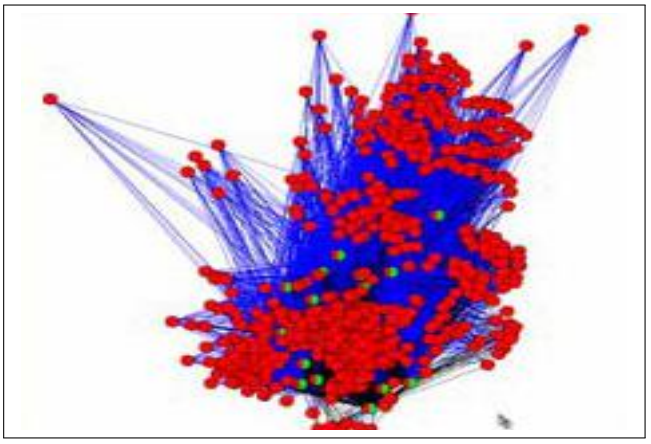


Figure 3.5 – Entrecroisements d'arcs et occlusion dans un graphe

des caractéristiques géométriques pour représenter aussi bien les motifs que leurs mesures d'intérêt.

3.4 Les outils géométriques

Cette section décrit des méthodes particulières de représentation de motifs basées sur des outils géométriques. Ces méthodes utilisent principalement la distance entre objets, la surface ou encore les coordonnées d'objets pour exprimer des mesures d'intérêt de motifs ou des relations entre les motifs.

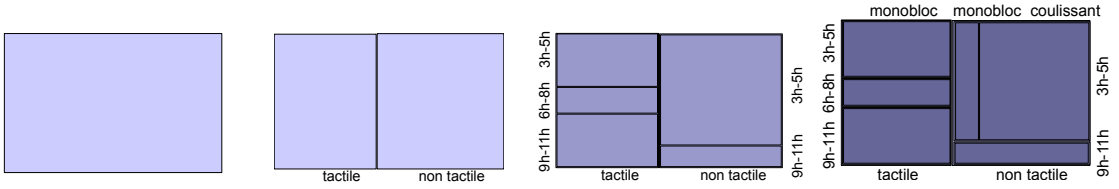
3.4.1 Les diagrammes en mosaïque

Les diagrammes en mosaïque ont été introduits dans [HK81] pour visualiser des tables de contingence. La table de contingence est un moyen permettant de représenter simultanément deux caractéristiques observées sur un ensemble de données. La figure 3.6 montre une table de contingence construit pour les attributs (caractéristiques) *Design*, *Écran* et *Autonomie* à partir des données du tableau 1.1. Il est composé des différentes combinaisons des valeurs de ces attributs. Dans chaque cellule est affiché le nombre de données qui contiennent les items en ligne et les items en colonne. Un diagramme en mosaïque est obtenu à partir d'une table de contingence en représentant chaque cellule du tableau par un rectangle. La construction du diagramme s'effectue de manière itérative en découpant à chaque étape les rectangles obtenus à l'étape précédente suivant un attribut. La figure 3.7 montre les différentes étapes de découpage pour obtenir le diagramme en mosaïque correspondant à la table de contingence de la figure 3.6. Au début, on dispose d'un grand rectangle qui représente toutes les données (cf. figure 3.6(a)). Ensuite, ce rectangle est découpé suivant l'attribut *Design*, ce qui donne le diagramme de la figure 3.6(b). Il est composé de deux rectangles correspondant aux 5 données contant la valeur *tactile* (à gauche) et aux 6 données contenant la valeur *non tactile* (à droite). A l'étape suivante, chacun de ces rectangles est divisé suivant l'attribut *Écran* (figure 3.6(c)). Les divisions se

Écran \ *Design* / *Autonomie*		monobloc	coulissant
tactile	3h-5h	2	0
	6h-8h	1	0
	9h-11h	2	0
non tactile	3h-5h	1	4
	6h-8h	0	0
	9h-11h	1	0

FIGURE 3.6 – Table de contingence

font alternativement en longueur puis en largeur d'une étape à l'autre. Enfin les rectangles obtenus sont divisés suivant l'attribut *Autonomie* pour donner le diagramme final de la figure 3.6(b).



(a) Rectangle initial (b) Division suivant *Design* (c) Division suivant *Écran* (d) Division suivant *Autonomie*

FIGURE 3.7 – Étapes de construction d'un diagramme en mosaïque

Les diagrammes en mosaïque[1] sont utilisés dans [HSW00] pour visualiser des règles de classification. Pour représenter une règle $\{(A_1, a_1), \cdots, (A_K, a_K)\} \Rightarrow \{(B, b)\}$, les auteurs construisent dans un premier temps un diagramme en mosaïque suivant les attributs $A_1, \cdots, A_K$ dont les valeurs apparaissent dans le corps de la règle. Puis, ils surlignent dans chaque rectangle la proportion des données qui contiennent la valeur b de la tête de la règle. La figure 3.8 montre le diagramme en mosaïque construit pour la règle $\{(\acute{E}cran, tactile), (Autonomie, 3h - 5h)\} \Rightarrow \{(Design, monobloc)\}$. Chaque portion de rectangle surlignée en rouge (en foncé) correspond à une règle. Parmi ces portions, on retrouve celle qui représente la règle $\{(\acute{E}cran, tactile), (Autonomie, 3h - 5h)\} \Rightarrow \{(Design, monobloc)\}$. L'espace qu'elle occupe équivaut à $\frac{2}{11}$ de la surface totale, ce qui correspond au support de la règle. D'autre part, elle occupe tout le rectangle correspondant aux valeurs *tactile* et $3h - 5h$, ce qui signifie qu'elle a une confiance égale à 1. La règle est visualisée dans un contexte global. En d'autres termes, elle est présentée avec toutes les autres règles qui partagent les mêmes attributs dans le corps et la même valeur dans la tête. Les diagrammes peuvent être construits de manière interactive. D'autre part, la taille des visualisations (le nombre de rectangles) peut être contrôlée indirectement en fixant le nombre d'attributs pour le corps des règles.

Cette méthode de visualisation est limitée aux règles de classification. De plus, elle

1. Mosaic plots

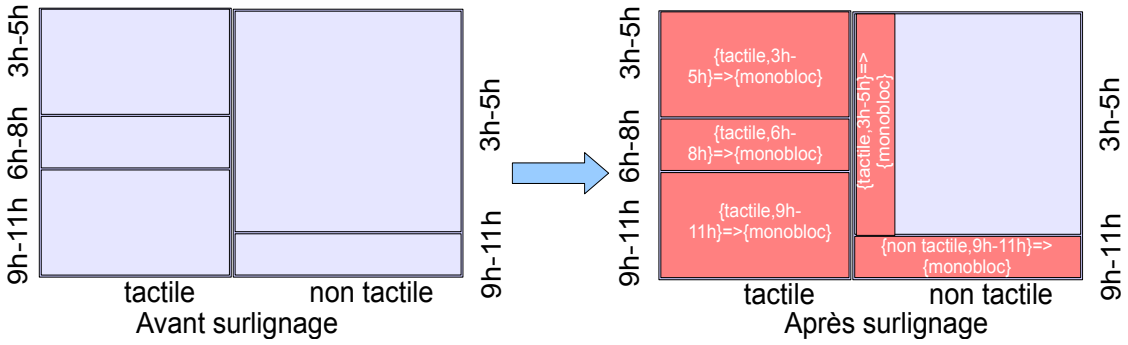


FIGURE 3.8 – Représentation de règles d'associations avec un diagramme en mosaïque

ne permet de représenter simultanément qu'une partie des ensembles de règles, i.e. les règles ayant la même tête et partageant les mêmes attributs dans le corps. Enfin, la seule navigation possible est l'interaction lors de la construction des diagrammes.

3.4.2 Les polylignes : FpViz

Dans [LIC08], les auteurs proposent une méthode basée sur les polylignes (succession de lignes) pour visualiser des itemsets fréquents. Le principe est de schématiser un itemset par une succession de lignes qui relient des items représentés par des cercles. Les polylignes sont placées dans un espace à deux dimensions qui représentent les items et leur support. La figure 3.9 montre une visualisation des itemsets du tableau 3.1. Les items sont placés sur l'axe des abscisses dans l'ordre décroisant de leur support. Le support d'un itemset correspond à celui du dernier nœud de sa polyligne (en allant de la gauche vers la droite). Cette représentation n'est cependant pas très pratique lorsqu'il y a beaucoup d'itemsets

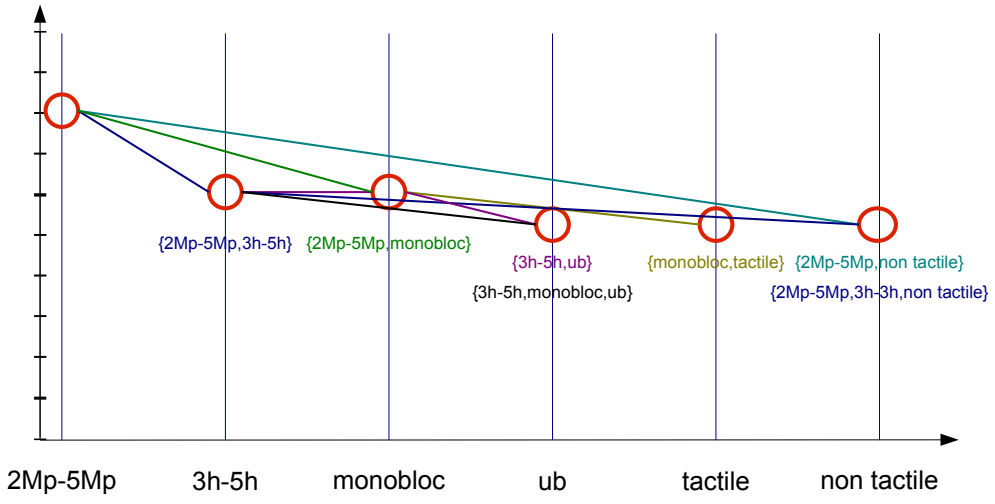


FIGURE 3.9 – Représentation d'itemsets avec des polylignes

car les polylignes fléchissent au niveau des cercles et ont souvent tendance à se croiser, ce qui rend difficile la différenciation des itemsets. Leung et Carmichael proposent une variante qui permet de supprimer les croisements [LC09] (cf. figure 3.10). Elle consiste à représenter chaque itemset sur une ligne horizontale. Les lignes sont affichées par ordre de support. Le dernier item (selon l'ordre des items) de chaque itemset est représenté par un disque. Cependant, lorsqu'il y a beaucoup d'itemsets, le nombre de lignes devient très important. Les auteurs proposent de compresser la représentation en fusionnant les lignes des itemsets ayant le même support pour donner une seul ligne par valeur de support. La ligne d'un ensemble d'itemsets de même support est obtenue en projetant toutes les lignes des itemsets sur une même ligne horizontale. Si un item est le dernier d'un itemset, toujours selon l'ordre des items, alors il est représenté par un disque dans la ligne résultante de la projection. La figure 3.11 présente une compression des lignes de la figure 3.10. Par exemple,

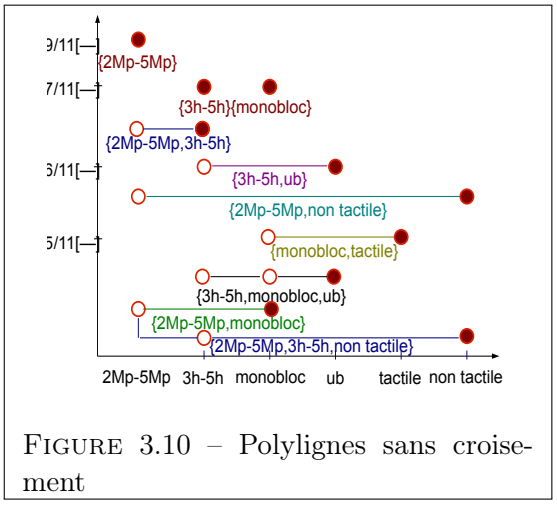


FIGURE 3.10 – Polylignes sans croisement

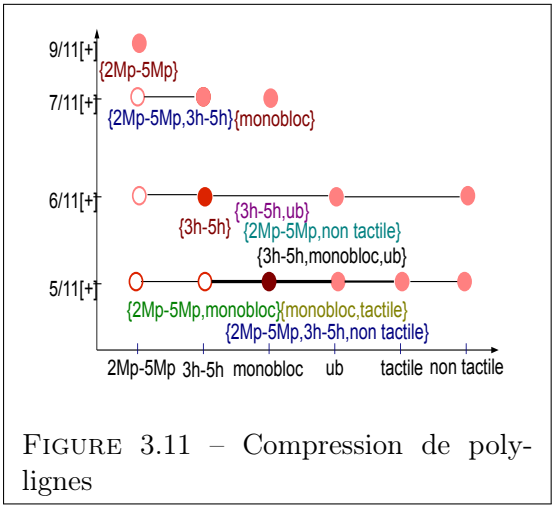


FIGURE 3.11 – Compression de polylignes

la ligne correspondant au support $\frac{5}{11}$ concentre les itemsets $\{2Mp-5Mp, monobloc\}$, $\{monobloc, tactile\}$ et $\{2Mp-5Mp, 3h-5h, non\ tactile\}$. La couleur des cercles indique la fréquence d'apparition des items qui leur sont associés dans les itemsets visualisés. Le cercle d'un item est de couleur claire si cet item apparaît dans un seul itemset. Cette couleur s'assombrit graduellement en fonction de l'occurrence de celui-ci. L'épaisseur d'une ligne indique le nombre de lignes horizontales qui sont projetées. Une ligne peut être détaillée en cliquant sur le signe + qui se trouve à côté de son support, ce qui permet de voir toutes les lignes qu'elle représente comme elles sont affichées dans la figure 3.10. Par ailleurs, cette méthode de visualisation permet de filtrer les motifs en faisant varier le supportS minimal et maximal ou les cardinalités minimale et maximale des itemsets. Des itemsets peuvent être sélectionnés en spécifiant des items qu'ils contiennent. Il est aussi possible de détailler les itemsets. Par exemple, en sélectionnant un segment entre deux nœuds, les itemsets contenant les items de ces nœuds sont affichés.

La fusion des lignes introduit une confusion car les lignes obtenues peuvent contenir des segments entre deux items alors que ces items n'apparaissent pas ensemble dans un itemset. Par exemple, il y a une ligne qui relie les items *ub* et *tactile* alors qu'il n'existe pas dans l'ensemble un itemset contenant $\{ub, tactile\}$. Les visualisations donnent une vue globale par rapport au support des motifs. Par contre elles deviennent très grandes, s'il y

a beaucoup de supports différents ou d'items qui apparaissent dans les itemsets.

3.4.3 Les coordonnées parallèles

Les coordonnées parallèles ont été introduites en 1981 par Inselberg [Ins81, Ins00] comme une méthode de représentation de données multidimensionnelles [2]. Elles sont adaptées dans [BB04] pour visualiser des règles d'association. La figure 3.12 montre des coordonnées parallèles réalisées pour les règles du tableau 3.3 qui partagent la même tête. Les axes représentent les différents items et les lignes brisées correspondent aux règles. Une ligne rencontre un axe en fonction de la valeur d'Utilité de l'Item (IU) [BD03] correspondant à l'axe, par rapport à la règle représentée par cette ligne. L'UI permet d'évaluer l'influence d'un item sur la conclusion d'une règle donnée. Étant donné un item a et une règle $X \Rightarrow Y$, l'utilité de a par rapport à $X \Rightarrow Y$ est définie par la formule suivante :

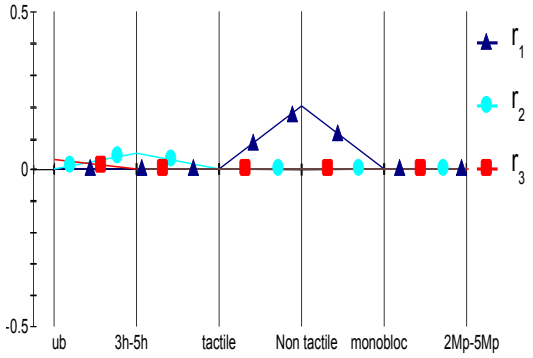


Figure 3.12 – Représentation de règles d'association avec des coordonnées parallèles

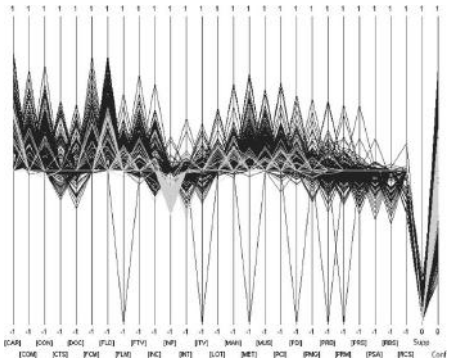


Figure 3.13 – Occlusion des coordonnées parallèles

	r_4	r_5	r_6
Corps (X)	$\{non\ tactile\}$	$\{3h-5h\}$	$\{ub\}$
Tête (Y)	$\{2Mp-5Mp\}$	$\{2Mp-5Mp\}$	$\{2Mp-5Mp\}$
$IU_r(ub)$	0	0	$\frac{1}{54}$
$IU_r(3h-5h)$	0	$\frac{1}{21}$	0
$IU_r(tactile)$	0	0	0
$IU_r(non\ tactile)$	$\frac{2}{11}$	0	0
$IU_r(\{monobloc\})$	0	0	0
$IU_r(\{2Mp-5Mp\})$	0	0	0

Table 3.3 – IU d'items par rapport à des règles d'associations qui partagent la même tête

$$IU_{X\Rightarrow Y}(a) = \frac{conf(X \Rightarrow Y, \mathcal{D}) - conf(X\backslash\{a\} \Rightarrow Y, \mathcal{D})}{\max(conf(X \Rightarrow Y, \mathcal{D}); conf(X\backslash\{a\} \Rightarrow Y, \mathcal{D}))} \quad (3.1)$$

où $conf(X \Rightarrow Y, \mathcal{D})$ et $conf(X\backslash\{a\} \Rightarrow Y, \mathcal{D})$ sont respectivement la confiance de la règle $X \Rightarrow Y$ et celle de la règle obtenue en supprimant l'item a de X. La valeur $IU_{X\Rightarrow Y}(a)$

2. Données décrites par plusieurs attributs

appartient à l'intervalle $]-1,1]$. Elle est positive si l'absence de a de X entraîne une diminution de la confiance et négative dans le cas contraire. Les règles peuvent être filtrées en fixant un seuil de confiance ou un seuil d'utilité d'item par rapport à chaque règle pour tous les items affichés sur les axes.

Les coordonnées parallèles sont aussi utilisées dans [Yan05] pour représenter des itemsets et des règles d'association. La figure 3.14 montre les règles d'association du tableau 3.2

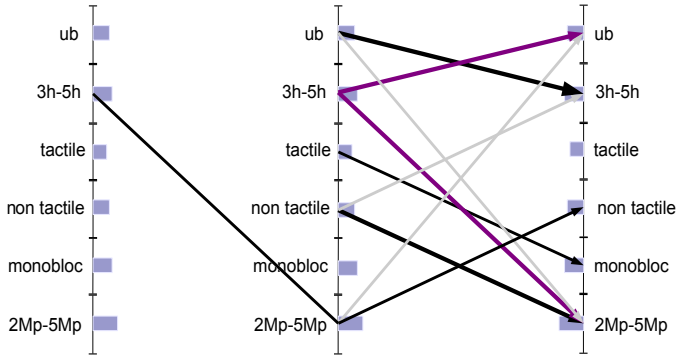


FIGURE 3.14 – Représentation de règles d'association avec des coordonnées parallèles

visualisées avec la méthode proposée par Yan. Un itemset correspond à une ligne brisée qui relie des items se trouvant sur les axes verticaux. Une règle est représentée par une ligne brisée qui passe par les items contenus dans sont corps suivi d'une flèche qui la relie à une seconde ligne brisée qui passe par les items contenus dans sa tête. Dans notre exemple, il n'y a pas les secondes lignes car les règles ont toutes un seul item dans la tête. Les items placés à gauche des axes apparaissent dans le corps des règles alors que ceux qui sont affichés à droite apparaissent dans la tête des règles. Le support des règles est exprimé par l'épaisseur des courbes et la confiance est symbolisée par la couleur graduelle des courbes. Le support de chaque item est exprimé par la longueur de la petite barre qui est placée en face.

Notons que si deux ou plusieurs règles ont des items en commun, on peut confondre les lignes qui les représentent. Par exemple, il y 2 règles qui contiennent les items $3h-5h$ et $2Mp-5Mp$ dans leur corps. Les deux lignes reliant ces items sont donc superposées, ce qui a pour effet la visualisation les règles $\{3h-5h, 2Mp-5Mp\} \Rightarrow \{non\ ; tactile\}$ et $\{2Mp-5Mp\} \Rightarrow \{ub\}$ au lieu des règles $\{3h-5h, 2Mp-5Mp\} \Rightarrow \{non\ ; tactile\}$ et $\{3h-5h, 2Mp-5Mp\} \Rightarrow \{ub\}$. Pour mieux les distinguer, les auteurs utilisent des courbes de Bézier à la place des lignes brisées qui rencontrent les axes avec des angles différents.

Ce mode de représentation nécessite que tous les items soient placés sur chaque axe, ce qui peut rendre la visualisation inexploitable s'il y a beaucoup d'items. Les auteurs proposent de remplacer ces items par les nœuds racines de taxonomies d'items [3]. Ainsi, un nœud racine qui est présent dans une taxonomie peut être détaillé en affichant ses nœuds fils. Ces nœuds peuvent à leur tour être détaillés jusqu'aux derniers de leurs descendants. Cependant, ce n'est pas dans tous les domaines que l'on peut organiser les items en taxonomie. D'autre part, le nombre d'axes est dépendant du nombre d'items dans le corps des règles.

3. classification d'items sous forme d'arborescence

Par exemple, on visualise un ensemble contenant des règleS avec deux items dans le corps sauf une qui en a six, alors on aurait six axes pour le corps des règles au lieu de deux si cette règle n'existait pas.

Les coordonnées parallèles sont limitées lorsqu'il y a beaucoup de motifs à représenter. En effet, il devient difficile de distinguer les motifs à cause de l'occlusion (cf. figure 3.13). Notons également qu'elles ne donnent pas une vue globale des ensembles de motifs.

3.4.4 Métaphore visuelle

Blanchard et ses collègues proposent une métaphore visuelle pour représenter des ensembles de règles d'association [BGB03]. Les règles sont visualisées par classe. Chaque classe est identifiée par un itemset et contient deux types de règles : les règles spécifiques et les règles générales. Les règles spécifiques sont celles dont le corps est l'identifiant de la classe. Les règles générales sont celles dont l'union de la tête et du corps correspond à l'identifiant de la classe. Seules les règles avec un seul item dans la tête sont considérées. Par exemple, si nous considérons la classe identifiée par $\{3h-5h, 2Mp-5Mp\}$ et constituée des règles $r_5 : \{3h-5h\} \Rightarrow \{2Mp-5Mp\}$, $r_7 : \{3h-5h, 2Mp-5Mp\} \Rightarrow \{ub\}$ et $r_9 : \{3h-5h, 2Mp-5Mp\} \Rightarrow \{non\ tactile\}$ alors r_5 est une règle spécifique et r_7 et r_9 sont des règles générales.

Les auteurs définissent une relation de spécialisation qui permet de naviguer entre les classes. Etant données deux classes C_1 et C_2 identifiées respectivement par $ID1$ et $ID2$, C_1 est plus spécifique que C_2 si $ID2 \subseteq ID1$. Le tableau 3.4 affiche les classes obtenues à partir des règles du tableau 3.2. Les relations de spécialisation entre les classes sont décrites par une hiérarchie visualisée dans la figure 3.15. Notons qu'une règle peut appartenir à plusieurs classes de niveau différent dans la hiérarchie.

Classe	Identifiant	Règles
C_1	$\{ub\}$	r_1, r_6
C_2	$\{3h-5h\}$	r_2, r_5
C_3	$\{non\ tactile\}$	r_4, r_8
C_4	$\{tactile\}$	r_3
C_5	$\{2Mp-5Mp\}$	
C_6	$\{monobloc\}$	
C_7	$\{ub, 3h-5h\}$	r_1, r_2
C_8	$\{non\ tactile, 3h-5h\}$	r_8
C_9	$\{tactile, monobloc\}$	r_3
C_{10}	$\{non\ tactile, 2Mp-5Mp\}$	r_4
C_{11}	$\{3h-5h, 2Mp-5Mp\}$	r_5, r_7, r_9
C_{12}	$\{ub, 2Mp-5Mp\}$	r_6
C_{13}	$\{3h-5h, 2Mp-5Mp, ub\}$	r_7
C_{14}	$\{3h-5h, 2Mp-5Mp, non\ tactile\}$	r_9

Table 3.4 – Classes de règles d'association

Une règle est représentée par une sphère placée sur un cône comme on peut le voir sur la figure 3.16. La surface visible de la sphère correspond au support de la règle, la longueur du cône représente la confiance. Les règles sont visualisées par classe. Chaque classe est divisée en deux arènes : une arène pour les règles spécifiques et une autre pour

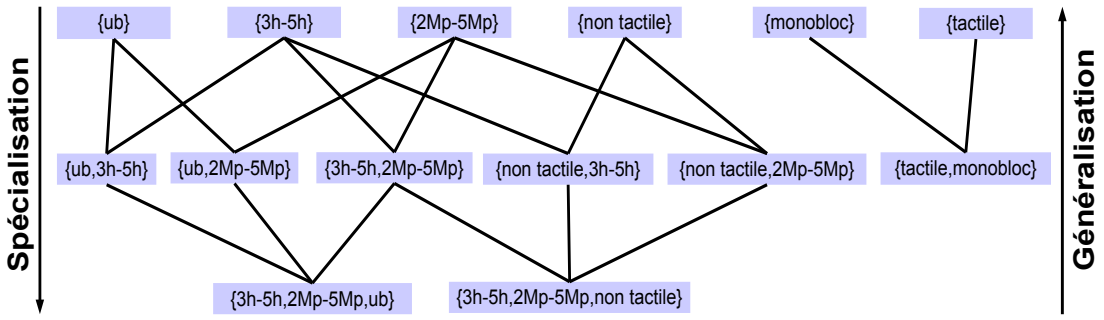


FIGURE 3.15 – Relations de spécialisation/généralisation entre classes de règles

les règles générales. Dans chaque arène, les règles sont placées à des hauteurs différentes en fonction de leur intensité d'implication. L'intensité d'implication est une mesure qui évalue l'imprévisibilité d'une règle [BKGG03]. Plus une règle est imprévisible, plus son intensité d'implication est grande. Les règles qui ont la plus grande intensité d'implication sont placées en bas. L'intensité diminue au fur et à mesure qu'on monte en hauteur (voir figure 3.17). La couleur d'une sphère et de son cône exprime une moyenne pondérée du support, de la confiance et de l'intensité d'implication de cette règle. Cette moyenne donne une vue globale de la règle. Plus elle est élevée, plus la couleur est foncée. L'utilisateur peut interagir avec la visualisation en filtrant les règles selon des mesures d'intérêt tels que le support, la confiance, etc. Il peut aussi passer d'une classe à une autre en cliquant sur une règle. Par exemple, en cliquant sur $r_5 : \{3h-5h\} \Rightarrow \{2Mp-5Mp\}$ de la classe d'identifiant $\{3h-5h, 2Mp-5Mp\}$, on passe à la classe d'identifiant $\{3h-5h\}$ qui est le corps de r_7. Inversement, en cliquant sur $r_7 : \{3h-5h, 2Mp-5Mp\} \Rightarrow \{ub\}$, on passe à la classe d'identifiant $\{3h-5h, 2Mp-5Mp, ub\}$ qui est l'union de la tête et du corps de r_7.

Cette visualisation ne permet pas d'avoir une vue globale des ensembles de motifs car seule une partie de ces ensembles est représentée dans les arènes.

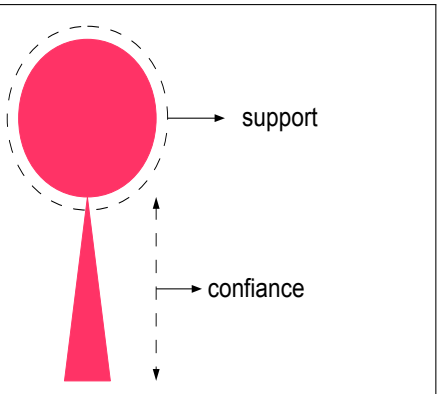


FIGURE 3.16 – Une règle d'association représentée par une sphère au-dessus d'un cône

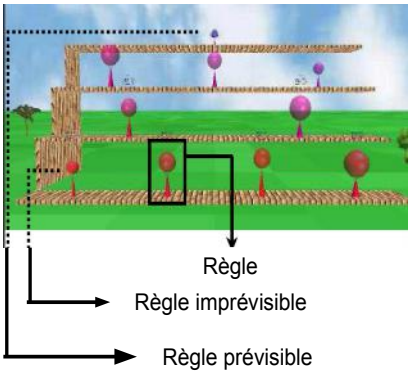


FIGURE 3.17 – Représentation d'une classe de règles d'association avec une métaphore visuelle

3.5 Les matrices

Les matrices constituent certainement le moyen le plus utilisé pour représenter des motifs. Elles permettent essentiellement de mettre en évidence les mesures de qualité des motifs. Nous distinguons deux types de matrices : les matrices à deux dimensions (2D) [BMR99, YN04, ZLTX05, CJR06, CRC07, CHYN07] et les matrices à trois dimensions (3D) [BKK97, CZ03, YN04, WWT99, ZLTX05].

3.5.1 Les matrices 2D

Boulicaut et ses collègues ont proposé une représentation matricielle d'ensembles de règles d'association [BMR99]. Le principe est de placer des itemsets sur chaque axe de la matrice. Une règle est représentée par une cellule. La tête de la règle est l'itemset affiché en ligne et son corps est l'itemset affiché en colonne ou inversement. Dans chaque cellule correspondant à une règle, un couple de valeurs qui représentent son support et sa confiance est affiché comme on le voit dans le tableau 3.18.

	{3h-5h}	{ub}	{monobloc}	{2Mp-5Mp}	{non tactile}
{ub}	$\frac{6}{11}$,1				
{3h-5h}		$\frac{6}{11}$,$\frac{6}{7}$			
{tactile}			$\frac{5}{11}$,1		
{non tactile}				$\frac{6}{11}$,1	
{3h-5h}				$\frac{6}{11}$,$\frac{6}{7}$	
{ub}				$\frac{5}{11}$,$\frac{5}{6}$	
{3h-5h,2Mp-5Mp}		$\frac{5}{11}$,$\frac{5}{6}$			
{non tactile}	$\frac{5}{11}$,$\frac{5}{6}$				
{3h-5h,2Mp-5Mp}					$\frac{5}{11}$,$\frac{5}{6}$

FIGURE 3.18 – Représentation d'un ensemble de règles d'association avec une matrice

Dans [YN04], les auteurs expriment la confiance des règles par une couleur graduelle des cellules. Plus la confiance est grande, plus la couleur de la cellule est sombre. La figure 3.19 montre une représentation des règles du tableau 3.2 avec une matrice 2D. Par ailleurs, une interface qui permet d'interagir avec la représentation est proposée. Elle offre la possibilité de sélectionner les items qui doivent figurer dans le corps et la tête des règles, de fixer un seuil de support et de confiance et de sélectionner les règles individuellement pour ensuite les visualiser avec une matrice 2D ou 3D.

Dans [CJR06, CRC07], les auteurs affichent plusieurs mesures d'intérêt dans une cellule en utilisant des portions colorées. Une mesure est représentée par une portion dont la couleur est graduelle en fonction de la valeur. La figure 3.20 montre une représentation où le support et la confiance sont visualisés par deux triangles.
Couturier et ses collègues proposent d'employer des outils issus du domaine d'IHM [4] pour mettre en évidence le contenu des cellules. Plus précisément, ils utilisent un FEV [5] [Fur06]. Comme le montre la figure 3.21, une cellule est détaillée en déformant la matrice par

4. Interface Homme Machine
5. FishEyeView

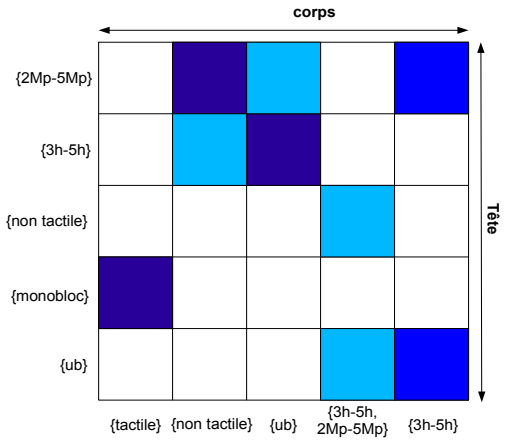


FIGURE 3.19 – Matrice 2D avec représentation du support des règles

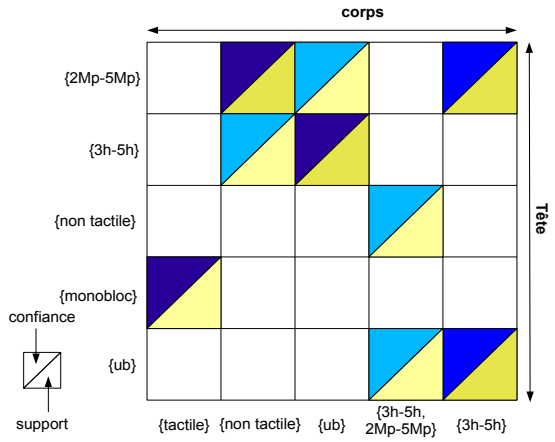


FIGURE 3.20 – Matrice 2D avec représentation du support et de la confiance des règles

rapport à un point focal : les cellules qui sont loin du point de focal sont réduites alors que les cellules qui sont près du foyer sont agrandies.

Cette approche est étendue dans [CHYN07] pour fournir une vue globale d'un ensemble de règles d'association. Plus précisément, il s'agit de visualiser un résumé d'un ensemble de règles. Le principe est de diviser l'ensemble de règles en plusieurs clusters puis de choisir un représentant pour chaque cluster. Le représentant d'un cluster est sélectionné suivant une mesure d'intérêt (le support, la confiance, le lift, etc). Par exemple, on peut sélectionner

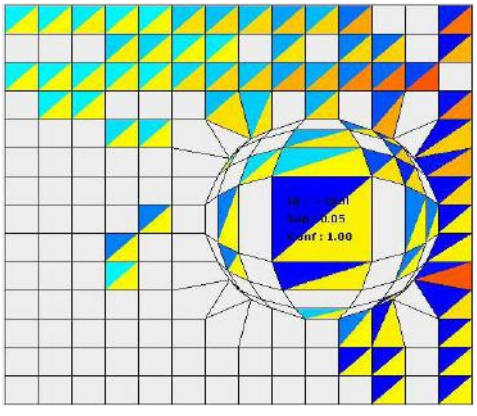


FIGURE 3.21 – Détail d'une règle d'association avec un FEV

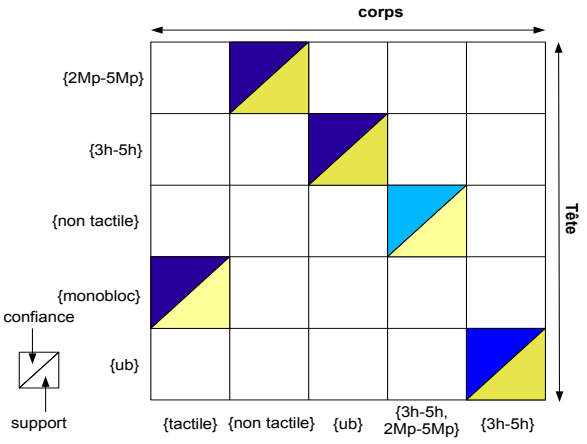
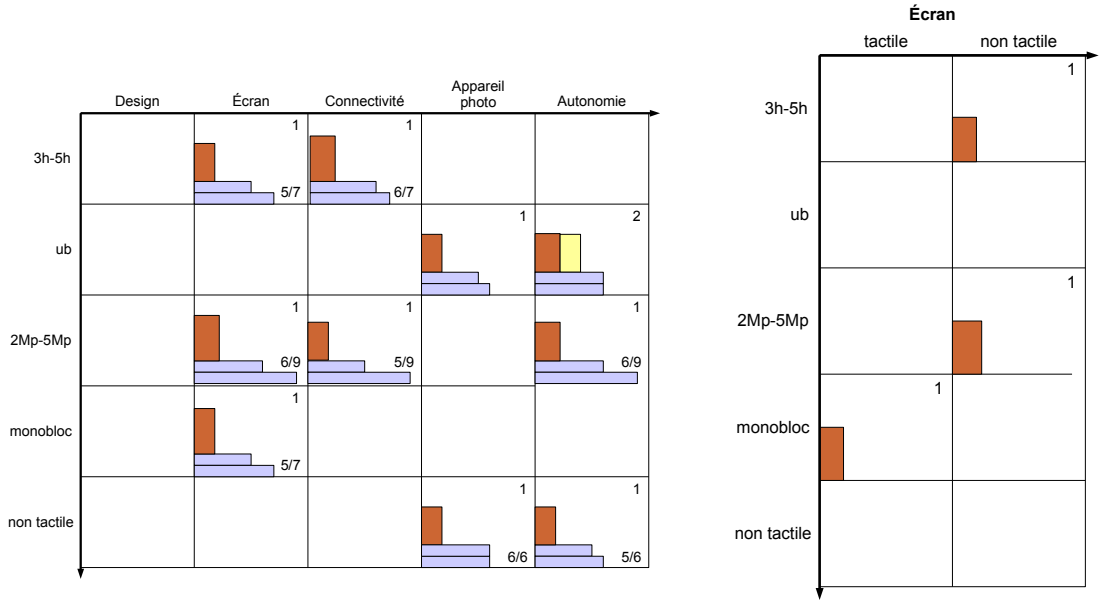


FIGURE 3.22 – Matrice 2D avec représentation du support et de la confiance du représentant des clusters par des couleurs graduelles

la règle ayant la plus grande valeur pour la mesure choisie. Les représentants sont ensuite visualisés par les cellules de la matrice. Les itemsets sur les axes correspondent à la tête

et au corps des représentants de clusters. Chaque cellule visualise les mesures d'intérêt du représentant correspondant. La figure 3.22 montre un exemple de représentation avec cette méthode. L'ensemble de règles est divisé en cinq clusters : $\mathcal{C}_1 = \{r_1, r_8\}$, $\mathcal{C}_2 = \{r_2, r_7\}$, $\mathcal{C}_3 = \{r_3\}$, $\mathcal{C}_4 = \{r_4, r_5, r_6\}$ et $\mathcal{C}_5 = \{r_9\}$. Chaque cluster contient des règles qui ont la même tête et son représentant est la règle qui a la plus grande confiance.
D'autre part, un cluster peut être détaillé en sélectionnant la cellule correspondant à son représentant. Les règles de ce cluster sont alors visualisées avec une matrice 2D comme celle de la figure 3.20 ou une matrice 3D (cf. section 3.5.2).

Une approche similaire à celle qui est décrite précédemment est présentée dans [ZLTX05]. Elle s'adresse en particulier aux règles de classification. Les cellules représentent des clusters constitués de règles qui partagent la même tête, i.e la même classe. La figure 3.23(a) montre une représentation de notre ensemble de règles suivant cette méthode. En abscisse, nous avons des attributs et en ordonnée, nous avons des classes, i.e des



(a) Représentation matricielle suivant tous les attributs

(b) Représentation matricielle suivant l'attribut Design

Figure 3.23 – Représentation matricielle de clusters de règles d'association

têtes de règles. Chaque cellule de la matrice affiche des règles ayant la même tête et ayant dans leur corps une valeur du domaine de l'attribut en abscisse. Une règle est représentée par une barre verticale dont la largeur et la hauteur correspondent respectivement au support et à la confiance de la règle. Le nombre de règles est affiché dans la cellule. Deux autres barres horizontales sont ajoutées dans la partie inférieure de chaque cellule pour exprimer le nombre de données couvertes par les règles représentées dans cette cellule et le nombre total de données qui contiennent une valeur du domaine de l'attribut en abscisse et la valeur en ordonnée. Par exemple, dans la cellule correspondant à *Écran* et $3h - 5h$, nous avons la règle $\{non\ tactile\} \Rightarrow \{3h - 5h\}$. Son support est $\frac{5}{11}$ et sa confiance est $\frac{5}{6}$.

Il y a 5 données qui sont couvertes par cette règles et il y a 7 données qui contiennent une valeur de $dom(\acute{E}cran)$ et l'item $3h - 5h$.
Pour détailler une règle, il suffit de poser la souris sur la barre correspondante qui provoque l'affichage d'une description textuelle de la règle. Notons qu'une règle est représentée dans autant de cellules qu'elle a d'items dans le corps. Les attributs et les classes sont choisis par l'utilisateur. Les règles peuvent être détaillées individuellement (en posant la souris sur une barre) ou par classe (en sélectionnant une cellule).
Cette visualisation peut aussi être détaillée par rapport à un attribut en affichant les valeurs de son domaine en ordonnée. La figure 3.23(b) montre une représentation des règles suivant l'attribut *Écran*. Dans chaque cellule, il y a au plus une règle qui est représentée. Elle contient dans son corps la valeur affichée en abscisse.
Globalement, cette représentation est difficile à interpréter car il y a beaucoup d'informations qui sont visualisées à la fois. Cette interprétation est encore plus difficile lorsque le nombre de règles est important. D'autre part, les règles sont représentées de manière redondante, i.e. autant de fois qu'elles ont d'items dans leur corps.

La représentation avec les matrices 2D est intéressante si les itemsets à afficher sur les axes ne sont pas nombreux. Par contre, les matrices deviennent très grandes lorsqu'il y a beaucoup d'itemsets, ce qui ne permet pas d'avoir une vue globale de l'ensemble de motifs.

3.5.2 Les matrices 3D

La visualisation d'ensembles de règles d'association avec des matrices 3D a été introduite dans [BKK97] avec MineSet [6]. Ces matrices appelées matrices item à item sont composées de trois axes dont deux sont utilisés pour afficher les items se trouvant dans la tête et le corps des règles. Le troisième axe sert à exprimer les mesures d'intérêt des motifs. Une règle est schématisée par une barre dont la couleur graduelle exprime son support et

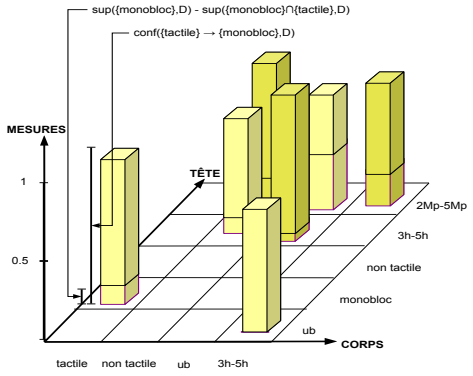


FIGURE 3.24 – Matrice item à item en 3D

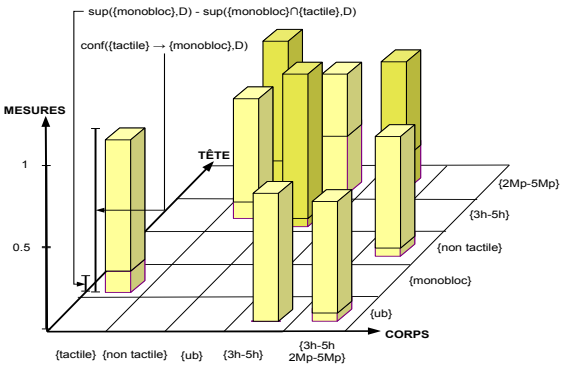


FIGURE 3.25 – Matrice itemset à itemset en 3D

la hauteur correspond à sa confiance. La probabilité d'avoir l'item de la tête sans celui du corps de la règle est aussi représentée par une portion de la barre. La figure 3.24 montre une

6. Logiciel disponible sur http://www.sgi.com/software/mineset/index.html. Il fournit plusieurs outils pour visualiser et explorer des données et des résultats d'extraction de connaissances.

représentation des règles item à item du tableau 3.2. L'utilisateur peut effectuer plusieurs opérations telles que la sélection ou le filtrage de règles.
Cette représentation est améliorée en matrice itemset à itemset dans [CHYN07]. Dans cette approche, les items sont regroupés pour pouvoir visualiser des règles avec plusieurs items dans la tête et dans le corps. La figure 3.25 est obtenue en ajoutant les règles r_7 et r_9.
Ces matrices souffrent généralement d'occlusion lorsque l'ensemble de règles est dense à cause de la hauteur des barres qui varie en fonction des mesures d'intérêt des motifs.

Une approche différente est proposée dans [CZ03] où la représentation en 3D est focalisée sur le détail de la tête des règles. Le premier axe correspond au corps des règles, le second axe représente des items qui sont présents dans la tête des règles et le dernier axe est utilisé pour les mesures de qualité. La figure 3.26 illustre ce mode de représentation. Si un item est présent dans la tête d'une règle, un cube est matérialisé dans la cellule qui correspond à l'intersection de cet item et de la règle. Le support et la confiance de la règle sont visualisés sur le dernier axe. Pour diminuer le nombre de motifs à afficher

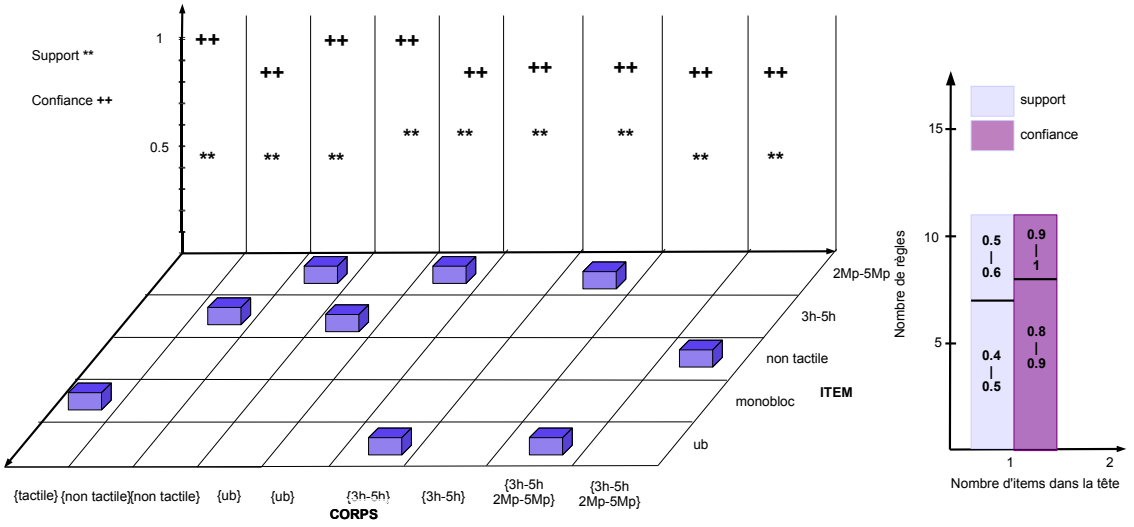


Figure 3.26 – Matrice itemset à item en 3D

simultanément dans la matrice, un diagramme qui montre une vue d'ensemble des règles est d'abord présenté à l'utilisateur. Les règles sont organisées par classe selon le nombre d'items contenus dans la tête. Chaque classe est symbolisée par une barre divisée en intervalles de support (partie gauche) et de confiance (partie droite). Par exemple, dans le diagramme de la figure 3.26, toutes les règles ont un item dans la tête. Parmi ces règles, trois ont une confiance appartenant à l'intervalle $[0.9, 1]$ et huit ont une confiance appartenant à l'intervalle $[0.8, 0.9[$. En cliquant sur la seule classe du diagramme, on obtient la matrice de gauche qui donne plus de détails sur les règles de cette classe.

En définitive, les matrices sont particulièrement adaptées à la représentation des règles d'association. En effet, elles permettent de mettre en évidence les relations entre la tête et le corps d'une règle mais aussi entre plusieurs règles. Elles ne sont d'ailleurs utilisées que pour ce type de motif. Cependant, le nombre d'items ou d'itemsets affichés sur les axes est

dépendant des motifs à visualiser. La taille des matrices n'est donc pas maitrisée et peut atteindre des proportions très élevées, ce qui ne permet pas d'avoir une vue globale des ensembles de motifs. D'autre part les matrices sont souvent très éparses.

3.6 Les cubes

Un cube est une structure communément utilisée pour analyser des données multi-dimensionnelles. Typiquement, un cube est composé de dimensions. Chaque dimension correspond à un attribut pouvant appartenir à une hiérarchie. Une dimension est visualisée par un axe gradué par les valeurs du domaine de cet attribut. Une cellule du cube est identifiée par un n-uplet de valeurs appartenant au domaine des attributs et correspond à une donnée ou un ensemble de données. Les cubes sont adaptés dans [LZBX06] pour l'exploration d'ensembles de règles de classification. Dans leur approche, une dimension est destinée à l'attribut de classe (l'attribut qui apparaît dans la tête des règles) et les autres dimensions sont utilisées pour afficher les attributs que l'on retrouve dans le corps des règles. Une règle est représentée par une cellule. Sa tête correspond à une valeur du domaine de l'attribut de classe. Son corps contient les valeurs du domaine des autres attributs. La cellule d'une règle contient le nombre de données couvertes par cette dernière.

		Prix					
		< 100	$100 - 200$	$200 - 300$	> 300		
Écran	tactile	0	1	1	3	monobloc	*Design*
		0	0	0	0	coulissant	
	non tactile	2	1	1	0	monobloc	
		1	2	1	0	coulissant	

FIGURE 3.27 – Représentation d'un ensemble de règles de classification sous forme de cube

Les auteurs adaptent certaines opérations OLAP [7] pour explorer les règles. Il s'agit plus précisément des opérations *Rollup*, *Drilldown*, *Slice* et *Dice*. Dans leur approche, l'opération *Rollup* consiste à monter dans la hiérarchie d'une dimension ou à enlever une dimension et l'opération *Drilldown* consiste à effectuer l'opération inverse, i.e. à descendre dans la hiérarchie d'une dimension ou ajouter une dimension. L'opération Slice effectue la sélection d'une dimension du cube et l'opération Dice est une extension du Slice qui permet de sélectionner deux dimensions ou plus. Les figures 3.28 et 3.29 montrent des cubes obtenus en appliquant respectivement un *R*ollup et un Slice sur le cube de la figure 3.27. Notons que les opérations *Rollup* et *Drilldown* sont différentes des opérations classiques. En effet, l'ajout ou la suppression d'une dimension ne permet pas d'agréger les règles mais elle permet plutôt de passer d'un sous-ensemble à un autre.

Ces travaux ont été étendus récemment dans [BSML09] où les auteurs proposent une représentation plus intuitive des cellules. Les valeurs dans les cellules sont remplacées par des rectangles dont la surface est proportionnelle au nombre de données couvertes par les règles et leur couleur (graduelle) varie en fonction du lift de ces règles.

7. On Line Analytical Processing

		Prix			
		< 100	$100 - 200$	$200 - 300$	> 300
Écran	tactile	0	1	1	3
	non tactile	3	3	2	0

FIGURE 3.28 – *R*ollup : suppression de la dimension *Design*

Prix			
< 100	$100 - 200$	$200 - 300$	> 300
3	4	3	3

FIGURE 3.29 – Slice : sélection de la dimension *Prix*

Ces approches sont très similaires à celles que nous proposons mais elles sont limitées aux règles de classification et les représentations obtenues donnent une vue incomplète de l'ensemble de règles, car seule une partie de ces règles est représenté. D'autre part, elles ne sont pas adaptées s'il y a beaucoup de règles à visualiser.

3.7 Synthèse sur les méthodes de représentation visuelle

Dans cette section, nous résumons les méthodes décrites précédemment en nous basant sur les critères que nous avons énoncés dans la section 3.1. Le tableau 3.5 montre un récapitulatif des caractéristiques de ces méthodes. Les colonnes présentent les différents critères et les lignes contiennent les références des approches étudiées. Une croix dans une cellule indique que la méthode en ligne possède la caractéristique en colonne. La légende du tableau donne la signification des abréviations et acronymes qui sont utilisés dans le tableau.

Type de motifs La majorité de ces méthodes est destinée aux règles d'association, en particulier aux règles de classification. La structure des itemsets fréquents étant assez proche de celle des données classiques, ils peuvent aussi être explorés avec des outils de visualisation de données. Ce qui pourrait expliquer le nombre faible de représentations proposées pour les itemsets fréquents en particulier.

Richesse de la représentation Toutes les méthodes permettent de représenter les mesures d'intérêt des motifs. D'autre part, les liens entre les motifs sont représentés avec presque toutes les méthodes sauf pour celles qui utilisent les tableaux [ZP03]. Par contre, seules deux méthodes permettent d'avoir une vue globale des ensembles de motifs. La première méthode [CHYN07] s'apparente à notre approche dans le sens où la visualisation est un résumé de l'ensemble qui est visualisé. Cependant, notre approche apporte en plus la possibilité de voir les ensembles de motifs à plusieurs niveaux de détail. La deuxième méthode permet d'avoir une vue globale des motifs suivant leur support [LC09]. Mais cette représentation introduit souvent des faux positifs. En effet, elle montre des liens entre des items qui n'apparaissent pas ensemble dans un itemset fréquent.

Lisibilité A l'exception de la méthode proposée dans [CHYN07], aucune méthode n'est adaptée pour représenter de grands ensembles de motifs, cela se traduit par des visualisations très grandes qui ne peuvent pas tenir sur un écran et/ou par des occlusions.

Interprétabilité Une des méthodes de visualisation fournit des visualisations qui ne sont pas consistantes. Plus précisément, il s'agit des méthodes proposées dans [Yan05] et [LC09]. En effet, dans l'approche décrite dans [Yan05] qui est basée sur les coordonnées parallèles, les lignes de différentes règles peuvent se chevaucher, ce qui provoque une confusion introduisant des faux positifs. Dans la seconde méthode basée sur les polylignes, la projection de plusieurs polylignes (représentant des règles de même support) sur une ligne fait apparaître des connexions entre les nœuds qui n'existent pas dans les lignes d'origine, ce qui se traduit par la représentation d'un itemset qui n'existe pas dans l'ensemble qui est visualisé. Par ailleurs, la méthode exposée dans [ZLTX05] fournit des visualisations redondantes. En effet, dans cette méthode, une règle est visualisée dans autant de cellules qu'elle a d'items dans son corps qui appartiennent aux attributs sélectionnés. Donc les informations sur cette règle sont partagées dans toutes les cellules où elle apparaît. Par exemple, si les attributs *Autonomie* et *Appareil photo* sont sélectionnés alors la règle $\{3h-5h, 2Mp-5Mp\} \Rightarrow \{ub\}$ sera représentée dans deux cellules, à savoir celle correspondant à l'item *ub* et à l'attribut *Autonomie* et celle correspondant à l'item *ub* et à l'attribut *Appareil photo* (cf. figure 3.23(a)).

Navigation Toutes les méthodes permettent d'effectuer les navigations classiques telles que le filtrage et la sélection. Par contre, seules les méthodes basées sur les cubes permettent de naviguer entre plusieurs niveaux d'agrégation [LZBX06, BSML09].

Contrôle de la taille des visualisations La majorité des méthodes permettent de contrôler la taille des visualisations de manière directe ou de manière indirecte. Le contrôle indirect passe souvent par le choix d'attributs, d'imets ou d'imetsets à partir desquels les visualisations sont construites [CHYN07, ZLTX05, BKK97, LZBX06, BSML09] ou encore le filtrage et la sélection. Les méthodes qui ne permettent pas de contrôler la taille des visualisations ne sont évidemment pas adaptées pour l'exploration de grands ensembles de motifs.

		Motifs		Richesse de la représentation			Lisibilité		Interprétabilité		Navigation		Contrôle de la taille	
		IF	RA	Liens	Mesures	Glob.	TG	Occl.	Cons	Red	FS	Agr.	Direct	Indirect
Tableaux	[ZP03]		×		2		×		×		×		×	
Graphes	[KMR+94]		×	×	2		×	×	×		×		×	
	[BB04]		×	×	2			×	×		×			
Formes géométriques	[HK81]		×	×	2		×		×		×		×	
	[LIC08]	×		×	1		×	×	×		×	×		
	[LC09]	×		×	1	×	×	×			×	×		
	[BB04]		×	×	1			×	×		×			
	[Yan05]	×	×	×	2		×	×			×			
	[BGB03]		×	×	3		×		×		×			
Matrices	[BMR99]		×	×	2		×		×		×			
	[YN04]		×	×	1		×		×		×		×	
	[CHYN07]		×	×	2	×			×		×			×
	[ZLTX05]		×	×	2		×		×	×	×			×
	[BKK97]		×	×	3		×		×		×			×
	[CZ03]		×	×	2		×		×		×			
Cubes	[LZBX06]		×	×	1		×		×		×			×
	[BSML09]		×	×	2		×		×		×			×

Table 3.5 – Comparaison des méthodes de visualisation

- **IF :** itemsets fréquents
- **RA :** règles d'association
- **Liens :** représentation de liens entre les motifs
- **Mesures :** nombre de mesures d'intérêt des motifs qui peuvent être représentées avec la visualisation
- **Occl. :** occlusion de la visualisation
- **TG. :** visualisation trop grande
- **Glob. :** capacité de la visualisation à fournir une vue globale des ensembles de motifs
- **Cons. :** consistance de la visualisation
- **Red. :** redondance de l'information qui est visualisée
- **FS :** filtrage et sélection
- **Agr. :** agrégation

3.8 Conclusion

Dans ce chapitre nous avons passé en revue diverses méthodes de visualisation. Une synthèse sur ces méthodes nous a permis de faire quelques constats : Le premier constat est que les visualisations deviennent inexploitables lorsque le nombre de motifs à visualiser est trop important. Le deuxième constat est que peu de visualisations donnent une vue globale des ensembles de motifs. En effet, la majorité des méthodes visualisent les motifs au niveau le plus détaillé.

Conclusion

Nous avons étudié dans cette partie des méthodes de construction de résumés et des méthodes de représentation visuelle d'ensembles de motifs. A l'issue de cette étude, nous identifions deux insuffisances majeures :

- Il n'y a pas de cadre qui est proposé pour explorer les motifs à partir de leur résumé.
- Les méthodes de visualisation proposées sont limitées lorsqu'il s'agit de représenter un ensemble de motifs volumineux.

Dans la deuxième partie de ce mémoire, nous présentons un cadre générique pour la construction de résumés pour l'exploration de grands ensembles de motifs. Nous nous inspirons de l'approche décrite dans [CK05] où les auteurs définissent un résumé d'ensemble de transactions comme étant un ensemble d'itemsets tel que chaque itemset représente un sous-ensemble de transactions et chaque transaction est représentée par au moins un itemset. Notre contribution repose sur l'extension suivante de la fonction de résumé (cf. définition 1.3.3) qui permet d'obtenir des résumés ayant les mêmes propriétés de couverture que ceux qui sont proposés dans cette approche.

Définition 3.8.1 (Fonction de résumé couvrante) *Soient $\mathcal{P}$ et $\mathcal{S}$ deux langages de motifs. Une fonction de résumé couvrante est une fonction de résumé $\Psi_\alpha : 2^{\mathcal{P}} \rightarrow 2^{\mathcal{S}}$ qui associe à un ensemble de mofifs $P \subseteq \mathcal{P}$, un ensemble de motifs $S \subseteq \mathcal{S}$ tel que :*

(i) Chaque motif de P est couvert par au moins un motif de S ;

(ii) Chaque motif de S couvre au moins un motif de P ;

(iii) $|S| \leq |P|$;

(iv) Le paramètre α permet de contrôler la taille de S.

L'ensemble de motifs S est un résumé P.

Cette définition offre la possibilité de prendre en compte n'importe quel type de motifs aussi bien pour les ensembles à résumer que pour les résumés. De plus, elle assure une couverture complète des ensembles qui sont résumés.

D'autre part, notre cadre permet de construire des résumés qui peuvent être représentés sous forme de cube, ce qui permet de les interpréter facilement et de manière intuitive mais aussi de les explorer avec les opérateurs de navigation OLAP.

Deuxième partie

Contribution à l'exploration de grands ensembles de motifs

Introduction

La partie précédente a montré les insuffisances des méthodes de construction de résumés et celles des représentations visuelles d'ensembles de motifs. Toutes ces méthodes convergent vers le même objectif qui est de faciliter l'exploration de grands ensembles de motifs. Cependant, les méthodes de construction de résumés souffrent du manque de technique d'exploration adaptée et les méthodes de visualisation sont limitées lorsqu'il s'agit de représenter de grands ensembles de motifs. Dans cette partie, notre adoptons une démarche qui regroupe les avantages de ces deux approches en proposant un cadre générique qui permet de construire des résumés de grands ensembles de motifs qui peuvent être représentés sous forme de cubes. Le résumé donne une vue globale et supporte de grands ensembles de motifs et la structuration sous forme de cube permet d'avoir une meilleure lisibilité des résumés mais aussi de pouvoir utiliser les opérateurs OLAP. Nous présentons dans le chapitre 4 une description de notre cadre et nous l'instancions pour résumer des ensembles de règles d'association dans le chapitre 5.

Chapitre 4

Cadre générique pour la construction de résumés d'ensembles de motifs

Les cubes ont largement fait leur preuve dans le domaine de l'analyse de données multidimensionnelles qui proviennent des entrepôts de données [1] [CNCL10, PLPP10]. Ces données sont habituellement représentées sous forme de cubes et explorées en appliquant des opérations de navigation OLAP. Nous avons vu dans le chapitre 2 que les cubes ont été utilisés ces dernières années pour représenter des ensembles de règles de classification [LZBX06, BSML09]. Le but de ces approches est d'utiliser des adaptations d'opérations OLAP pour l'exploration des règles. Les représentations obtenues ne donnent cependant pas une vue globale des règles et s'appliquent en particulier aux règles de classification. Dans ce chapitre nous proposons un cadre générique permettant de construire des résumés d'ensembles de motifs qui peuvent être représentés sous forme de cubes. Dans la section 4.1, nous donnons des définitions préliminaires relatives aux cubes de données. Ensuite, nous présentons notre cadre dans la section 4.2. Puis nous décrivons dans la section 4.3 des opérateurs OLAP qui permettent de naviguer dans un espace de résumés. Enfin, nous proposons deux méthodes de construction de résumés dans la section 4.4. Le tableau 4.1 décrit les notations que nous utilisons dans ce chapitre.

4.1 Une brève présentation des cubes de données

Les cubes de données sont construits à partir de modélisations multidimensionnelles d'entrepôts de données.

4.1.1 Modélisations multidimensionnelles

Les modélisations multidimensionnelles les plus utilisées sont : le schéma en étoile [Kim96], le schéma en flocon [JLS99] et le schéma en constellation [CD97]. Un schéma

1. Un entrepôt de données (Data Warehouse) est une collection de données orientées pour un sujet, intégrées, non volatiles et historisées, organisées pour le support du processus d'aide à la décision. Cette définition est donnée par William Harvey (Bill) Inmon.

Notation	Description
$\mathcal{P}$	Langage de motifs
P	Ensemble de motifs ($P \subseteq \mathcal{P}$)
p	Motif ($p \in \mathcal{P}$)
$\mathcal{A}$	Ensemble d'attributs
A	Attribut ($A \in \mathcal{A}$)
$dom(A)$	Domaine de A
a	Valeur du domaine d'un attribut A ($a \in dom(A)$)
$dom(A)^+$	Domaine étendu de A
c	Cube
D	Dimension d'un cube
μ	Fonction qui calcule les valeurs à afficher dans les cellules d'un cube
$Ref(c)$	domaine de définition de la fonction μ du cube c
C	Schéma du cube c
$Ref(C)$	Ensemble de toutes les références définies sur C
$\mathcal{C}_\mathcal{A}$	Ensemble de tous les schémas définis sur $\mathcal{A}$
$\mathcal{L}_\mathcal{A}$	Langage des références définies sur l'ensemble des schémas $\mathcal{C}_\mathcal{A}$
$\mathcal{S}_\mathcal{A}$	Ensemble des résumés basés sur un Schéma défini sur $\mathcal{A}$
S_C	Résumé basé sur le schéma C

Table 4.1 – Notations utilisées dans ce chapitre

en étoile est composé d'une table de faits et de tables de dimensions. Les faits sont ce sur quoi va porter l'analyse et les dimensions sont les axes avec lesquels on veut faire l'analyse des données. Notons que les tables de faits et de dimensions sont des relations (cf. chapitre 1, section 1.1).

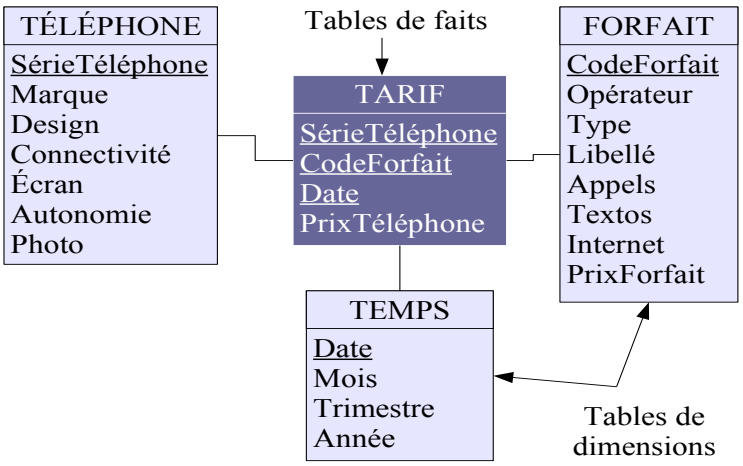


Figure 4.1 – Modélisation multidimensionnelle : exemple de schéma en étoile

Exemple 4.1.1 *La figure 4.1 montre un exemple de schéma en étoile. On y distingue la table de faits Tarif au centre reliée aux tables de dimensions Téléphone, Forfait et Temps par le mécanisme de clé étrangère. Les clés étrangères (attributs identifiants) SérieTéléphone, CodeForfait et Date provenant des tables de dimensions constituent la clé primaire*

de la table de faits. L'attribut PrixTéléphone qui ne fait pas partie de la clé primaire de la table de faits est un attribut mesure.

Dans un schéma en étoile, chaque dimension est décrite par une seule table dont les attributs représentent les différents niveaux de granularité possibles. Par exemple, dans la table de dimensions $Temps$, les attributs $Date$, $Mois$, $Trimestre$ et $Année$ représentent le temps du niveau le plus détaillé ($Date$) au niveau le plus agrégé ($Année$).

Exemple 4.1.2 *La figure 4.2 montre un cube construit à partir du schéma en étoile de la figure 4.1. Les cellules contiennent le prix (le plus bas) par trimestre des téléphones vendus avec un forfait durant l'année 2010. Ce prix est représenté dans la table de faits par l'attribut prixTéléphone.*

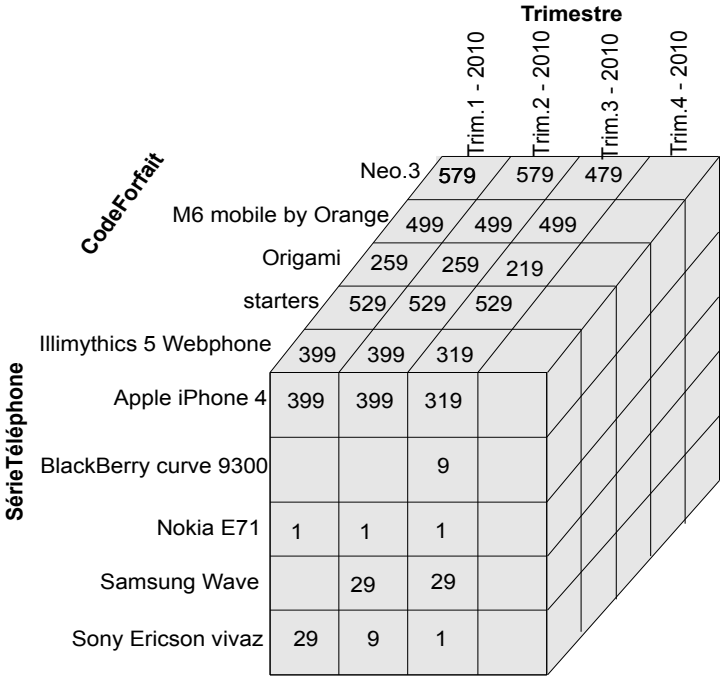


Figure 4.2 – Exemple de cube de données : $c = \langle SérieTéléphone, CodeForfait, Temps, \mu \rangle$

Le schéma en flocon est caractérisé par le fait que les tables de dimensions sont normalisées pour éviter les redondances d'information. Cette normalisation se traduit par la description de chaque dimension par une succession de tables (reliées à l'aide de clés étrangères) qui traduit la granularité de l'information. Par exemple, la table de dimensions $Temps$ peut être normalisée en quatre tables : une pour la date, une pour le mois, une pour le trimestre et une pour l'année.

Enfin, le schéma en constellation est composé de plusieurs tables de faits qui ont des tables de dimensions en commun. Par exemple, on peut avoir une table de faits supplémentaire qui porte sur les appels émis par les clients et qui est reliée à la table de dimensions $Forfait$.

Les faits sont représentés sous forme de cubes où chaque axe correspond à une dimension. Les valeurs du domaine des attributs mesures d'une table de faits sont habituellement des valeurs numériques calculées à partir d'un ensemble de données. La section suivante formalise la notion de cube.

4.1.2 Les cubes de données dans notre contexte

Dans notre approche, pour avoir un cadre général, nous définissons un cube à partir d'un ensemble d'attributs. Cet ensemble peut être formé des attributs des tables de dimensions qui proviennent de n'importe quel schéma ou même des tables d'une base de données relationnelles. Pour exprimer l'absence de valeur pour les attributs, nous introduisons une valeur nulle dans le domaine de chaque attribut.

Soit A un attribut quelconque, nous notons $dom(A)^+$ le domaine de l'attribut A auquel nous ajoutons une valeur nulle noté $nulle_A$, i.e. $dom(A)^+ = dom(A) \cup \{nulle_A\}$. L'ensemble $dom(A)^+$ est appelé le domaine étendu de A. Intuitivement, un cube est constitué d'attributs qui représentent les dimensions et d'une fonction qui calcule les valeurs à afficher dans les cellules. La définition suivante formalise la notion de cube dans notre contexte.

Définition 4.1.1 (Cube) *Étant donnés un ensemble d'attributs $\mathcal{A}$, un cube défini sur $\mathcal{A}$ est un $N+1$-uplet $c = \langle D_1 \cdots, D_N, \mu \rangle$ tel que $\{D_1, \cdots, D_N\} \subseteq \mathcal{A}$ et μ est une fonction partielle définie de $Ref(c) \subseteq dom(D_1)^+ \times \cdots \times dom(D_N)^+$ vers $\mathbb{R}$.*
Le schéma de c noté C est le N-uplet composé des dimensions de c, i.e. $C = \langle D_1, \cdots, D_N \rangle$.

Exemple 4.1.3 *Le cube de la figure 4.2 correspond au 4-uplet $c = \langle SérieTéléphone, CodeForfait, Trimestre, \mu \rangle$ où μ est la fonction qui retourne le prix minimal d'un téléphone vendu avec un forfait pendant un trimestre. Son schéma est $C = \langle SérieTéléphone, CodeForfait, Trimestre \rangle$.*

Dans notre cadre, nous utilisons des schémas définis sur un ensemble d'attributs pour définir nos résumés. Dans la suite de ce chapitre, nous notons $\mathcal{C}_\mathcal{A}$ le langage des schémas définis sur l'ensemble d'attributs $\mathcal{A}$. Étant donnés deux schémas $C, C' \in \mathcal{C}_\mathcal{A}$, C' est plus spécifique que C si tous les attributs de C sont aussi dans C'. La définition suivante est une formalisation de la notion de cellule.

Définition 4.1.2 (Cellule) *Étant donné un cube $c = \langle D_1, \cdots, D_N, \mu \rangle$, une cellule de c est un $N+1$-uplet $\langle a_1, \cdots, a_N, f \rangle$ tel que $f = \mu(a_1, \cdots, a_N)$ où $\langle a_1, \cdots, a_N \rangle$ appartient au domaine de définition de μ.*

Exemple 4.1.4 *$\langle Apple\ iPhone\ 4, Illimythics\ 5\ Webphone, Trim.1\ -\ 2010, 399 \rangle$ est une cellule de c qu'on peut voir dans le cube de la figure 4.2. En effet, elle contient une valeur qui montre que $\langle Apple\ iPhone\ 4, Illimythics\ 5\ Webphone, Trim.1\ -\ 2010 \rangle$ appartient à $Ref(c)$.*

Les cellules d'un cube sont identifiées par des références qui sont définies comme suit :

Définition 4.1.3 (Référence) *Étant donné un cube c de schéma $C = \langle D_1, \cdots, D_N \rangle \in \mathcal{C}_\mathcal{A}$, une référence de c définie sur C est un N-uplet $\langle a_1, \cdots, a_N \rangle$ qui appartient à $Ref(c)$.*

Exemple 4.1.5 $\langle Apple\ iPhone\ 4, Illimythics\ 5\ Webphone, Trim.1\ -\ 2010\rangle$ *est une référence définie sur le schéma* $\langle SérieTéléphone, CodeForfait, Trimestre\rangle$ *du cube décrit dans l'exemple 4.1.3.*

Nos résumés sont des ensembles dont les motifs appartiennent à un langage de références. Dans la suite de ce chapitre, nous notons $\mathcal{L}_{\mathcal{A}}$ le langage des références définies sur l'ensemble des schémas de $\mathcal{C}_{\mathcal{A}}$. Pour définir nos résumés, nous utilisons une relation de couverture qui repose sur la relation de spécialisation/généralisation suivante :

Définition 4.1.4 (Spécialisation/généralisation de références) *Étant données deux références* $s = \langle a_1, \cdots, a_N\rangle$ *et* $s' = \langle a'_1, \cdots, a'_{N'}\rangle$ *appartenant à* $\mathcal{L}_{\mathcal{A}}$*, s' est plus spécifique que s* ($s \preceq s'$) *si* $\{a_1, \cdots, a_N\} \subseteq \{a'_1, \cdots, a'_{N'}\}$.

En d'autres termes, s' est plus spécifique que s si s' contient toutes les valeurs de s.

Exemple 4.1.6 *La référence* $s' = \langle Apple\ iPhone\ 4, Illimythics\ 5\ Webphone, Trim.1\ -\ 2010\rangle$ *est plus spécifique que* $s = \langle Apple\ iPhone\ 4, Illimythics\ 5\ Webphone\rangle$ *car elle contient toutes les valeurs de s.*

Dans la section suivante, nous présentons notre cadre générique qui repose sur les notions définies précédemment.

4.2 Des résumés pour explorer de grands ensembles de motifs

Cette section présente notre première contribution qui est la définition d'un cadre générique pour la construction de résumés d'ensembles de motifs, qui peuvent être représentés sous forme de cubes. Notre choix a porté sur les cubes pour deux raisons. La première raison est que la structuration des résumés sous forme de cubes permet de faciliter leur compréhension et leur interprétation. La deuxième raison est que les opérateurs de navigation OLAP peuvent être utilisés pour explorer les résumés. Intuitivement, nos résumés sont des ensembles de références définis sur un même schéma qui permet de les représenter sous forme de cubes. Nous utilisons une relation de couverture pour relier les motifs des ensembles qui sont résumés aux références d'un résumé. Nous présentons dans la section 4.2.1 les propriétés des relations de couverture qui peuvent être utilisées pour construire de tels résumés. Ensuite, nous proposons dans la section 4.2.2 une formalisation de la notion de résumé basé sur un schéma.

4.2.1 Relation de couverture entre motifs et références de cubes

Notre approche repose sur une relation de couverture entre un langage contenant les motifs qu'on souhaite résumer et un langage de références à partir duquel les motifs du résumé sont choisis. Dans la pratique, le langage des références et la relation de couverture sont définis à partir du langage des motifs. Nous avons vu dans notre état de l'art que les résumés qui permettent de couvrir de manière redondante un motif de l'ensemble qui est résumé sont généralement difficiles à interpréter. En effet, lorsqu'on souhaite avoir des

informations sur un motif particulier qui est couvert par plusieurs motifs du résumé, on se heurte au problème du choix du motif du résumé qu'il faut considérer. Pour éviter de couvrir plusieurs fois un même motif, nous utilisons des relations de couverture ayant la propriété de non redondance définie comme suit :

Définition 4.2.1 (Non redondance) *Étant donnés un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_\mathcal{A}$ et une relation de couverture $\triangleleft$ entre $\mathcal{P}$ et $\mathcal{L}_\mathcal{A}$, $\triangleleft$ est non redondante si pour toutes références s et s' définies sur un schéma $C \in \mathcal{C}_\mathcal{A}$, $couverture(s, \mathcal{P}) \cap couverture(s', \mathcal{P}) = \emptyset$.*

Cette définition signifie que si une relation de couverture $\triangleleft$ entre un langage de motifs $\mathcal{P}$ et un langage de références $\mathcal{L}_\mathcal{A}$ est non redondante alors des références définies sur un même schéma ne couvrent pas conjointement un motif de $\mathcal{P}$.
D'autre part, nous souhaitons pouvoir couvrir totalement tout ensemble de motifs avec des références définies sur un même schéma. Pour atteindre cet objectif, nous utilisons une relation de couverture ayant la propriété de complétude définie comme suit :

Définition 4.2.2 (Complétude) *Étant donnés un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_\mathcal{A}$ et une relation de couverture $\triangleleft$ entre $\mathcal{P}$ et $\mathcal{L}_\mathcal{A}$, $\triangleleft$ est complète si pour tout cube c de schéma schéma $C \in \mathcal{C}_\mathcal{A}$ et pour tout motif $p \in \mathcal{P}$, il existe une référence s définie sur C telle que $s \triangleleft p$.*

En d'autres termes, quel que soit le schéma C défini sur $\mathcal{A}$ et le motif p pris dans $\mathcal{P}$, il existe une référence définie sur C qui couvre p. Dans ce contexte, pour assurer la complétude de notre relation de couverture, nous ajoutons au langage de références $\mathcal{L}_\mathcal{A}$, une référence notée $s_{\langle\rangle}$ qui est associée au schéma $C = \langle\rangle$ et qui couvre tous les motifs de $\mathcal{P}$.

Exemple 4.2.1 *Considérons un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_\mathcal{A}$ et une relation de couverture $\triangleleft$ entre $\mathcal{P}$ et $\mathcal{L}_\mathcal{A}$ décrits ci-dessous.*

- *$\mathcal{P}$ est constitué des requêtes fréquentes qu'on peut effectuer sur la base de données relationnelles $\mathcal{D}$ de schéma $sch(\mathcal{D}) = \{Téléphone, Forfait, Temps, Tarif\}$ avec :*
  - $sch(Téléphone) = \{SérieTéléphone, Marque, Design, Connectivité, Écran, Autonomie, Appareil\ Photo\}$;
  - $sch(Forfait) = \{CodeForfait, Opérateur, Type, Libellé, Appels, Textos, Internet, PrixForfait\}$;
  - $sch(Temps) = \{Date, Mois, Trimestre, Année\}$;
  - $sch(Tarif) = \{SérieTéléphone, CodeForfait, Date, PrixTéléphone\}$.
- *$\mathcal{A}$ est l'ensemble des attributs des tables de $\mathcal{D}$.*
- *$\triangleleft$ est définie telle qu'une référence $X = \langle a_1, \cdots, a_N\rangle$ couvre une requête $q = \pi_{A_1,\ldots,A_I}(\sigma_{B_1=b_1,\ldots,B_J=b_J}(R_1 \bowtie \ldots \bowtie R_K))$ si $\{a_1, \cdots, a_N\} \subseteq \{b_1, \cdots, b_J\}$.*

Nous voyons clairement que $\triangleleft$ est complète. En effet, quel que soit le schéma appartenant à $\mathcal{C}_\mathcal{A}$, les références définies sur ce schéma couvrent toutes les requêtes de $\mathcal{P}$. D'autre part, $\triangleleft$ est non redondante car une requête de $\mathcal{P}$ ne peut pas être couverte par plus d'une référence définie sur un schéma qui appartient à $\mathcal{C}_\mathcal{A}$.

Ainsi, si la relation de couverture est non redondante et complète, tout motif de $\mathcal{P}$ est couverte par une seule référence d'un schéma de $\mathcal{C}_\mathcal{A}$ donné.

Dans la suite de ce chapitre, nous utilisons une relation de couverture qui est non redondante et complète. Nous présentons dans la section suivante une formalisation de la notion de résumé basé sur un schéma.

4.2.2 Résumés basés sur un schéma

Intuitivement, un résumé basé sur un schéma d'un ensemble de motifS est un ensemble de références qui couvrent la totalité de l'ensemble de motifs. La définition suivante formalise cette notion :

Définition 4.2.3 (Résumé basé sur un schéma) *Soient un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_{\mathcal{A}}$, une relation de couverture $\triangleleft$ entre $\mathcal{P}$ et $\mathcal{L}_{\mathcal{A}}$, un ensemble de motifs $P \subseteq \mathcal{P}$ et un schéma C appartenant à $\mathcal{C}_{\mathcal{A}}$, le résumé S_C de P basé sur C étant donnée $\triangleleft$ est l'ensemble des références définies sur C qui couvrent au moins un motif de P.*

La couverture totale de l'ensemble de motifs est obtenue grâce à la relation de couverture qui est complète. En effet, pour tout motif de l'ensemble à résumer, il existe au moins une référence du schéma considéré qui le couvre (cf. définition 4.2.2). D'autre part, la non redondance de la relation de couverture permet de garantir que chaque motif est couvert par une seule référence du résumé, ce qui permet d'avoir une meilleure lisibilité des résumés (cf. définition 4.2.1).

Exemple 4.2.2 *Considérons l'ensemble P des requêtes décrites dans le tableau 4.2, le langage de références $\mathcal{L}_{\mathcal{A}}$ où $\mathcal{A}$ est l'ensemble des attributs des tables de la base de données $\mathcal{D}$ décrits dans l'exemple 4.2.1, la relation de couverture $\triangleleft$ non redondante et complète définie dans le même exemple et le schéma $C = \langle SérieTéléphone, CodeForfait, Trimestre \rangle$. Le résumé de P basé sur C est l'ensemble des références décrit dans le tableau 4.3 avec leur couverture. Nous observons que chaque requête de P est couverte par une seule référence définie sur C.*

Le lemme suivant montre qu'un résumé basé sur un schéma d'un ensemble de motifs couvre tous les motifs de cet ensemble.

Lemme 4.2.1 *Considérons un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_{\mathcal{A}}$, une relation de couverture $\triangleleft$ entre $\mathcal{L}_{\mathcal{A}}$ et $\mathcal{P}$, un ensemble de motifs $P \subseteq \mathcal{P}$ et un résumé $S_C \subseteq \mathcal{L}_{\mathcal{A}}$ de P basé sur un schéma. Chaque motif de p est couvert par au moins une référence de S_C.*

Preuve 4.2.1 *La relation de couverture $\triangleleft$ étant complète (cf. définition 4.2.2), il existe pour tout motif de P, une référence définie sur C qui le couvre. Cette référence appartient au résumé S_C basé sur le schéma C car S_C contient toutes les références définies sur C qui couvrent au moins un motif.*

Le lemme suivant montre qu'un résumé basé sur un schéma est synthétique.

q_1	$\pi_{PrixForfait,PrixTéléphone}(\sigma_{SérieTéléphone=\text{"}Apple\ iPhone\ 4\text{"},Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Illimythics\ 5\ Webphone\text{"}}(Forfait \bowtie Temps \bowtie Tarif))$
q_2	$\pi_{PrixTéléphone}(\sigma_{SérieTéléphone=\text{"}Apple\ iPhone\ 4\text{"},Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Illimythics\ 5\ Webphone\text{"}}(Temps \bowtie Tarif))$
q_3	$\pi_{PrixForfait}(\sigma_{SérieTéléphone=\text{"}Apple\ iPhone\ 4,\text{"}Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Illimythics\ 5\ Webphone\text{"}}(Forfait \bowtie Temps))$
q_4	$\pi_{PrixForfait,PrixTéléphone}(\sigma_{SérieTéléphone=\text{"}Apple\ iPhone\ 4,\text{"}Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Illimythics\ 5\ Webphone\text{"},Appels=3h}(Forfait \bowtie Temps \bowtie Tarif))$
q_5	$\pi_{PrixForfait}(\sigma_{SérieTéléphone=\text{"}Apple\ iPhone\ 4,\text{"}Trimestre=\text{"}Trim.2-2010\text{"},CodeForfait=\text{"}Neo.3\text{"}}(Forfait \bowtie Temps))$
q_6	$\pi_{PrixTéléphone}(\sigma_{SérieTéléphone=\text{"}Apple\ iPhone\ 4,\text{"}Trimestre=\text{"}Trim.2-2010\text{"},CodeForfait=\text{"}Neo.3\text{"}}(Temps \bowtie Tarif))$
q_7	$\pi_{PrixForfait,Appels}(\sigma_{SérieTéléphone=\text{"}BlackBerry\ curve\ 9300\text{"},Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Illimythics\ 5\ Webphone\text{"}}(Forfait \bowtie Temps))$
q_8	$\pi_{PrixForfait}(\sigma_{SérieTéléphone=\text{"}BlackBerry\ curve\ 9300\text{"},Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Illimythics\ 5\ Webphone\text{"}}(Forfait \bowtie Temps))$
q_9	$\pi_{PrixForfait,Textos,Internet}(\sigma_{SérieTéléphone=\text{"}BlackBerry\ curve\ 9300\text{"},Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Illimythics\ 5\ Webphone\text{"}}(Forfait \bowtie Temps))$
q_{10}	$\pi_{PrixForfait,PrixTéléphone}(\sigma_{CodeForfait=\text{"}Neo.3\text{"},SérieTéléphone=\text{"}SamsungWave\text{"}}(Forfait \bowtie Tarif))$
q_{11}	$\pi_{PrixForfait,PrixTéléphone}(\sigma_{Trimestre=\text{"}Trim.1-2010\text{"},CodeForfait=\text{"}Neo.3\text{"}}(Forfait \bowtie Temps \bowtie Tarif))$
q_{12}	$\pi_{PrixForfait,PrixTéléphone}(\sigma_{Trimestre=\text{"}Trim.2-2010\text{"},SérieTéléphone=\text{"}SamsungWave\text{"}}(Forfait \bowtie Temps \bowtie Tarif))$

TABLE 4.2 – Ensemble de requêtes fréquentes

Référence	Couverture
$s_1 = \langle Apple\ iPhone\ 4, Illimythics\ 5\ Webphone, Trim.1-2010\rangle$	$\{q_1, q_2, q_3, q_4\}$
$s_2 = \langle Apple\ iPhone\ 4, Neo.3, Trim.2-2010\rangle$	$\{q_5, q_6\}$
$s_3 = \langle BlackBerry\ curve\ 9300, Illimythics\ 5\ Webphone, Trim.1-2010\rangle$	$\{q_7, q_8, q_9\}$
$s_4 = \langle Samsung\ Wave, Neo.3, nulle_{Trimestre}\rangle$	$\{q_{10}\}$
$s_5 = \langle nulle_{SérieTéléphone}, Neo.3, Trim.1-2010\rangle$	$\{q_{11}\}$
$s_6 = \langle Samsung\ Wave, nulle_{CodeForfait}, Trim.2-2010\rangle$	$\{q_{12}\}$

TABLE 4.3 – Résumé basé sur un schéma d'un ensemble de requêtes

Lemme 4.2.2 *Considérons un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_\mathcal{A}$, une relation de couverture $\triangleleft$ entre $\mathcal{L}_\mathcal{A}$ et $\mathcal{P}$, un résumé basé sur un schéma $S_C \subseteq \mathcal{L}_\mathcal{A}$ d'un ensemble de motifs $P \subseteq \mathcal{P}$. Le résumé S_C a une taille plus petite que celle de $\mathcal{P}$, i.e. $|S_C| \leq |P|$.*

Preuve 4.2.2 *Chaque référence de S_C couvre au moins un motif de P (cf. définition 4.2.3) et pour tout motif de P, il existe une seule référence de S_C qui le couvre car $\triangleleft$ est non redondante et complète (cf. définitions 4.2.1 et 4.2.2). Par conséquent, le nombre de références de S_C est au plus égale au nombre motifs dans P.*

Nous montrons dans la section 4.4 qu'un résumé basé sur un schéma est un résumé au sens de la définition 3.8.1 (cf. page 85), i.e. qu'il est un obtenu a partir d'une fonction de résumé couvrante.

Par ailleurs, les résumés basés sur un schéma ont la propriété de minimalité suivante :

Propriété 4.2.1 (Minimalité) *Considérons un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_\mathcal{A}$, une relation de couverture non redondante et complète $\triangleleft$ entre $\mathcal{P}$ et $\mathcal{L}_\mathcal{A}$. Soient un ensemble de motifs $P \subseteq \mathcal{P}$ et l'ensemble $S_C \subseteq \mathcal{L}_\mathcal{A}$ de références définies sur $C \in \mathcal{C}_\mathcal{A}$. Si S_C est un résumé de P basé sur C alors il n'existe pas de sous-ensemble S'_C de S_C tel que S'_C est un résumé de P basé sur C. On dit que S_C est minimal.*

Preuve 4.2.3 *Puisqu'un motif de P ne peut pas être couvert par plus d'une référence de S_C (conséquence de la propriété de non redondance de $\triangleleft$, cf. définition 4.2.1), alors aucun sous-ensemble de S_C ne couvre tous les motifs de P. Donc les sous-ensembles de S_C ne sont pas des résumés de P.*

Ainsi, les résumés basés sur un schéma sont les plus petits résumés possibles qui couvrent la totalité des ensembles qu'ils représentent.

Nous nous intéressons à présent à la manière dont un résumé basé sur un schéma peut être représenté avec un cube. Le schéma sur lequel sont définies les références d'un résumé nous sert de support pour représenter ce résumé. Il suffit juste de lui ajouter une fonction pour obtenir un cube. Plus précisément, un résumé S_C d'un ensemble de motifs P basé sur le schéma C peut être représenté avec un cube de schéma C dont la fonction μ est définie de S_C vers $\mathbb{R}$ et associe à une référence du résumé, une valeur réelle décrivant les motifs couverts par cette référence (par exemple, le nombre de motifs couverts).

Exemple 4.2.3 *Considérons le résumé $S_C = \{s_1, \cdots, s_6\}$ de l'ensemble des requêtes décrites dans le tableau 4.2 dont les références sont présentées dans le tableau 4.3. Ce résumé est basé sur le schéma $C = \langle SérieTéléphone, CodeForfait, Trimestre \rangle$. Nous définissons une fonction μ de S_C vers $\mathbb{R}$ qui à une référence $s \in S_C$, associe le nombre de requêtes couvertes par s. Ce cube est présenté dans la figure 4.3. Pour une meilleure lisibilité, les dimensions $SérieTéléphone$ et $CodeForfait$ sont imbriquées sur un axe.*

Un ensemble de motifs peut avoir beaucoup de résumés basés sur un schéma. Plus précisément, il y a autant de résumés que de schémas dans $\mathcal{C}_\mathcal{A}$, i.e. $2^{2^{|\mathcal{A}|}}$ résumés. Ces résumés sont des représentations à différents niveaux de détail de l'ensemble. Nous nous intéressons dans la section 4.3 à la navigation entre ces résumés.

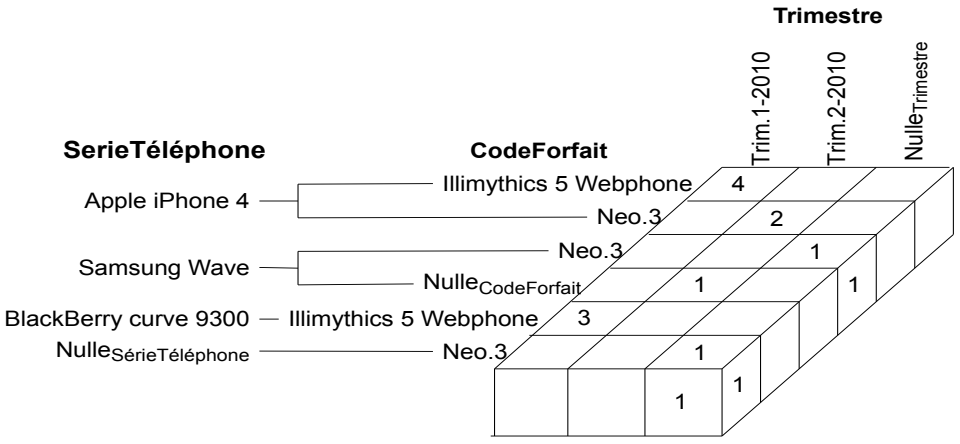


Figure 4.3 – Résumé basé sur un schéma représenté sous forme de cube

4.3 Navigation entre des résumés basés sur un schéma

Les opérateurs classiques de navigation OLAP peuvent être utilisés pour explorer les résumés basés sur un schéma d'un ensemble de motifs. En particulier, les opérateurs de granularité *R*ollup et *D*rilldown permettent de modifier le niveau de granularité des données qui sont représentées dans un cube. *R*ollup consiste à passer d'un niveau plus détaillé à un niveau plus agrégé et *D*rilldown est l'inverse de *R*ollup. Nous adaptons ces opérateurs à notre contexte en nous basant sur la relation de spécialisation entre les résumés, définie comme suit :

Définition 4.3.1 (Spécialisation/généralisation de résumés basés sur un schéma) *Soient un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_\mathcal{A}$, une relation de couverture $\triangleleft$ entre $\mathcal{P}$ et $\mathcal{L}_\mathcal{A}$ et deux résumés S_C, $S_{C'}$ d'un ensemble de motifs $P \subseteq \mathcal{P}$ étant donnée $\triangleleft$ qui sont basés respectivement sur $C = \langle D_1, \cdots, D_N \rangle$ et $C' = \langle D'_1, \cdots, D'_{N'} \rangle$ appartenant à $\mathcal{C}_\mathcal{A}$. $S_{C'}$ est plus spécifique que S_C si $\{D_1, \cdots, D_N\} \subseteq \{D'_1, \cdots, D'_{N'}\}$.*

Exemple 4.3.1 *Le résumé de l'ensemble des requêtes du tableau 4.2 basé sur le schéma $C' = \langle SérieTéléphone, CodeForfait, Trimestre \rangle$ est plus spécifique que celui qui est basé sur le schéma $C = \langle SérieTéléphone, CodeForfait \rangle$ car tous les attributs de C apparaissent dans C'.*

Considérons un ensemble de motifs P , un ensemble d'attributs $\mathcal{A}$, l'ensemble des résumés de P basés sur un schéma appartenant à $\mathcal{C}_\mathcal{A}$, noté $\mathcal{S}_\mathcal{A}$, la relation de spécialisation/généralisation $\preceq$ entre les résumés et les lois internes $\vee$ et $\wedge$ telles que pour tous $S_C, S_{C'} \in \mathcal{S}_\mathcal{A}$, $S_C \vee S_{C'} = S_{C \cup C'}$ et $S_C \wedge S_{C'} = S_{C \cap C'}$. L'ensemble $(\mathcal{C}_\mathcal{A}, \vee, \wedge, \preceq)$ est un treillis de borne inférieure $S_{\langle\rangle}$ et de borne supérieure $S_{\langle\mathcal{A}\rangle}$. Ce treillis est isomorphe au treillis des parties de $\mathcal{A}$. Donc il possède les mêmes propriétés que les treillis des parties d'un ensemble.

Exemple 4.3.2 *Considérons l'ensemble d'attributs $\mathcal{A} = \{CodeForfait, SérieTéléphone, Trimestre\}$. La figure 4.4 montre le treillis des résumés d'un ensemble de motifs basés*

sur un schéma défini sur $\mathcal{A}$.

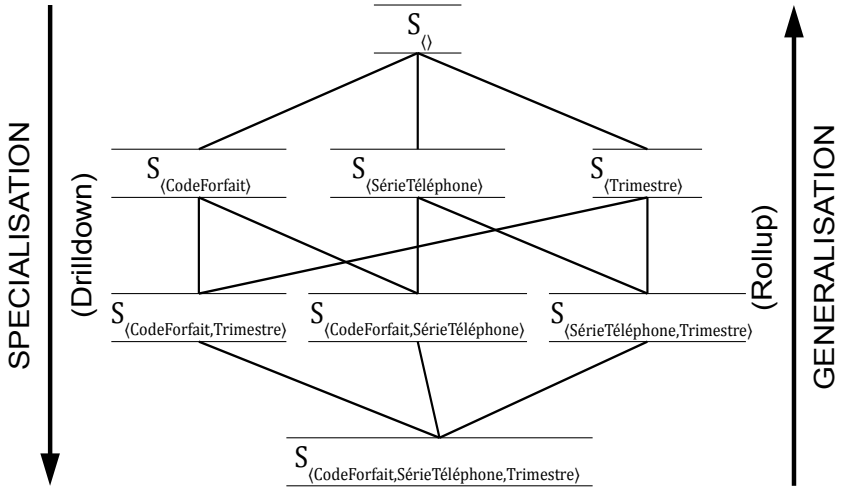


FIGURE 4.4 – Treillis des résumés définis sur $\{SérieTéléphone, CodeForfait, Trimestre\}$

Dans notre approche, la navigation entre les résumés d'un ensemble de motifs consiste à se déplacer dans ce treillis en utilisant des opérateurs de navigation. *R*ollup permet de remonter dans le treillis en allant d'un résumé plus spécifique vers un résumé plus général. *D*rilldown permet d'effectuer l'opération inverse. Les définitions 4.3.2 et 4.3.3 formalisent les opérateurs *R*ollup et *D*rilldown dans notre contexte.

Définition 4.3.2 (Opérateur *R*ollup) *Soient un ensemble d'attributs $\mathcal{A}$, un langage de motifs $\mathcal{P}$, une relation de couverture $\triangleleft$ entre $\mathcal{L}_\mathcal{A}$ et $\mathcal{P}$, un ensemble de motifs $P \subseteq \mathcal{P}$ et l'ensemble $\mathcal{S}_\mathcal{A}$ des résumés de P basé sur un schéma défini sur $\mathcal{A}$. L'opérateur Rollup est une fonction $Rollup : \mathcal{S}_\mathcal{A} \times \mathcal{A} \to \mathcal{S}_\mathcal{A}$ qui associe à un résumé S_C de P basé sur le schéma C étant donnée $\triangleleft$ et un attribut $A \in \mathcal{A}$, le résumé $S_{C \setminus A}$ de P basé sur le schéma $C \setminus A$.*

Exemple 4.3.3 *Considérons le résumé $S_{C'}$ défini sur $\mathcal{A} = \{CodeForfait, SérieTéléphone, Trimestre\}$ avec $C' = \langle SérieTéléphone, CodeForfait, Trimestre\rangle$, l'opération $Rollup(S_C, Trimestre)$ founit le résumé $S_{\langle SérieTéléphone, CodeForfait\rangle}$, qui est plus général que $S_{C'}$.*

Définition 4.3.3 (Opérateur *D*rilldown) *Soient un ensemble d'attributs $\mathcal{A}$, un langage de motifs $\mathcal{P}$, une relation de couverture $\triangleleft$ entre $\mathcal{L}_\mathcal{A}$ et $\mathcal{P}$, un ensemble de motifs $P \subseteq \mathcal{P}$ et l'ensemble $\mathcal{S}_\mathcal{A}$ des résumés de P basé sur un schéma défini sur $\mathcal{A}$. L'opérateur Drilldown est une fonction $Drilldown : \mathcal{S}_\mathcal{A} \times \mathcal{A} \to \mathcal{S}_\mathcal{A}$ qui associe à un résumé S_C de P basé sur le schéma C étant donnée $\triangleleft$ et un attribut $A \in \mathcal{A}$, le résumé $S_{C \cup A}$ de P basé sur le schéma $C \cup A$.*

Exemple 4.3.4 *Considérons le résumé S_C défini sur $\mathcal{A} = \{CodeForfait, SérieTéléphone, Trimestre\}$ avec $C = \langle SérieTéléphone, CodeForfait\rangle$, l'opération $Drilldown(S_C,$*

$Trimestre$) *founit le résumé* $S_{\langle SérieTéléphone, CodeForfait, Trimestre\rangle}$, *qui est plus spécifique que* S_C.

Pour une meilleure lisibilité, nous notons dans la suite de ce chapitre l'opération $\mathcal{O}(S_C, A)$ par $\mathcal{O}_A(S_C)$ où $\mathcal{O}$ est un *Rollup* ou un *Drilldown*. La propriété 4.3.1 montre que n'importe quel résumé du treillis peut être atteint avec les opérateurs *Rollup* et *Drilldown*.

Propriété 4.3.1 (Atteignabilité) *Étant donnés un langage de motifs* $\mathcal{P}$, *un langage de références* $\mathcal{L}_\mathcal{A}$, *une relation de couverture* $\triangleleft$ *entre* $\mathcal{P}$ *et* $\mathcal{L}_\mathcal{A}$, *un ensemble de motifs* $P \subseteq \mathcal{P}$ *et deux résumés* $S_C, S_{C'} \subseteq \mathcal{L}_\mathcal{A}$ *de* P. *Il existe une séquence finie d'opérateurs de granularité* $\mathcal{O}^1, \cdots, \mathcal{O}^N$ *et une séquence finie d'attributs* $A_1, \cdots, A_N$,$\{A_1, \cdots, A_N\} \subseteq \mathcal{A}$, *telles que* $S_{C'} = (\mathcal{O}^1_{A_1} \circ \cdots \circ \mathcal{O}^N_{A_N})(S_C)$.

Preuve 4.3.1 *Considerons les résumés* S_C *et* $S_{C'}$ *de* P *basés respectivement sur les schémas* $C = \langle D_1, \cdots D_N\rangle$ *et* $C' = \langle D'_1 \cdots D'_{N'}\rangle$. *En effectuant une séquence de Rollup à partir de* S_C *avec les attributs* $D_1, \cdots, D_N$, *nous obtenons le résumé* $S_{\langle\rangle} = (Rollup_{D_1} \circ \cdots \circ Rollup_{D_N})(S_C)$ *de* P *basé sur le schema* $\langle\rangle$. *Maintenant, si nous effectuons une séquence de Drilldown à partir de* $S_{\langle\rangle}$ *avec les attributs de* C', *nous obtenons* $S_{\langle D'_1 \cdots D'_{N'}\rangle} = (Drilldown_{D'_1} \circ \cdots \circ Drilldown_{D'_N})(S_{\langle\rangle})$. *Ainsi nous avons au moins une séquence finie d'opérateurs et une séquence finie d'attributs qui permettent d'atteindre* $S_{C'}$ *à partir de* S_C, *i.e* $S_{C'} = (Drilldown_{D'_1} \circ \cdots \circ Drilldown_{D'_N}, Rollup_{D_1} \circ \cdots \circ Rollup_{D_N})(S_C)$.

Ainsi, on peut explorer tout le treillis des résumés en utilisant les opérateurs *Rollup* et *Drilldown*.

Exemple 4.3.5 *Considérons les résumés* S_C *et* $S_{C'}$ *avec* $C = \langle CodeForfait, Trimestre\rangle$ *et* $C' = \langle SérieTéléphone\rangle$. *La figure 4.5 montre la séquence d'opérations qui permet d'atteindre* $S_{C'}$ *à partir de* S_C. *On applique* $Rollup_{Trimestre}(S_{\langle CodeForfait, Trimestre\rangle})$ *qui nous donne le résumé* $S_{\langle CodeForfait\rangle}$. *Puis on applique* $Rollup_{CodeForfait}(S_{\langle CodeForfait\rangle})$ *qui nous donne le résumé* $S_{\langle\rangle}$. *Enfin on applique* $Drilldown_{SérieTéléphone}(S_{\langle\rangle})$ *pour obtenir le résumé final* $S_{\langle SérieTéléphone\rangle}$.

Notons qu'il peut exister plusieurs séquences distinctes d'opérations qui donnent le même résultat. Par exemple, nous pouvons atteindre autrement $S_{\langle SérieTéléphone\rangle}$ à partir de $S_{\langle CodeForfait, Trimestre\rangle}$ en faisant $(Rollup_{CodeForfait} \circ Drilldown_{SérieTéléphone} \circ Rollup_{Trimestre})(S_{\langle CodeForfait, Trimestre\rangle})$.
Dans la section suivante, nous nous intéressons à la construction de résumés basés sur un schéma. Nous y présentations deux fonctions de résumé. La première fonction construit un résumé en utilisant un schéma donné. La deuxième fonction explore le treillis des résumés avec l'opérateur *Drilldown* et retourne un résumé dont la taille maximale est donnée en paramètre.

4.4 Fonction de résumé

Dans cette section, nous décrivons deux fonctions qui permettent de construire des résumés basés sur un schéma : la première permet de contrôler la taille du résumé de

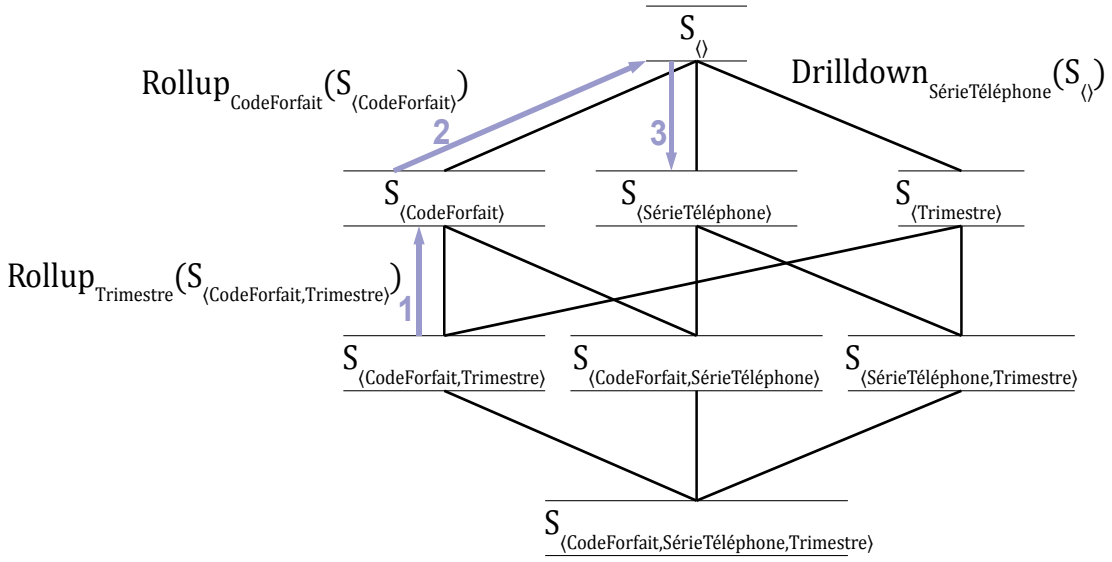


FIGURE 4.5 – atteignabilité de $S_{\langle CodeForfait,Trimestre\rangle}$ à partir de $S_{\langle SérieTéléphone\rangle}$

manière indirecte en spécifiant le schéma à partir duquel il sera construit tandis que la seconde permet de fixer directement la taille maximale du résumé (cf. définition 1.3.3, page 38).

4.4.1 Contrôle indirect de la taille des résumés

Le choix du schéma sur lequel est basé un résumé influe sur la taille de ce dernier. En effet la propriété suivante montre que les schémas ayant moins d'attributs permettent d'obtenir les résumés les plus petits.

Propriété 4.4.1 (Décroissance) *Étant donnés un langage de motifs $\mathcal{P}$, un langage de références $\mathcal{L}_\mathcal{A}$ et une relation de couverture $\triangleleft$ entre $\mathcal{L}_\mathcal{A}$ et $\mathcal{P}$, soient un ensemble de motifs $P \subseteq \mathcal{P}$ et deux résumés de P basés respectivement sur les schémas $C \in \mathcal{C}_\mathcal{A}$ et $C' \in \mathcal{C}_\mathcal{A}$ étant donnée $\triangleleft$. Si C' est plus spécifique que C alors $|S_C| \leq |S_{C'}|$.*

Preuve 4.4.1 *Soient deux résumés S_C et $S_{C'}$ de P basés respectivement sur les schémas $C = \langle D_1, \cdots, D_N\rangle$ et $C' = \langle D_1, \cdots, D_N, D_{N+1}\rangle$. Notons que C' est plus spécifique que C. Pour toute référence $s' = \langle a_1, \cdots, a_N, a_N + 1\rangle \in S_{C'}$, il existe une seule référence $s = \langle a_1, \cdots, a_N\rangle$ de schéma C qui est plus générale que s' et qui appartient à S_C car elle couvre au moins les règles couvertes par s' (cf. définition 4.2.3 et définition 4.2.1).*
D'autre part, pour toute référence de $s \in S_C$, il existe au moins une référence de schéma C' qui est plus spécifique que s et qui appartient à $S_{C'}$ car si tel n'est pas le cas, les références de $S_{C'}$ ne couvriraient pas tous les motifs de P (cf. définition 4.2.2).
Par conséquent, il y a au moins autant de références dans $S_{C'}$ que dans S_C, d'où $|S_C| \leq |S_{C'}|$.

Exemple 4.4.1 *Considérons l'ensemble P des requêtes décrites dans le tableau 4.2, le tableau 4.4 présente le résumé de P basé sur $C' = \langle Trimestre \rangle$. Ce résumé est plus petit (3 références) que celui qui est présenté dans la tableau 4.3 (6 références).*

	$S_{\langle Trimestre \rangle}$	$Couverture$
$Références$	$s'_1 = \langle Trim.1 - 2010 \rangle$	$\{q_1, q_2, q_3, q_4, q_7, q_8, q_9, q_{11}\}$
	$s'_2 = \langle Trim.2 - 2010 \rangle$	$\{q_5, q_6\}$
	$s'_3 = \langle nulle_{Trimestre} \rangle$	$\{q_{10}, q_{12}\}$

Table 4.4 – Le résumé de P basé sur le schéma $\langle Trimestre \rangle$

Dans ce contexte, le paramètre qui permet de contrôler la taille du résumé est le schéma sur lequel il est basé. Plus le schéma est spécifique, plus le résumé est petit. Ainsi, notre fonction de construction de résumé est définie comme suit :

$$\Psi_C : 2^{\mathcal{P}} \rightarrow 2^{Ref(C)} \tag{4.1}$$

où C est un schéma défini sur $\mathcal{A}$, $\mathcal{P}$ est un langage de motifs et $Ref(C)$ est l'ensemble de toutes les références définies sur C. Cette fonction associe à un ensemble de motifs P, un résumé basé sur le schéma C :

$$\Psi_C(P) = \{s \in Ref(C) \mid \exists p \in P, s \triangleleft p\} \tag{4.2}$$

Propriété 4.4.2 *La fonction Ψ_C est une fonction de résumé couvrante.*

Preuve 4.4.2 *Considérons un ensemble de motifs $P \subseteq \mathcal{P}$. $\Psi_C(P)$ est le résumé de P basé sur le schéma C. Donc nous avons les propriétés suivantes :*

(i) Chaque référence de $\Psi_C(P)$ couvre au moins un motif de P (cf. définition 4.2.3) ;

(ii) Chaque motif de P est couvert par une référence de $\Psi_C(P)$ (cf. lemme 4.2.1) ;

(iii) La taille de $\Psi_C(P)$ est plus petite que celle de P (cf. lemme 4.2.2).

De plus, nous avons la propriété (iv) car pour tous $C, C' \in \mathcal{C}_{\mathcal{A}}$, si C' est plus spécifique que C alors $|\Psi_C(P)| \leq |\Psi_{C'}(P)|$ (cf. propriété 4.4.1).

Par conséquent, Ψ_C est une fonction de résumé couvrante.

L'algorithme 4.4.1 est une traduction de cette fonction. Il prend en entrée un schéma et un ensemble de motifs et il retourne un résumé de cet ensemble basé sur le schéma fourni.

A chaque étape, une référence s est prise dans l'ensemble S qui contient au départ toutes les références de $Ref(C)$. Si s couvre un motif de P, alors ce motif ainsi que tous les autres motifs de P couverts par s sont retirés de P et s est supprimé de S. Cette opération est itérée jusqu'à ce qu'il ne reste plus de motifs dans P. Pour un schéma donné, le nombre de motifs dans le résumé obtenu varie entre 1 et $|Ref(C)|$. Cette taille ne peut donc pas être maîtrisée de manière directe. Nous proposons dans la section suivante une fonction qui permet de fixer la taille maximale du résumé d'un ensemble de motifs.

Algorithm 4.4.1 Ψ_C (P) : algorithme de construction de résumé d'un ensemble de motifs étant donnée un schéma

Entrées P {Un ensemble de motifs}, C {Un schéma}
Sortie S_C {Un résumé de P basé sur C}

```
1: S ← Ref(C)
2: Tant que P ≠ ∅ Faire
3:   prendre une référence s ∈ S
4:   Si couverture(s, P) ≠ ∅ Alors
5:     S_C ← S_C ∪ {s}
6:     P ← P\couverture(s, P)
7:     S ← S\{s}
8:   Fin Si
9: Fin Tant que
10: Retourner S_C
```

4.4.2 Contrôle direct de la taille des résumés

Nous souhaitons à présent construire des résumés en ayant la possibilité de fixer directement leur taille. Un ensemble de motifs peut avoir plusieurs résumés basés sur un schéma. Cependant, la taille de ces résumés ne peut pas être maîtrisée directement. Pour atteindre notre objectif, nous proposons d'utiliser une mesure pour évaluer la qualité de ces résumés. Ainsi, le résumé qui sera retenu sera celui qui a la meilleure qualité et dont la taille ne dépasse pas la taille maximale fixé.

Un résumé qui est plus spécifique apporte plus d'information qu'un résumé plus général sur l'ensemble de motifs qui est résumé. Il nous semble donc naturel qu'une mesure de qualité reflète cette propriété sur les résumés en fournissant une meilleure qualité pour un résumé plus spécifique. Par conséquent, les mesures à utiliser pour évaluer la qualité des résumés basés sur un schéma doivent être monotones.

Définition 4.4.1 (Monotonie) *Soit ϕ une mesure de qualité, ϕ est mononone si pour tous résumés S_C et $S_{C'}$ d'un ensemble de motifs P, si $S_{C'}$ est plus spécifique que S_C, alors $\phi(P, S_{C'}) \geq \phi(P, S_C)$*

Par exemple, la mesure de qualité qui à un ensemble de motifs et un résumé de celui-ci associe le ratio entre la taille du résumé et la taille de l'ensemble de motifs est monotone car ce ratio décroît au fur et mesure que la taille du résumé diminue.

A présent, nous disposons de tous les outils nécessaires pour définir notre fonction qui permet de construire des résumés avec un contrôle direct de la taille. Cette fonction est définie comme suit :

$$\Psi_{N,\mathcal{A}} : 2^{\mathcal{P}} \rightarrow 2^{\mathcal{L}_{\mathcal{A}}}$$

où N est la taille maximale du résumé désiré, $\mathcal{A}$ est l'ensemble d'attributs sur lequel les schémas sont définis et $\mathcal{L}_{\mathcal{A}}$ est l'ensemble de références des schémas définis sur $\mathcal{A}$. $\Psi_{N,\mathcal{A}}$ associe à un ensemble de motifs P, un résumé basé sur le schéma C :

$$\Psi_{N,\mathcal{A}}(P) = \arg\max\{\phi(P, S_C) \mid C \in \mathcal{C}_{\mathcal{A}}, |S_C| \leq N\} \tag{4.3}$$

Propriété 4.4.3 *La fonction $\Psi_{N,\mathcal{A}}$ est une fonction de résumé couvrante.*

Preuve 4.4.3 *Considérons un ensemble de motifs $P \subseteq \mathcal{P}$. $\Psi_{N,\mathcal{A}}(P)$ est le résumé de P basé sur le schéma C. Donc nous avons les propriétés (i), (ii) et (iii) de la définition 3.8.1 (cf. preuve 4.4.2). Nous avons également la propriété (iv) car pour tous $P \subseteq \mathcal{P}$, si $\Psi_{N,\mathcal{A}}(P) \leq N$. Par conséquent, $\Psi_{N,\mathcal{A}}$ est une fonction de résumé couvrante.*

Le problème qui consiste à rechercher le résumé de meilleure qualité d'un ensemble de motifs, étant donné une taille maximale, est NP-complet. Sa NP-complétude peut être prouvée en utilisant le problème du sac à dos. Nous proposons donc l'algorithme glouton 4.4.2 qui donne une solution approchée du meilleur résumé.

`Algorithm 4.4.2` $\Psi_{N,\mathcal{A}}(P)$ `: algorithme de résumé d'un ensemble de motifs étant donné la taille maximale du résumé`

`Entrées` P `{Un ensemble de motifs},` $\mathcal{A}$ `{L'ensemble d'attributs sur lequel les schémas sont définis},` N `{Le nombre maximal de motifs du résumé}`
`Sortie` S_C `{Un résumé de` P `basé sur un schéma` $C \subseteq \mathcal{A}$`}`

```
1:  i ← 0
2:  S_{C_i} ← ∅
3:  Répéter
4:     i ← i + 1
5:     S_{C_i} ← S_{C_{i-1}}
6:     Pour tout A ∈ 𝒜 Faire
7:        S_C ← Drilldown_A(S_{C_{i-1}})
8:        Si φ(P, S_C) > φ(P, S_{C_i}) et |S_C| ≤ N Alors
9:           S_{C_i} ← S_C
10:       Fin Si
11:    Fin Pour
12: Jusqu'à φ(P, S_{C_{i-i}}) = φ(P, S_{C_i})
13: Retourner  φ(P, S_C)
```

Notre algorithme prend en entrée un ensemble de motifs, un ensemble d'attributs à partir duquel le schéma du résumé est construit et une valeur réelle correspondant à la taille maximale du résumé désiré. Il utilise l'opérateur de navigation *D*rilldown pour explorer l'espace des résumés et une mesure de qualité qui a la propriété de monotonie. Le principe consiste à sélectionner à chaque étape un résumé dont le schéma est obtenu en ajoutant un attribut au résumé obtenu à l'étape précédent. Ce résumé est celui qui optimise la mesure de qualité utilisée. Nous décrivons ci-dessous les étapes de l'algorithme.

Nous commençons par construire un premier résumé d'ordre 0 basé sur le schéma vide, i.e. le schéma sans dimension (ligne 2). Ensuite, à chaque étape, nous construisons les résumés d'ordre i, i.e les résumés basés sur un schéma obtenu en ajoutant un attribut au schéma du résumé d'ordre $i-1$ qui a été retenu à l'étape précédente (lignes 3 à 12). Le résumé d'ordre i retenu à l'étape i est celui qui a la meilleure qualité (ligne 8). Pour construire un résumé d'ordre i, nous utilisons l'opérateur *Drilldown* auquel nous fournissons le résumé d'ordre $i-1$ et l'attribut à ajouté au schéma du résumé d'ordre $i-1$ pour obtenir celui du

résumé d'ordre i (ligne 7). Nous reviendrons sur cet algorithme dans le chapitre 5 où nous proposons une mesure de qualité pour des résumés d'ensembles de règles d'association. Nous y présenterons un exemple concret d'utilisation de l'algorithme avec cette mesure de qualité.

4.5 Conclusion

Nous avons défini dans ce chapitre un langage de références permettant de résumer des ensembles de motifs à différents niveaux de détail. Ces résumés peuvent être représentés sous forme de cubes. D'autre part, nous avons vu que les opérateurs de navigation OLAP, en particulier les opérateurs de granularité, permettent de naviguer entre ces résumés. Enfin, nous avons proposé deux fonctions de résumé. L'une permet de résumer un ensemble de motifs en donnant le schéma sur lequel il est basé, ce qui permet de manipuler indirectement la taille du résumé. L'autre permet de fixer directement la taille maximale du résumé. L'intérêt de notre cadre est qu'il offre une navigation simple et intuitive de grands ensembles de motifs tout en gardant constamment une vue globale de ces ensembles. Nous exposons dans le chapitre 5 une instanciation de ce cadre avec le langage des règles d'association.

Chapitre 5

Construction de résumés de grands ensembles de règles d'association

Nous avons défini dans le chapitre précédent un cadre formel permettant de construire des résumés d'ensemble de motifs. Dans ce chapitre, notre objectif est d'instancier ce cadre pour résumer des ensembles de règles d'association. Pour atteindre notre objectif, nous proposons dans ce chapitre les composantes suivantes :

- Un langage de règles d'association contenant les règles des ensembles à résumer.
- Un langage de références qui contient les références qui vont constituer les résumés.
- Une relation de couverture non redondante et complète entre le langage de règles d'association et le langage de références. Cette relation intervient dans la construction des résumés. Elle est aussi utilisée dans nos fonctions de résumé avec un contrôle indirect et avec un contrôle direct de la taille des résumés.
- Une mesure de qualité de résumé monotone qui pourra être utilisée dans notre algorithme 4.4.1 pour construire des résumés avec un contrôle direct de leur taille.

Nous décrivons dans la section 5.1 le langage de règles d'association, le langage de références et la relation de couverture. Ensuite, nous présentons dans la section 5.2 un exemple de construction de résumé basé sur un schéma d'un ensemble de règles d'association, avec un contrôle indirect de la taille du résumé. Dans la section 5.3, nous proposons une mesure de qualité basée sur l'entropie de Shannon. Nous présentons quelques propriétés de cette mesure, en particulier sa propriété de monotonie qui permettra de l'utiliser dans notre algorithme pour construire des résumés avec un contrôle direct de leur taille. Un exemple de construction de résumé avec un contrôle direct de la taille est présenté dans la section 5.4. Enfin, nous détaillons nos expérimentations qui portent sur l'évaluation du temps d'exécution de notre algorithme et de la qualité des résumés obtenus sur des bases de règles génériques. Le travail exposé dans ce chapitre a été présenté dans [NDG+09] et [NDG+10].

5.1 Résumés d'ensembles de règles d'association

Dans cette section, nous présentons les langages de règles d'association et de références ainsi que la relation de couverture entre ces deux langages. Notre langage de règles d'association est l'ensemble des règles d'associations définie sur un ensemble d'attributs $\mathcal{A}$.

Rappelons qu'une règle d'association définie sur $\mathcal{A}$ est une implication de la forme $X \Rightarrow Y$ où X et Y sont des itemsets définis sur $\mathcal{A}$ (cf. définition 1.2.4, page 34). Dans la suite, nous notons $\mathcal{R}_{\mathcal{A}}$ le langage des règles définies sur $\mathcal{A}$. La définition suivante introduit la relation de spécialisation entre les règles d'association que nous utilisons dans notre contexte.

Définition 5.1.1 *Étant données deux règles d'association $r : X \Rightarrow Y$ et $r' : X' \Rightarrow Y'$ appartenant au langage $\mathcal{R}_{\mathcal{A}}$, r' est plus spécifique que r si $X \subseteq X'$ et $Y' \subseteq Y$.*

Considérons le langage $\mathcal{R}_{\mathcal{A}}$ des règles définies sur $\mathcal{A}$, notre langage de références $\mathcal{L}'_{\mathcal{A}}$ est l'ensemble des références de schéma défini sur $\mathcal{A}' = \{Corps.A \mid A \in \mathcal{A}\} \cup \{Tête.A \mid A \in \mathcal{A}\}$ auquel nous ajoutons la référence $s_{\langle\rangle}$ tel que pour tout $A \in \mathcal{A}$, $dom(Corps.A) = dom(Tête.A) = dom(A)$ et $s_{\langle\rangle}$ est la référence de schéma $\langle\rangle$. Nous notons une référence de schéma $C = \langle Corps.B_1, \cdots, Corps.B_I, Tête.H_1, \cdots, Tête.H_J\rangle$ définie sur $\mathcal{A}'$ par le N-uplet $\langle b_1, \cdots, b_I, h_1, \cdots h_J\rangle$ où $b_i \in dom(B_i)^+$, $i \in \{1, \cdots, I\}$ et $h_j \in dom(H_j)^+$, $j \in \{1, \cdots, J\}$ s'il n'y a pas d'ambigüité sur les valeurs du domaine des attributs ou si le schéma est précisé.
La relation de spécialisation entre les références est définie comme suit :

Définition 5.1.2 *Étant données deux références $s = \langle b_1, \cdots, b_I, h_1, \cdots h_J\rangle$ et $s' = \langle b'_1, \cdots, b'_{I'}, h'_1, \cdots h'_{J'}\rangle$ appartenant au langage $\mathcal{L}'_{\mathcal{A}}$, s' est plus spécifique que s si $\{b_1, \cdots, b_I\} \subseteq \{b'_1, \cdots, b'_{I'}\}$ et $\{h_1, \cdots h_J\} \subseteq \{h'_1, \cdots h'_{J'}\}$.*

La référence $s_{\langle\rangle}$ est plus générale que toutes les références de $\mathcal{L}'_{\mathcal{A}}$. La relation de couverture entre le langage de règles d'association et le langage de références est définie comme suit :

Définition 5.1.3 (Relation de couverture) *Étant donnés un ensemble d'attributs $\mathcal{A}$, le langage $\mathcal{R}_{\mathcal{A}}$ des règles d'association et le langage des références associées $\mathcal{L}'_{\mathcal{A}}$, une référence $s = \langle b_1, \cdots, b_I, h_1, \cdots, h_J\rangle \in \mathcal{L}'_{\mathcal{A}}$ définie sur le schéma $C = \langle Corps.B_1, \cdots, Corps.B_I, Tête.H_1, \cdots, Tête.H_J\rangle$ couvre une règle $r : X \Rightarrow Y \in \mathcal{R}_{\mathcal{A}}$ ($s \triangleleft r$) si pour tout $a = b_i$, $i \in \{1, \cdots, I\}$ (respectivement $a = h_j$, $j \in \{1, \cdots, J\}$) :*
- *Si $a \neq nulle_{B_i}$ (respectivement $a \neq nulle_{H_j}$), $a \in X$ (respectivement $a \in Y$) ;*
- *Sinon, X (respectivement Y) ne contient pas de valeur appartenant au domaine de l'attribut $Corps.B_i$ (respectivement $Tête.H_j$).*

Intuitivement, une référence couvre une règle si ses valeurs non nulles sont contenues dans la règle, i.e. si les valeurs correspondant aux attributs préfixées par *Corps* et *Tête* sont respectivement dans le corps et la tête de la règle.

Exemple 5.1.1 *Considérons l'ensemble de règles d'association R du tableau 3.2. Rappelons que ces règles sont formées avec les itemsets fréquents fermés extraits à partir de la relation du tableau 1.1. Ces règles sont définies sur l'ensemble d'attributs $\mathcal{A} = \{Marque, Design, Connectivité, Écran, Autonomie, Appareil\ photo, Prix\}$. Soit la référence $s_1 = \langle non\ tactile\rangle$ de schéma $\langle Corps.Écran\rangle$. Elle couvre les règles r_4 : $\{non\ tactile\} \Rightarrow \{2Mp - 5Mp\}$ et $r_8 : \{non\ tactile\} \Rightarrow \{3h - 5h\}$ de R car ces deux règles contiennent l'item non tactile dans leur corps. D'autre part, la référence $\langle nulle_{Écran}\rangle$ du même schéma couvre les règles r_1, r_2, r_5, r_6, r_7 et r_9 car elles ne contiennent pas dans leur corps une valeur appartenant au domaine de l'attribut $Corps.Écran$.*

Propriété 5.1.1 (Non redondance et complétude) *La relation de couverture $\triangleleft$ entre $\mathcal{R}_\mathcal{A}$ et $\mathcal{L}'_\mathcal{A}$ est non redondante et complète.*

Preuve 5.1.1 *Étant donnés le langage de règles d'association $\mathcal{R}_\mathcal{A}$, le langage de références $\mathcal{L}'_\mathcal{A}$, la relation de couverture $\triangleleft$ entre $\mathcal{R}_\mathcal{A}$ et $\mathcal{L}'_\mathcal{A}$, soient une règle $r : X \Rightarrow Y$ appartenant à $\mathcal{R}_\mathcal{A}$ et un schéma $C = \langle Corps.B_1, \cdots, Corps.B_I, Tête.H_1, \cdots, Tête.H_J\rangle$ défini sur $\mathcal{A}'$. Considérons les ensembles $\mathcal{X} = \cup_{B\in\{B_1,\cdots,B_I\}} dom(B)$ et $\mathcal{Y} = \cup_{H\in\{H_1,\cdots,H_J\}} dom(H)$. La référence $s = \langle b_1, \cdots, b_I, h_1, \cdots h_J\rangle$ telle que $\{b_1, \cdots, b_I\} = (X \cap \mathcal{X}) \cup \{nulle_B \mid (B \in \{Corps.B_1, \cdots, Corps.B_I\}) \wedge (dom(B) \cap X = \emptyset)\}$ et $\{h_1, \cdots, h_J\} = (Y \cap \mathcal{Y}) \cup \{nulle_H | (H \in \{H_1, \cdots, Corps.H_J\}) \wedge (dom(H) \cap Y = \emptyset)\}$ est définie sur C et elle couvre r. Donc $\triangleleft$ est complète.*

D'autre part, il est clair que s'il existait une autre référence définie sur C qui couvrait r alors elle aurait les mêmes valeurs que s. Par conséquent, $\triangleleft$ est non redondante.

Exemple 5.1.2 *La règle $r_4 : \{non\,tactile\} \Rightarrow \{2Mp - 5Mp\}$ est couverte uniquement par la référence $s_1 = \langle non\,tactile\rangle$ de schéma $\langle Corps.Écran\rangle$.*

A présent, nous disposons des outils nécessaires pour construire des résumés d'ensembles de règles d'association avec un contrôle indirect de leur taille. La section suivante présente un exemple de construction d'un résumé basé sur un schéma donné.

5.2 Exemple de construction de résumé basé sur un schéma

Reprenons l'ensemble de règles d'association R du tableau 3.2. Soit le schéma $C = \langle Corps.Écran, Tête.Appareil\,photo\rangle$ défini sur $\mathcal{A}'$, le résumé de R basés sur C est construit en générant d'abord toutes les références de schéma C. Ces références sont listées avec leur couverture dans le tableau 5.1. Ensuite, celles qui couvrent au moins une règle sont sélectionnées pour former le résumé.

Référence	Couverture
$\langle tactile, nulle_{Appareil\,photo}\rangle$	r_3
~~$\langle tactile, 2Mp - 5Mp\rangle$~~	
~~$\langle tactile, 6Mp - 9Mp\rangle$~~	
~~$\langle tactile, 10Mp - 14Mp\rangle$~~	
$\langle non\,tactile, nulle_{Appareil\,photo}\rangle$	r_8
$\langle non\,tactile, 2Mp - 5Mp\rangle$	r_4
~~$\langle non\,tactile, 6Mp - 9Mp\rangle$~~	
~~$\langle non\,tactile, 10Mp - 14Mp\rangle$~~	
$\langle nulle_{Écran}, nulle_{Appareil\,photo}\rangle$	r_1, r_2, r_7, r_9
$\langle nulle_{Écran}, 2Mp - 5Mp\rangle$	r_5, r_6
~~$\langle nulle_{Écran}, 6Mp - 9Mp\rangle$~~	
~~$\langle nulle_{Écran}, 10Mp - 14Mp\rangle$~~	

Table 5.1 – Résumé basé sur le schéma $\langle Corps.Écran, Tête.Appareil\,photo\rangle$

Le résumé S_C peut être représenté avec le cube $c = \langle Corps.Écran, Tête.Appareil\ photo, \mu \rangle$ que montre la figure 5.1 où $\mu : S_C \Rightarrow \mathbb{R}$ associe à une référence, le support le plus élevé parmi les supports des règles couvertes par la référence.

	Appareilphoto=null	Appareilphoto=2Mp-5Mp
Ecran=null	0.45454545454545453	0.45454545454545453
Ecran=tactile	0.45454545454545453	
Ecran=non tactile	0.45454545454545453	0.5454545454545454

Figure 5.1 – Cube de schéma $\langle Corps.Écran, Tête.Appareil\ photo \rangle$

Comme nous l'avons précisé dans le chapitre précédent, nous ne pouvons pas maîtriser la taille des résumés basés sur un schéma si nous nous limitons à spécifier le schéma. Nous avons donc proposé une fonction de résumé qui permet de contrôler directement la taille du résumé désiré. Cette fonction correspond à l'algorithme 4.4.2 (cf. page 106) qui utilise une mesure de qualité de résumé qui est monotone. Nous présentons dans la section 5.3 une nouvelle mesure de qualité monotone appelée homogénéité.

5.3 Une mesure de qualité de résumé

Dans cette section, nous proposons une mesure de qualité pour les résumés basés sur un schéma et nous présentons des propriétés de cette mesure.

5.3.1 L'homogénéité d'un résumé basé sur un schéma

Notre mesure de qualité repose sur l'entropie de Shannon [Sha48] qui est une fonction mathématique utilisée en théorie de l'information pour évaluer la quantité d'information contenue ou délivrée par une source d'information. Plus précisément, nous utilisons l'entropie conditionnelle qui correspond à la quantité d'information apportée par une source V si on connaît déjà la source U. Dans un cadre probabiliste, les sources d'information correspondent à des variables aléatoires discrètes.

Etant donné un ensemble d'attributs $\mathcal{A}$, considérons le langage des règles d'association $\mathcal{R}_\mathcal{A}$, le langage des références $\mathcal{L}_\mathcal{A}$ et la relation de couverture $\triangleleft$ non redondante et complète entre $\mathcal{R}_\mathcal{A}$ et $\mathcal{L}_\mathcal{A}$. Dans notre approche, un ensemble de règles $R \subseteq \mathcal{R}_\mathcal{A}$ est bien représenté par un résumé $S_C \subseteq \mathcal{L}_\mathcal{A}$ basé sur le schéma C défini sur $\mathcal{A}'$ si la couverture de chaque référence de S_C est homogène, i.e. les règles qu'elle couvre ont sensiblement les mêmes valeurs dans la tête et le corps. Intuitivement, la qualité de S_C est la qualité globale des couvertures des références de S_C. Nous l'appelons plus simplement l'homogénéité de S_C. Pour évaluer cette homogénéité, nous adaptons l'entropie conditionnelle en considérant les couvertures des références de S_C comme la source connue et les valeurs qui apparaissent dans la tête et le corps des règles de R comme la source inconnue. Nous allons à présent décrire les variables aléatoires correspondant respectivement à ces sources.

Soit U la variable aléatoire discrète définie de R vers S_C qui associe à une règle $X \Rightarrow Y \in R$, la référence $s \in S_C$ telle que s couvre $X \Rightarrow Y$. U est bien une fonction car chaque règle de R est couverte par une seule référence de S_C.

A présent, nous allons définir la variable aléatoire de notre source inconnue. Pour cela nous utilisons la notation X^+ pour exprimer l'extension d'un itemset dans $\mathcal{A}$. Plus précisément, X^+ correspond à l'itemset formé par les valeurs de X et les valeurs nulles des attributs de $\mathcal{A}$ qui ne sont pas représentés dans X. i.e. $X^+ = X \cup \{nulle_A \mid (A \in \mathcal{A}) \wedge (dom(A) \cap X = \emptyset\}$. Soit V la variable aléatoire discrète définie de R vers $dom(\mathcal{A})^+$ qui associe à une règle $X \Rightarrow Y \in R$ et un attribut A, la valeur a si $a \in (X \cup Y)^+$. V est bien une fonction car une règle ne peut pas contenir plus d'une valeur pour le même attribut. $V = a$ représente l'évènement « l'item a appartient à l'extension de l'union de la tête et du corps de la règle $X \Rightarrow Y$ ».

L'homogénéité globale de S_C correspond à l'entropie de V sachant U pondérée par le nombre d'attributs de $\mathcal{A}$. Elle est définie comme suit :

$$\phi(R, S_C) = \frac{1}{|\mathcal{A}|} \sum_{s \in S_C} \varphi(R, s) \tag{5.1}$$

avec $\varphi(R, s) = \sum_{a \in dom(\mathcal{A})^+} p(V = a, U = s) \ln[p(V = a \mid U = s)]$ où $p(V = a, U = s)$ et $p(V = a \mid U = s)$ sont respectivement la probabilité conjointe et la probabilité conditionnelle des évènements $V = a$ et $U = s$. $\phi(R, S_C)$ peut également être notée $frac1|\mathcal{A}|I(V \mid U)$ où $I(V \mid U)$ est l'entropie conditionnelle de V sachant U. ϕ mesure l'homogénéité globale du résumé S_C. Sa valeur est négative ou nulle. Elle vaut à 0 si R est parfaitement homogène, i.e. si dans chaque groupe $couverture(s, R)$, les règles contiennent les mêmes valeurs pour tous les attributs de $\mathcal{A}$. Plus S_C est homogène, plus la valeur de ϕ est élevée. Dans l'exemple suivant, nous détaillons le calcul de l'homogénéité pour la couverture d'une référence.

Exemple 5.3.1 *Considérons l'ensemble de règles R du tableau 3.2 et la référence $\langle$non tactile$\rangle$ de son résumé basé sur le schéma $\langle Corps.Écran \rangle$. Cette référence couvre les règles $r_4 : \{non\ tactile\} \Rightarrow \{2Mp - 5Mp\}$ et $r_8 : \{non\ tactile\} \Rightarrow \{3h - 5h\}$. Pour calculer l'homogénéité de couverture($\langle$non tactile$\rangle, R$), nous considérons uniquement les valeurs qui apparaissent dans r_4 et r_8 et les valeurs nulles des attributs correspondant à ces valeurs, i.e. $a_1 = non\ tactile$, $a_2 = 2Mp - 5Mp$, $a_3 = 3h - 5h$, $a_4 = nulle_{Écran}$, $a_5 = nulle_{Appareil\ photo}$ et $a_6 = nulle_{Autonomie}$. En effet, les autres valeurs ne sont d'aucune utilité pour le calcul car les probabilités qui leur sont associées sont nulles. Ainsi, l'homogénéité de couverture($\langle$non tactile$\rangle, R$) est donnée par la formule suivante :*

$$\varphi(R, \langle non\ tactile \rangle) = \sum_{a \in dom(\mathcal{A})^+} p(V = a, U = \langle non\ tactile \rangle) \ln[p(V = a \mid U = \langle non\ tactile \rangle)]$$

Sa valeur correspond à :

$$\begin{aligned}
\varphi(R, \langle non\ tactile \rangle) &= p(V = a_1, U = non\ tactile) \ln[p(V = a_1 \mid U = \langle non\ tactile \rangle)] \\
&\quad + p(V = a_2, U = non\ tactile) \ln[p(V = a_2 \mid U = \langle non\ tactile \rangle)] \\
&\quad + p(V = a_3, U = non\ tactile) \ln[p(V = a_3 \mid U = \langle non\ tactile \rangle)] \\
&\quad + p(V = a_4, U = non\ tactile) \ln[p(V = a_4 \mid U = \langle non\ tactile \rangle)] \\
&\quad + p(V = a_5, U = non\ tactile) \ln[p(V = a_5 \mid U = \langle non\ tactile \rangle)] \\
&\quad + p(V = a_6, U = non\ tactile) \ln[p(V = a_6 \mid U = \langle non\ tactile \rangle)] \\
&= \frac{2}{9} \ln(\frac{2}{2}) + \frac{1}{9} \ln(\frac{1}{2}) + \frac{1}{9} \ln(\frac{1}{2}) + 0 + \frac{1}{9} \ln(\frac{1}{2}) + \frac{1}{9} \ln(\frac{1}{2}) \\
&= -0.308
\end{aligned}$$

L'exemple suivant décrit le calcul de l'homogénéité du résumé d'un ensemble de motifs.

Référence	Couverture
$\langle tactile \rangle$	r_3
$\langle non\ tactile \rangle$	r_4, r_8, r_9
$\langle nulle_{Corps.Écran} \rangle$	r_1, r_2, r_5, r_6, r_7

Table 5.2 – Résumé d'un ensemble de règles basé sur le schéma $C = \langle Corps.Écran \rangle$

Exemple 5.3.2 *Soit le résumé basé sur le schéma $C = \langle Corps.Écran \rangle$ du tableau 5.2. Le tableau 5.3 montre pour chaque référence de C, le nombre de règles qui sont couvertes par cette référence et qui contiennent la valeur de l'attribut indiqué en colonne. Nous utilisons ces valeurs pour calculer l'homogénéité du résumé S_C.*

$$\begin{aligned}
\phi(R, S_C) &= \frac{1}{7} \times \left[\left(4 \times \frac{1}{9} \ln \frac{1}{1} \right) + \left(4 \times \frac{1}{9} \ln \frac{1}{2} + 3 \times \frac{2}{9} \ln \frac{2}{2} \right) \right] \\
&\quad + \frac{1}{7} \left(3 \times \frac{4}{9} \ln \frac{4}{6} + 3 \times \frac{2}{9} \ln \frac{2}{6} + \frac{6}{9} \ln \frac{6}{6} + \frac{5}{9} \ln \frac{5}{6} + \frac{1}{9} \ln \frac{1}{6} \right) \\
&= -0.251
\end{aligned}$$

L'homogénéité ϕ possède naturellement les propriétés de l'entropie. Nous présentons dans la section suivante quelques unes de ces propriétés.

5.3.2 Les propriétés de l'homogénéité

Nous décrivons dans cette section trois propriétés de l'homogénéité. Pour chacune d'elles, nous considérons que nous disposons d'un ensemble d'attributs $\mathcal{A}$, du langage $\mathcal{R}_\mathcal{A}$ des règles d'association définies sur $\mathcal{A}$, du langage $\mathcal{L}'_\mathcal{A}$ des références définies sur $\mathcal{A}'$, et d'une relation de couverture $\triangleleft$ non redondante et complète entre $\mathcal{R}_\mathcal{A}$ et $\mathcal{L}'_\mathcal{A}$.

La propriété suivante montre que l'homogénéité est mononone.

		Attributs										
		Connectivité		*Appareil photo*		*Écran*			*Autonomie*		*Design*	
Référence	Nombre de règles couvertes	ub	$nulle_{Connectivité}$	$2Mp - 5Mp$	$nulle_{Appareil\ photo}$	$tactile$	$non\ tactile$	$nulle_{Écran}$	$3h - 5h$	$nulle_{Autonomie}$	$monobloc$	$nulle_{Design}$
$\langle tactile \rangle$	1	0	1	0	1	1	0	0	0	1	1	0
$\langle non\ tactile \rangle$	2	0	2	1	1	0	2	0	1	1	0	2
$\langle nulle_{Corps.Écran} \rangle$	6	4	2	4	2	0	1	5	4	2	0	6
S_C	9	4	5	5	4	1	3	5	6	3	1	8

Table 5.3 – Nombre de règles couvertes par les références

Propriété 5.3.1 (Monotonie) *Soient S_C et $S_{C'}$ deux résumés d'un ensemble de règles $R \subseteq \mathcal{R}$ basés respectivement sur C et C'. Si $S_{C'}$ est plus spécifique que S_C alors $\phi(R, S_C) \leq \phi(R, S_{C'})$.*

Preuve 5.3.1 *Considérons les homogénéités de S_C et $S_{C'}$ correspondant respespectivement à $\phi(R, S_C) = \frac{1}{|\mathcal{A}|} I(V \mid U)$ et $\phi(R, S_{C'}) = \frac{1}{|\mathcal{A}|} I(V \mid U')$.*

Soient $C = \langle D_1, \cdots, D_N \rangle$ et $C' = \langle D_1, \cdots, D_N, D_{N+1} \rangle$ les schémas respectifs de S_C et $S_{C'}$. Remarquons que C' est plus spécifique que C. Soit une référence $s' = \langle a_1, \cdots, a_N, a_{N+1} \rangle \in S_{C'}$, la référence $s = \langle a_1, \cdots, a_N \rangle \in S_C$ appartient à S_C et couvre toutes les règles qui sont couvertes par s'. Si une règle est couverte par s ($U = s$) et si elle contient la valeur $a_{N+1} \in dom(D_{N+1})^+$ alors elle est couverte par s' ($U' = s'$).
Donc, l'événement $U' = s'$ correspond à la combinaison des événements $U = s$ et $W = a$ où W est une variable aléatoire qui associe à une règle $X \Rightarrow Y \in R$, la valeur $a \in dom(D_{N+1})^+$ telle que $a \in (X \cup Y)^+$. Ainsi nous pouvons réécrire l'homogénéité de $S_{C'}$ comme suit : $\phi(R, S_{C'}) = \frac{1}{|\mathcal{A}|} I(V \mid U, W)$.
Or, nous avons la propriété suivante de l'entropie : $I(V \mid U, W) \geq I(V \mid U)$ (propriété démontrée dans [Sha48]).
Par conséquent, $\phi(R, S_{C'})$ est supérieure ou égale à $\phi(R, S_C)$.

D'après la propriété 5.3.1, plus on descend dans le treillis des résumés d'un ensemble, plus l'homogénéité des résumé est grande et inversement plus on monte, plus l'homogénéité diminue.

Exemple 5.3.3 *Considérons les résumés S_C et $S_{C'}$ de R décrits dans l'exemple 5.2. Ils sont basés respectivement sur $\langle Corps.Écran \rangle$ et $\langle Corps.Écran, Tête.Appareil photo \rangle$. Notons que $S_{C'}$ est plus spécifique que S_C. Ainsi, l'homogénéité de S_C est inférieure à celle de $S_{C'}$: $\phi(R, S_C) = -0.251$ et $\phi(R, S_{C'}) = -0.159$.*

Il est donc facile de voir que l'homogénéité du résumé le plus spécifique, i.e. celui qui est basés sur le schéma contenant tous les attributs de $\mathcal{A}'$, est maximale.

Propriété 5.3.2 (Valeur maximale) *Soient un ensemble de règles $R \subseteq \mathcal{R}_\mathcal{A}$ et le résumé de $S_C \subseteq \mathcal{L}'_\mathcal{A}$ basé sur le schéma C. $\phi(R, S_C)$ est maximale si $C = \langle \mathcal{A}' \rangle$.*

Cette propriété signifie que ϕ est maximale si chaque référence du résumé couvre une seule règle.

Exemple 5.3.4 *La figure 5.2 montre le résumé de l'ensemble de règles du tableau 3.2 basé sur le schéma $C = \langle Corps.Écran, Corps.Connectivité, Corps.Appareil\ photo, Tête.Écran, Tête.Connectivité, Tête.Appareilphoto, Tête.Design, Tête.Autonomie \rangle$. Les cellules contiennent le nombre de règles couvertes par chaque référence. Ce résumé a une homogénéité nulle car chacune de ses références couvre une seule règle.*

			Design=null				Design=monobloc
			Ecran=null			Ecran=non tactile	Ecran=null
			Autonomie=3h-5h	Autonomie=null		Autonomie=null	Autonomie=null
			Appareilphoto=null	Appareilphoto=null	Appareilphoto=2Mp-5Mp	Appareilphoto=null	Appareilphoto=null
			Connectivité=null	Connectivité=ub	Connectivité=null	Connectivité=null	Connectivité=null
Ecran=null	Connectivité=ub	Appareilphoto=null	1		1		
	Connectivité=null	Appareilphoto=null		1	1		
		Appareilphoto=2Mp-5Mp		1		1	
Ecran=tactile	Connectivité=null	Appareilphoto=null					1
Ecran=non tactile	Connectivité=null	Appareilphoto=null	1		1		

Figure 5.2 – Exemple de résumé homogène d'un ensemble de règles d'association

D'autre part, le résumé le plus général, i.e. celui qui est basé sur le schéma vide, a la plus petite homogénéité.

Propriété 5.3.3 (Valeur minimale) *Soient un ensemble de règles $R \subseteq \mathcal{R}_\mathcal{A}$ et le résumé de $S_C \subseteq \mathcal{L}'_\mathcal{A}$ basé sur le schéma C. ϕ est minimale si $C = \langle\rangle$.*

Plus précisément, ϕ est minimale lorsque le résumé S_C est réduit à l'élément $s_{\langle\rangle}$ qui couvre toutes les règles de R.

Exemple 5.3.5 *L'homogénéité du résumé de l'ensemble de règles du tableau 3.2 basé sur le schéma $C = \langle\rangle$ vaut -0.47. Elle est plus petite que l'homogénéité de tous les résumés de cet ensemble, qui sont basés sur un schéma défini sur $\mathcal{A}'$.*

La propriété de monotonie nous permet d'utiliser l'homogénéité pour construire des résumés avec un contrôle direct de leur taille suivant l'algorithme 4.4.2. Nous présentons dans la section 5.4 un exemple de construction d'un résumé basé sur un schéma en illustrant les différentes étapes de cet algorithme.

5.4 Exemple de construction d'un résumé de taille maximale fixée

Dans cette section, nous appliquons à un ensemble de règles d'association, notre algorithme qui permet de construire un résumé étant donnée une taille maximale. Nous utilisons comme mesure de qualité l'homogénéité qui possède les propriétés de non redondance et de complétude.

Considérons l'ensemble de règles R du tableau 3.2, l'ensemble d'attributs $\mathcal{A} = \{Marque, Design, Connectivité, Écran, Autonomie, Appareilphoto, Prix\}$ et la taille maximale de résumé fixée à $N = 5$. La figure 5.3 montre les résumés générés pendant le processus de construction du résumé.

Dans un premier temps, nous générons le premier résumé $S_{\langle\rangle}$ qui est basé sur le schéma vide et nous calculons son homogénéité. Ce résumé est présenté au niveau 1 de la figure. Les valeurs entre les crochets correspondent respectivement à son homogénéité et à sa taille. Ce résumé contient une seule référence qui est vide et qui couvre toutes les règles de R. Puisque sa taille est inférieure à 5, nous la retenons comme étant le meilleur résumé de niveau 1 et nous passons à l'étape suivante. Notons que le meilleur résumé de chaque niveau est surligné sur la figure.

Au début de la deuxième étape, notre meilleur résumé, noté S_{C_2}, est le meilleur résumé de niveau 1, i.e. $S_{C_2} = S_{\langle\rangle}$. Pour chaque attribut $A \in \mathcal{A}$, nous construisons le résumé $S_{\langle A\rangle}$ basé sur le schéma obtenu en ajoutant A dans le schéma de $S_{\langle\rangle}$. Cette construction est effectuée en appliquant $Drilldown_A(S_{\langle\rangle})$. Le niveau 2 de la figure présente tous les résumés de niveau 2 qui sont générés. Si la taille du résumé obtenu est inférieure à 5, nous calculons son homogénéité puis nous la comparons à celle de S_{C_2}. Si elle est plus élevée que celle de S_{C_2} alors elle remplace S_{C_2}. A la fin de cette étape, nous avons le résumé $S_{C_2} = S_{\langle Corps.Écran\rangle}$ qui est le meilleur de niveau 2. Il contient 3 références et il a une homogénéité égale à 0.251.

A la troisième étape nous procédons comme nous l'avons fait à la deuxième étape. Nous obtenons comme meilleur $\langle Corps.Écran, Tête.Appareil\ photo\rangle$ avec 5 références et une homogénéité égale à 0.159. Nous nous arrêtons à cette étape car la taille maximale est atteinte.

Notons que l'opération $DrilldownA(S_C)$ est effectuée dans la pratique en générant d'abord toutes les références définie sur le nouveau schéma, i.e. le schéma obtenu en ajoutant l'attribut A au schéma C. Cette génération se fait en calculant toutes les combinaisons possibles entre les références du résumé en paramètre et la valeur correspondant à l'attribut en paramètre. Le nouveau résumé est ensuite obtenu en sélectionnant les références du nouveau schéma qui couvrent au moins une règle.

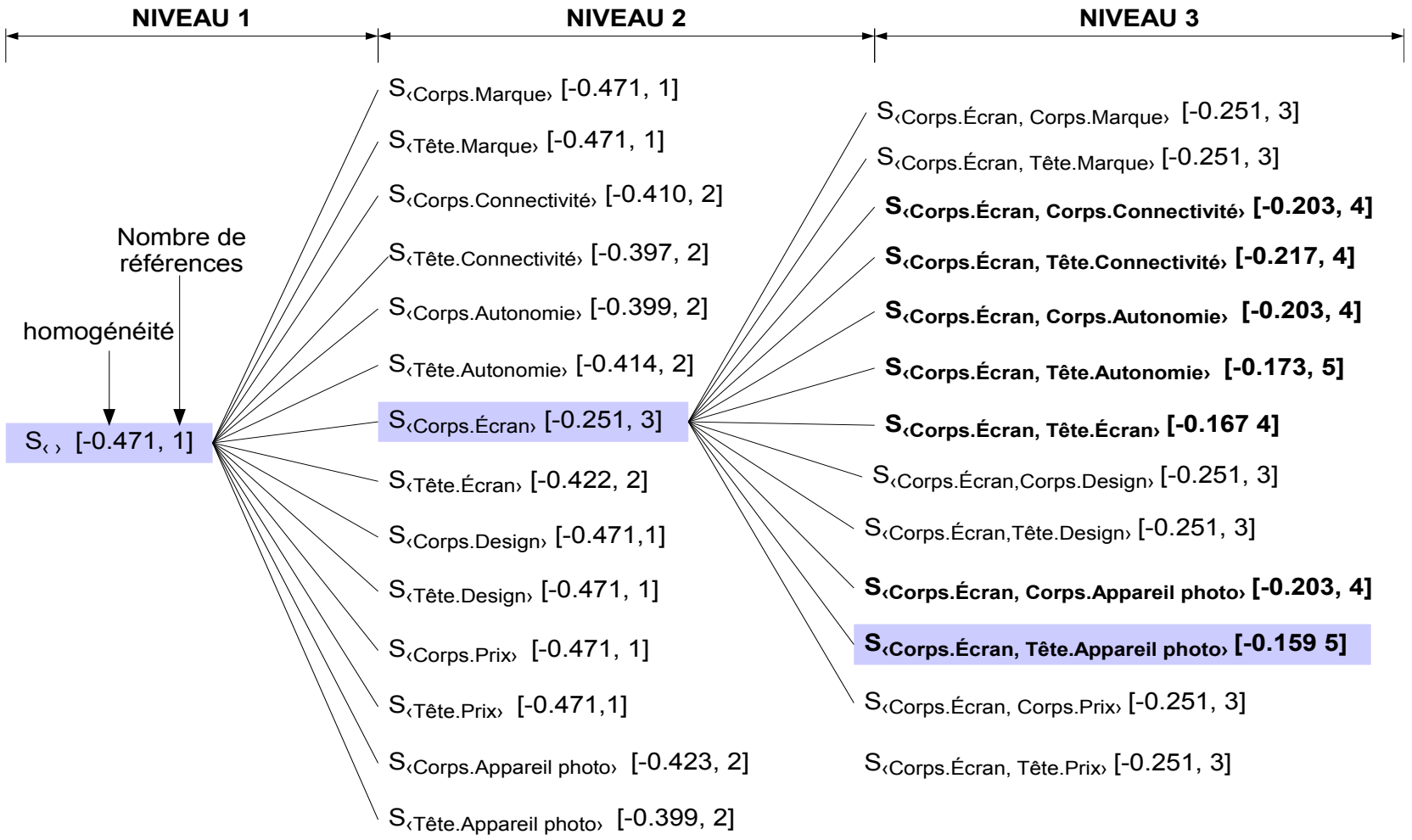


Figure 5.3 – Étapes de construction d'un résumé avec un contrôle direct de sa taille

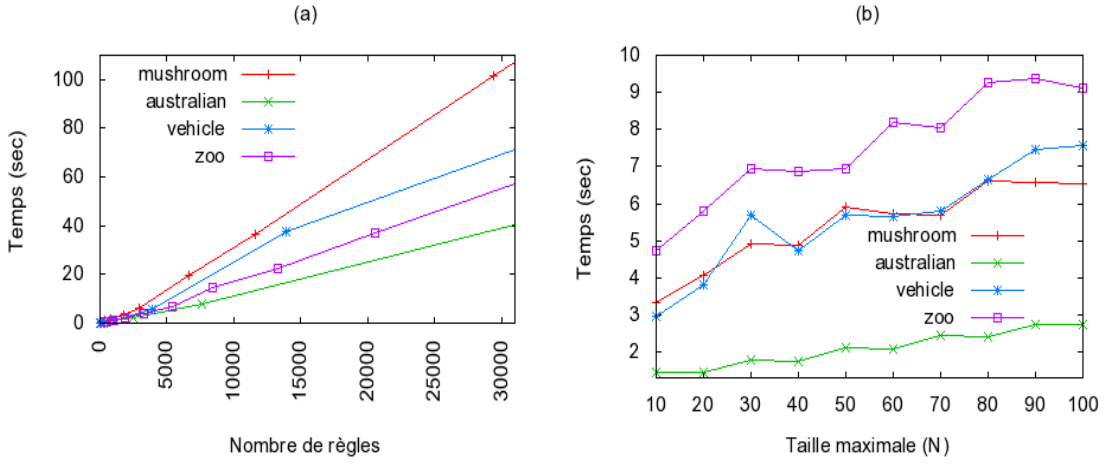


Figure 5.4 – Temps d'éxecution

5.5 Implémentations et expérimentations

Dans cette section, nous étudions les performances de l'algorithme correspondant à la fonction $\Psi_{N,\mathcal{A}}$ que nous avons décrit dans le chapitre 4. Nous utilisons des données discrétisées [1] provenant de UCI Machine Learning repository [2]. Nos tests sont effectués sur des bases génériques de règles d'association extraites avec l'algorithme CHARM [3] [ZH02]. Notre algorithme de construction de résumés est implémenté en java. Par ailleurs, toutes les expériences ont été réalisées avec un ordinateur Intel duo core 2GHz comportant 2 Go de mémoire vive sous Windows Vista. Nous évaluons les performances de l'algorithme en terme de temps d'exécution et d'homogénéité des résumés.

5.5.1 Temps d'exécution

Nous avons effectué les expériences sur le temps d'exécution en utilisant les ensembles de données décrits dans le tableau 5.4. Pour chaque ensemble, nous générons plusieurs ensembles de règles en faisant varier le seuil de support ($minsup$) tandis que le seuil de confiance ($minconf$) est fixé à 50%. La figure 5.4a reporte l'évolution du temps d'exécution en fonction du nombre de règles contenues dans les bases génériques, la taille maximale du résumé étant fixée à $N = 50$. Nous remarquons que le temps d'exécution augmente linéairement quelle que soit la base et il n'excède pas 120 secondes même si l'ensemble de règles contient plus de 30.000 règles. Par ailleurs, la pente des courbes dépend du nombre d'attributs dans $\mathcal{A}$. Plus il y a d'attributs dans $\mathcal{A}$, plus la pente est importante. Par exemple, la courbe de *mushroom* a la plus grande pente car les données sont définies avec plus d'attributs (23 attributs) que les données des autres bases. La figure 5.4b montre l'évolution du temps d'exécution en fonction de la taille maximale des résumés N qui

1. users.info.unicaen.fr/~frioult/uci
2. mlearn.ics.uci.edu/MLRepository.html
3. www.cs.rpi.edu/~zaki/software/

varie entre 10 et 100 sachant que $minsup = 25\%$ pour *mushroom* et *zoo*, $minsup = 15\%$ pour *australian* et *vehicle* et $minconf = 50\%$ pour toutes ces bases. Nous observons que pour tous les ensembles de règles obtenus, le temps d'exécution augmente de manière sous-linéaire en fonction de N. Cela s'explique par le fait que dans l'algorithme, chaque attribut ajouté au schéma du résumé n'est pas testé dans les étapes suivantes.

Données	$\|\mathcal{A}\|$	*minsup* 5%	8%	10%	15%	20%	25%	30%	35%	40%	45%	50%
mushroom	23	-	38406	29441	11629	6681	2915	1732	838	390	227	110
vehicle	19		31873	13890	3899	1066	339	52	4	0	0	0
australian	15	39060	-	7573	2437	1019	486	247	124	62	23	9
zoo	17	31053	-	20583	13253	8446	5382	3283	1864	957	569	300

Table 5.4 – Nombre de règles générées à partir des ensembles de données

5.5.2 Homogénéité des résumés

Dans cette section, nous évaluons l'homogénéité des résumés générés avec notre algorithme. Ainsi, étant donné un ensemble de règles R et une taille maximale de résumé N, nous comparons l'homogénéité des résumés de R basé sur un schéma suivant trois approches :

- **L'approche gloutonne (`gloutonne_RBS`) :** Notre algorithme qui produit une solution approchée du résumé de taille plus petit que N et de meilleure homogénéité.
- **L'approche optimale (`optimale`) :** Un algorithme exhaustif qui parcourt tout l'espace de recherche et qui retourne le résumé optimal.
- **L'approche intermédiaire (`moyenne`) :** Un algorithme exhaustif qui parcourt tout l'espace de recherche et retourne la moyenne des homogénéités des résumés les plus spécifiques dont la taille n'excède pas N.

Bien évidemment, l'algorithme qui génère la solution optimale est très couteux en mémoire et échoue pour les ensembles de règles définis sur un grand nombre d'attributs. Nous avons donc effectué ces expériences sur des ensembles de données avec un nombre restreint d'attributs : *cmc*, *glass*, *tic-tac-toe* et *abalone*. Le tableau 5.5.2 détaille le nombre de règles des bases génériques de règles obtenues à partir de ces ensembles en fonction d'un seuil de support et d'un seuil de confiance donnée.

	Données *cmc*	*glass*	*tic-tac-toe*	*abalone*
$\|\mathcal{A}\|$	10	10	10	9
minsup	8%	6%	5%	20%
minconf	80%	50%	50%	80%
Nombre de règles	1116	1210	1299	1003

Description des bases génériques

La figure 5.5 montre pour chaque base générique, l'homogénéité fournie par les trois approches détaillées précédemment (i.e. `gloutonne_RBS`, `optimale` et `moyenne`) en fonction

de la taille maximale du résumé. Comme prévu, nous remarquons que l'homogénéité augmente en fonction de la taille maximale N. Cette augmentation est due au fait que plus N augmente, plus le résumé est spécifique. Par ailleurs, nous observons que l'homogénéité des résumés produits par notre algorithme est proche de l'homogénéité des solutions optimales pour toutes les bases. Même si N augmente, la distance entre notre solution et la solution optimale reste modérée.

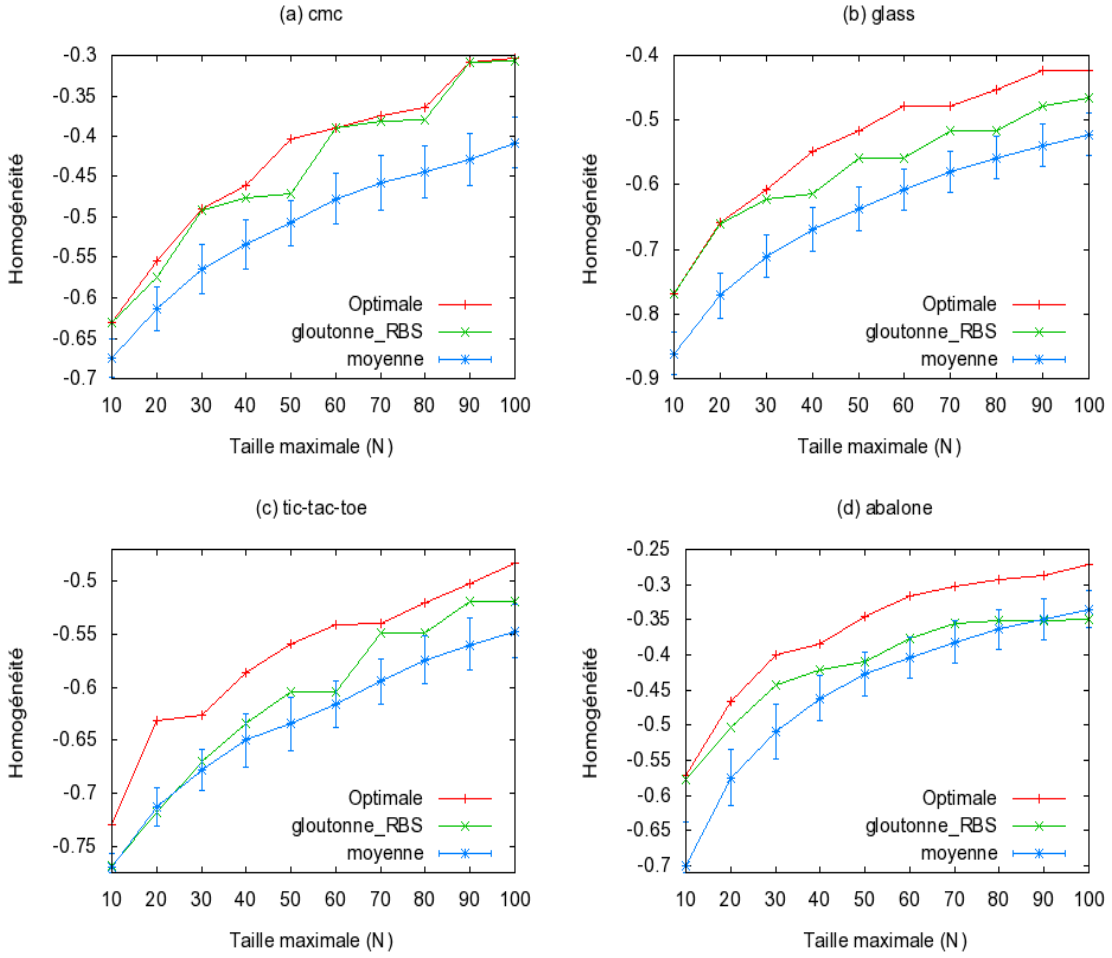


Figure 5.5 – Comparaison de `gloutonne_RBS` avec `optimale` et `moyenne`

D'autre part, nous remarquons que l'homogénéité des résumés obtenus avec l'algorithme `gloutonne_RBS` est pratiquement toujours plus élevée que l'homogénéité moyenne déterminée avec `moyenne`. De plus, elle est dans la plupart des cas au-dessus de l'intervalle de confiance de l'homogénéité moyenne. Nous observons même que sur la figure sur la figure 5.5a, cette homogénéité est optimale pour plusieurs valeurs de N (par exemple, pour les valeurs 30, 60 et 90).

En conclusion, notre algorithme donne des résultats satisfaisants tant sur le temps d'exécution que sur la qualité des résumés générés.

5.6 Conclusion

Nous avons proposé dans ce chapitre une instanciation de notre cadre générique avec les règles d'association. Nous avons également défini une mesure appelée homogénéité qui permet d'évaluer la qualité des résumés. Nous avons montré que l'homogénéité est monotone et qu'elle peut par conséquent être utilisée pour construire des résumés avec un contrôle direct de leur taille. Des tests effectués sur des ensembles génériques de règles ont montré que notre approche est satisfaisante. Notons que si nous regroupons les processus d'extraction et de construction de résumé, le temps de calcul sera légèrement plus important mais il restera quand même raisonnable du fait que les algorithmes d'extraction sont généralement très rapides.

Conclusion

Nous avons présenté dans cette partie un cadre générique pour la construction de résumés de grands ensembles de motifs. Ce cadre permet d'obtenir des résumés qui représentent les ensembles de motifs à plusieurs niveaux de détail et qui peuvent être structurés sous forme de cubes. L'originalité de notre approche est que les résumés donnent une vue globale des ensembles de motifs et les opérateurs de navigation OLAP permettent de les explorer. Nous avons présenté un exemple d'instanciation de notre cadre avec les requêtes. Nous avons proposé une instanciation plus complète avec les règles d'association. Dans ce contexte, nous avons proposé une relation de couverture entre les règles d'association et les références ainsi qu'une mesure de qualité appelée homogénéité que nous utilisons pour générer un résumé intéressant pouvant initialiser l'exploration. Ce travail ouvre des perspectives sur des instanciations avec d'autres types de motifs que nous évoquerons en conclusion de ce mémoire.

Conclusion générale et perspectives

Les algorithmes d'extraction de connaissances génèrent souvent de grandes quantités de motifs. Dans le cadre de cette thèse, nous nous sommes intéressés à l'exploration de grands ensembles de motifs. Une étude des méthodes de construction de résumés et de visualisation d'ensembles de motifs nous a permis de mettre en évidence certaines limites de ces approches. Ces limites sont notamment le manque d'organisation des résumés, l'absence de méthode d'exploration des ensembles de motifs à partir de leur résumé et la difficulté des méthodes de visualisation à représenter de grands ensembles de motifs. Nous avons proposé un cadre générique permettant de construire des résumés qui peuvent être représentés sous forme de cubes et qui permettent d'explorer des ensembles de motifs en utilisant les opérateurs de navigation OLAP. Nous exposons dans la suite un bilan de nos travaux et des perspectives pour nos futurs travaux.

Bilan

Définition d'un résumé Dans les travaux existants, la notion de résumé n'a jamais été clairement définie. Dans ce mémoire, nous avons pris l'initiative de proposer une définition d'un résumé que nous considérons comme étant une représentation synthétique dont la taille peut être contrôlée. En utilisant cette définition, nous avons classé les résumés en deux catégories : Les résumés génératifs et les résumés non génératifs. Les résumés génératifs permettent de régénérer les motifs avec ou sans leurs mesures d'intérêt tandis que les résumés non génératifs ne permettent pas de déduire les motifs. Notons que les résumés génératifs sont plus précisément des représentations condensées approximatives.

Un cadre générique de construction de résumés Le cadre que nous avons proposé ne se limite pas à un type particulier de motifs. Il est défini de manière générique afin de pouvoir être utilisé pour différents types de motifs. Nous l'avons décrit en utilisant comme exemple des requêtes et nous l'avons instancié et implémenté avec des règles d'association. De manière générale, pour instancier notre cadre, il suffit de remplir les conditions suivantes :

- Définir une relation de couverture entre le langage des motifs à résumer et le langage de références ;
- Définir une mesure de qualité des résumés.

Notons que la deuxième condition n'est pas nécessaire si l'on ne souhaite pas contrôler directement la taille des résumés.

Des résumés structurés Notre cadre permet de construire des résumés qui peuvent être structurés sous forme de cube. Plus précisément, un résumé est constitué de références de même schéma et ce schéma permet d'organiser les références dans un cube. Pour construire ces résumés, nous utilisons une relation de couverture définie entre le langage des motifs de l'ensemble à résumer et un langage de références de cubes. Ce langage de références est défini à partir de celui des motifs. Les résumés proposés ne contiennent pas de redondance, i.e. un motif est couvert par une seule référence du résumé. L'absence de redondance permet une meilleure interprétation des résumés. De plus tous les motifs de l'ensemble sont couverts, ce qui donne une vue globale des ensembles de motifs.

Une navigation intuitive Notre cadre permet de résumer un ensemble de motifs suivant plusieurs niveaux de détail. Le niveau de détail est fonction du schéma qui est utilisé. Plus ce schéma est spécifique, plus le résumé est détaillé. Par ailleurs, le fait que les résumés puissent être représentés avec des cubes permet d'utiliser les opérateurs de navigation OLAP. Notamment, les opérateurs de granularité *R*ollup et *D*rilldown permettent d'explorer tout l'espace des résumés d'un ensemble de motifs. L'opérateur *D*rilldown est utilisé pour passer d'un niveau plus détaillé vers un niveau moins détaillé tandis que l'opération *D*rilldown permet d'effectuer l'opération inverse.

Un contrôle direct de la taille des résumés Nous avons proposé un algorithme qui produit un résumé dont la taille maximale est fixée. Cet algorithme utilise l'opérateur *R*ollup pour explorer l'espace des résumés. Il fournit un résumé qui optimise une mesure de qualité donnée. Notons que l'optimum retourné est local. Nous avons présenté des tests sur des bases génériques de règles d'association. Les résultats ont montré qu'on peut obtenir des résumés dans des délais raisonnables. De plus, ces résumés ont une qualité très proche de celle des solutions optimales.

Un outil de construction de résumés Nous avons actuellement une application implémentée en Java qui permet de construire des résumés d'ensemble de règles d'association en spécifiant leur schéma ou leur taille. Ces résumés sont structurés sous forme de cubes qui sont visualisés en forme de tableaux croisés avec des JTAbles. Plusieurs mesures telles que le support, la confiance maximale ou encore l'entropie des règles couvertes par les références, peuvent être affichées dans les cellules. Un clique sur une cellule provoque l'affichage des règles qui sont couvertes par la référence correspondant à la cellule sous forme textuelle comme le montre la figure 5.6.

Perspectives

A court terme

Nos travaux peuvent être étendus à court terme en considérant les perspectives suivantes :

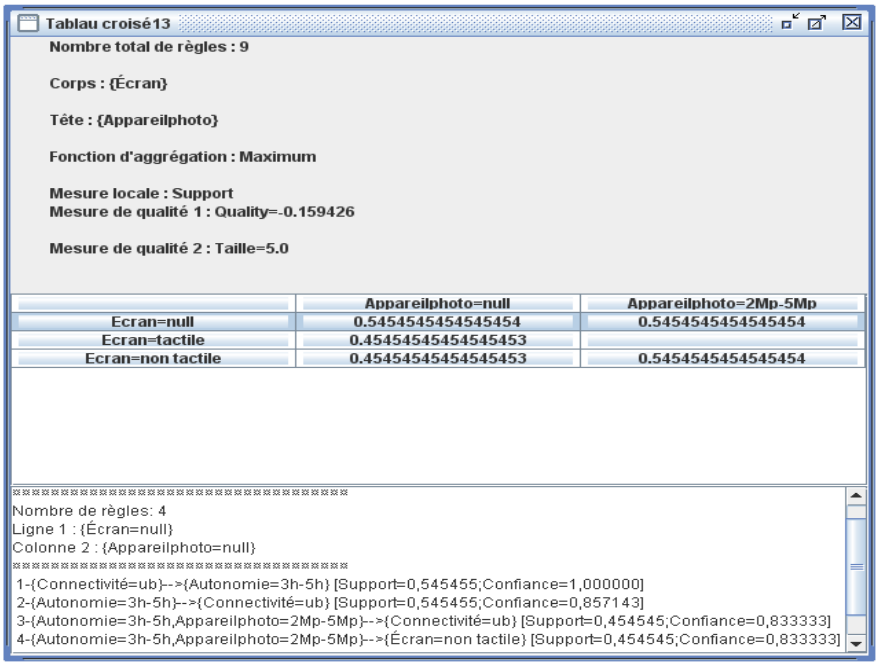


Figure 5.6 – Affichage des règles couvertes par une référence

Prise en compte des hiérarchies Actuellement, nous considérons des données décrites par un ensemble d'attributs. Nous projetons d'élargir notre cadre en introduisant des hiérarchies entre les attributs et entre les valeurs de leur domaine. Cette hiérarchisation nous permettra d'enrichir notre langage de résumés. Par exemple, au lieu d'utiliser un seul attribut $Temp$ pour le temps, on peut considérer les trois attributs $Date$, $Trimestre$ et $Année$ qui représentent le temps et qui sont hiérarchisés en $Date \rightarrow Trimestre \rightarrow Année$ où $Date$ est le niveau le plus bas et $Année$ est le niveau le plus élevé de la hiérarchie. Ainsi, en plus de pouvoir passer d'un résumé à un autre en ajoutant ou en supprimant un attribut quelconque du schéma du premier résumé, on peut naviguer en remplaçant un attribut par un autre se trouvant dans la même hiérarchie. Par exemple, dans un schéma de résumé contenant l'attribut $Année$, en remplaçant ce dernier par l'attribut $Trimestre$, nous obtenons un résumé qui est plus détaillé.

Amélioration de notre application Actuellement, notre algorithme n'a subit aucune optimisation. Nous envisageons donc de le rendre plus performant. En particulier, si nous considérons l'exécution de l'algorithme de construction de résumé avec un contrôle direct de la taille du résumé, nous constatons qu'à une étape donnée, on peut identifier des attributs qui peuvent ne plus être utilisés aux étapes suivantes car ils ne permettent pas d'améliorer la qualité. Par exemple, si nous reprenons les étapes de l'exécution de l'algorithme détaillées sur la figure 5.3 (cf. page 118), nous constatons qu'a la seconde étape, l'ajout de $Corps.Marque$, $Tête.Marque$, $Corps.Design$, $Tête.Design$, $Corps.Prix$ ou $Tête.Prix$ dans le schéma du résumé $S_{\langle\rangle}$ obtenu à la première étape ne permet pas d'améliorer l'homogénéité qui reste égale à -0.471. Ainsi, ces attributs peuvent être supprimés de l'ensemble

des attributs à ajouter lors de la troisième étape, ce qui réduit le nombre de résumés à explorer (nous n'explorons ainsi que les résumés mis en gras à la troisième étape).
D'autre part, nous n'avons implémenté dans notre prototype que l'opérateur *D*rilldown que nous utilisons dans l'algorithme de construction de résumé. Nous souhaitons implémenter l'opérateur *R*ollup mais aussi intégrer d'autres opérateurs tels que des opérateurs qui permettent de naviguer dans les hiérarchies.
Enfin, Le prototype que nous avons développé permet de visualiser les cubes par des tableaux croisés. En cliquant sur une cellule, nous obtenons la liste des règles couvertes par la référence correspondant à la cellule. Cependant, le nombre de règles peut être important, surtout si le résumé est très général. Nous projetons d'organiser ces règles en les triant selon leur support par exemple, mais aussi de les visualiser de manière plus intuitive en utilisant, par exemple, un code de couleur pour mettre en évidence leurs mesures d'intérêt ou leur composition.

A moyen et long termes

À moyen et à long termes nous projetons de travailler sur les perspectives suivantes :

Allègement de la contrainte de couverture totale Dans notre cadre, tous les motifs de l'ensemble à résumer doivent être couverts par les références du résumé. Cependant, il serait intéressant d'alléger cette contrainte en autorisant que certains motifs ne soit pas couvert afin d'avoir des résumés plus synthétiques.

Instanciation avec d'autres types de motifs Dans nos travaux, nous avons effectué une instanciation complète de notre cadre avec les règles d'association. Nous avons aussi effectué une ébauche d'instanciation avec les requêtes. Dans ce contexte, nous avons défini une relation de couverture entre les requêtes et les références. Ce travail peut être poursuivi en proposant une mesure de qualité pour les résumés de requêtes et en effectuant des tests sur des logs de requêtes provenant par exemple de SkyServer [4]. D'autre part, une instanciation avec les itemsets fréquents est triviale, il suffit de considérer les règles dont le corps est vide. Par ailleurs, il serait intéressant d'appliquer notre cadre à d'autres types de motifs tels que les motifs séquentiels ou encore les graphes. Notons qu'on peut rencontrer des difficultés pour représenter ces motifs avec un langage de références, d'où l'intérêt d'explorer d'autres langages de résumé.

Evolution de notre approche Notre cadre repose sur un langage de motifs dans lequel les motifs de l'ensemble à résumer sont pris, un langage de références dans lequel les motifs des résumés sont choisis et une relation de couverture entre le langage des motifs et celui des références. Cette relation de couverture possède deux propriétés qui permettent de représentés les résumés sous forme de cube. D'une part, un motif ne peut pas être couvert par deux références de même schéma (propriété de non redondance). D'autre part, pour tout schéma et pour tout motif, il existe une référence définie sur le schéma qui couvre le motif (propriété de complétude). Notons que ces propriétés font intervenir les schémas qui

4. http://cas.sdss.org/dr7/en/

permettent de représenter les résumés sous forme de cubes. On pourrait imaginer qu'on veuille structurer les résumés sous une autre forme, par exemple sous forme de graphes.

Bibliographie

[AGM04] Foto N. Afrati, Aristides Gionis, and Heikki Mannila. Approximating a collection of frequent sets. In *KDD*, pages 12–19, 2004.

[AIS93] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In *SIGMOD Conference*, pages 207–216, 1993.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[BB00] Jean-François Boulicaut and Artur Bykowski. Frequent closures as a concise representation for binary data mining. In *PAKDD*, pages 62–73, 2000.

[BB04] Dario Bruzzese and Paolo Buono. Combining visual techniques for association rules exploration. In *AVI*, pages 381–384, 2004.

[BBR00] Jean-Francois Boulicaut, Artur Bykowski, and Christophe Rigotti. Approximation of frequency queris by means of free-sets. In *PKDD '00 : Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 75–85, London, UK, 2000. Springer-Verlag.

[BCL01] Paolo Buono, Maria Francesca Costabile, and Francesca A. Lisi. Supporting data analysis through visualizations. In *Proceedings of the International Workshop on Visual Data Mining, in conjunction with 2nd European Conference on Machine Learning (ECML'01) and 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, pages 67–78, 2001.

[BD03] Dario Bruzzese and Cristina Davino. Visual post-analysis of association rules. *J. Vis. Lang. Comput.*, 14(6) :621–635, 2003.

[BGB03] Julien Blanchard, Fabrice Guillet, and Henri Briand. Exploratory visualization for association rule rummaging. In *KDD-03 Workshop on Multimedia Data Mining (MDM-03*, pages 107–114, 2003.

[BKGG03] J. Blanchard, P. Kuntz, F. Guillet, and R. Gras. Implication intensity : From the basic definition to the entropic version. In *Statistical Data Mining and Knowledge Discovery*, pages 475–493. C Press - Chapman and al., 2003.

[BKK97] Clifford Brunk, James Kelly, and Ron Kohavi. Mineset : An integrated system for data mining. In *KDD*, pages 135–138, 1997.

[BMR99] Jean-François Boulicaut, Patrick Marcel, and Christophe Rigotti. Query driven knowledge discovery in multidimensional data. In *DOLAP '99 : Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP*, pages 87–93, New York, NY, USA, 1999. ACM Press.

[BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets : Generalizing association rules to correlations. In *SIGMOD Conference*, pages 265–276, 1997.

[BMUT97] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD Conference*, pages 255–264, 1997.

[BPT+00] Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, and Lotfi Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. *Lecture Notes in Computer Science*, 1861 :972–986, 2000.

[BR98] Jr. Bayardo and J. Roberto. Efficiently mining long patterns from databases. In *SIGMOD '98 : Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 85–93, 1998.

[BR01] Artur Bykowski and Christophe Rigotti. A condensed representation to find frequent patterns. In *PODS*, 2001.

[BSML09] Axel Blumenstock, Franz Schweiggert, Markus Müller, and Carsten Lanquillon. Rule cubes for causal investigations. *Knowl. Inf. Syst.*, 18(1) :109–132, 2009.

[BTP+00] Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme, and Lotfi Lakhal. Levelwise search of frequent patterns with counting inference. In *BDA*, 2000.

[CD97] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and olap technology. *SIGMOD Record*, 26(1) :65–74, 1997.

[CG07] Toon Calders and Bart Goethals. Non-derivable itemset mining. *Data Min. Knowl. Discov.*, 14(1) :171–206, 2007.

[CHYN07] Olivier Couturier, Tarek Hamrouni, Sadok Ben Yahia, and Engelbert Mephu Nguifo. A scalable association rule visualization towards displaying large amounts of knowledge. In *IV*, pages 657–663, 2007.

[CJR06] Olivier Couturier and Vincent Chevrin José Rouillard. Une approche hybride pour une meilleure visualisation de grands ensembles de règles d'association. In *ERGO'IA 2006 :*, 2006.

[CK05] Varun Chandola and Vipin Kumar. Summarization - compressing data into an informative representation. In *Knowledge and Information Systems*, volume 12, pages 355–378. Springer London, 2005.

[CNCL10] Alain Casali, Sébastien Nedjar, Rosine Cicchetti, and Lotfi Lakhal. Constrained cube lattices for multidimensional database mining. *IJDWM*, 6(3) :43–72, 2010.

[Cod70] E. F. Codd. A relational model of data for large shared data banks. *Commun. ACM*, 13 :377–387, June 1970.

[CPL+09] Lucie Copin, Nicolas Pecheur, Anne Laurent, Yudi Augusta, Budi Sentana, Dominique Laurent, and Tao-Yuan Jen. Dbfrequentqueries : Extraction de requêtes fréquentes. In *EGC*, page 499, 2009.

[CRB04] Toon Calders, Christophe Rigotti, and Jean-François Boulicaut. A survey on condensed representations for frequent sets. In *Constraint-Based Mining and Inductive Databases*, pages 64–80, 2004.

[CRC07] Olivier Couturier, José Rouillard, and Vincent Chevrin. An interactive approach to display large sets of association rules. In *HCI (8)*, pages 258–267, 2007.

[CZ03] Sharma Chakravarthy and Hongen Zhang. Visualization of association rules over relational dbmss. In *SAC*, pages 922–926, 2003.

[DGLS04] Cheikh Talibouya Diop, Arnaud Giacometti, Dominique Laurent, and Nicolas Spyratos. Extraction itérative de requêtes fréquentes. *Ingénierie des Systèmes d'Information*, 9(3-4) :83–108, 2004.

[Dio03] Cheikh Talibouya Diop. *Etude et mise en oeuvre des aspects itératifs de l'extraction de règles d'association dans une base de données.* PhD thesis, Université de Tours, France, 2003.

[FPSS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3) :37–54, 1996.

[Fur06] George W. Furnas. A fisheye follow-up : further reflections on focus + context. In *CHI*, pages 999–1008, 2006.

[GH06a] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining : A survey. *ACM Comput. Surv.*, 38(3), 2006.

[GH06b] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining : A survey. *ACM Comput. Surv.*, 38(3), 2006.

[Goe03] Bart Goethals. Survey on frequent pattern mining, 2003.

[HK81] J. Hartigan and B. Kleiner. Mosaics for contingency. In *13th Symposium on the Interface*, pages 268–273, 1981.

[HPYM04] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation : A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1) :53–87, 2004.

[HSW00] Heike Hofmann, Arno Siebes, and Adalbert F. X. Wilhelm. Visualizing association rules with interactive mosaic plots. In *KDD*, pages 227–235, 2000.

[Ins81] Alfred Inselberg. N-dimensional graphics, part i - lines and hyperplanes. *IBM LASC Tech. Rep. G320-2711*, page 140, 1981.

[Ins00] Alfred Inselberg. Visualizing high dimensional datasets and multivariate relations (tutorial am-2). In *KDD '00 : Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–94, New York, NY, USA, 2000. ACM Press.

[JAAXR08] Ruoming Jin, Muad Abu-Ata, Yang Xiang, and Ning Ruan. Effective and efficient itemset pattern summarization : regression-based approaches. In *KDD*, pages 399–407, 2008.

[JLS99] H. V. Jagadish, Laks V. S. Lakshmanan, and Divesh Srivastava. What can hierarchies do for data warehouses ? In *VLDB*, pages 530–541, 1999.

[JLS08] Tao-Yuan Jen, Dominique Laurent, and Nicolas Spyratos. Mining all frequent projection-selection queries from a relational table. In *EDBT*, pages 368–379, 2008.

[JXL09] Ruoming Jin, Yang Xiang, and Lin Liu. Cartesian contour : a concise representation for a collection of frequent sets. In *KDD*, pages 417–426, 2009.

[Kei02] Daniel A. Keim. Information visualization and visual data mining. *IEEE Trans. Vis. Comput. Graph.*, 8(1) :1–8, 2002.

[KGLB00] Pascale Kunt, Fabrice Guillet, Rémi Lehn, and Henri Briand. A user-driven process for mining association rules. In *PKDD '00 : Proceedings of the 4th European Conference on Principles of Data Mining and KnowledgeDiscovery*, pages 483–489, London, UK, 2000. Springer-Verlag.

[KH06] Arno J. Knobbe and Eric K. Y. Ho. Pattern teams. In *PKDD*, pages 577–584, 2006.

[Kim96] Ralph Kimball. *The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses.* John Wiley, 1996.

[KK96a] Daniel A. Keim and Hans-Peter Kriegel. Visualization techniques for mining large databases : A comparison. *IEEE Trans. Knowl. Data Eng.*, 8(6) :923–938, 1996.

[KK96b] Daniel A. Keim and Hans-Peter Kriegel. Visualization techniques for mining large databases : A comparison. *IEEE Trans. on Knowl. and Data Eng.*, 8 :923–938, December 1996.

[KMR+94] Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and A. Inkeri Verkamo. Finding interesting rules from large sets of discoveredassociation rules. In *CIKM*, pages 401–407, 1994.

[KR05] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics).* Wiley-Interscience, 2005.

[Kry98] Marzena Kryszkiewicz. Representative association rules. In *PAKDD*, pages 198–209, 1998.

[Kul59] Solomon Kullback. *Information Theory and Statistics.* Wiley, New York, 1959.

[LC09] Carson Kai-Sang Leung and Christopher L. Carmichael. Fpviz : a visualizer for frequent pattern mining. In *KDD Workshop on Visual Analytics and Knowledge Discovery*, pages 30–39, 2009.

[LHM98] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *KDD*, pages 80–86, 1998.

[LHM99] Bing Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In *KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125–134, New York, NY, USA, 1999. ACM.

[LIC08] Carson Kai-Sang Leung, Pourang Irani, and Christopher L. Carmichael. Fisviz : A frequent itemset visualizer. In *PAKDD*, pages 644–652, 2008.

[LZBX06] Bing Liu, Kaidi Zhao, Jeffrey Benkler, and Weimin Xiao. Rule interestingness analysis using olap operations. In *KDD '06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 297–306, New York, NY, USA, 2006. ACM.

[Mac67] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[MM03] Taneli Mielikäinen and Heikki Mannila. The pattern ordering problem. In *PKDD*, pages 327–338, 2003.

[MT96] Heikki Mannila and Hannu Toivonen. Multiple uses of frequent sets and condensed representations (extended abstract). In *KDD*, pages 189–194, 1996.

[NDG+09] Marie Ndiaye, Cheikh T. Diop, Arnaud Giacometti, Patrick Marcel, and Arnaud Soulet. Construction et exploration de résumés de grands ensembles de règles d'association. In *BDA*, 2009.

[NDG+10] Marie Ndiaye, Cheikh T. Diop, Arnaud Giacometti, Patrick Marcel, and Arnaud Soulet. Cube based summaries of large association rule sets. In *ADMA*, 2010.

[NLHP98] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Han, and Alex Pang. Exploratory mining and pruning optimizations of constrained association rules. In *SIGMOD Conference*, pages 13–24, 1998.

[OES06] Carlos Ordonez, Norberto F. Ezquerra, and Cesar A. Santana. Constraining and summarizing association rules in medical data. *Knowl. Inf. Syst.*, 9(3) :1–2, 2006.

[PG09] Ardian Kristanto Poernomo and Vivekanand Gopalkrishnan. Cp-summary : a concise representation for browsing frequent itemsets. In *KDD*, pages 687–696, 2009.

[PLPP10] Yoann Pitarch, Anne Laurent, Marc Plantevit, and Pascal Poncelet. Fenêtres sur cube. *Ingénierie des Systèmes d'Information*, 15(1) :9–33, 2010.

[PTB+05] Nicolas Pasquier, Rafik Taouil, Yves Bastide, Gerd Stumme, and Lotfi Lakhal. Generating a condensed representation for association rules. *J. Intell. Inf. Syst.*, 24(1) :29–60, 2005.

[SC08] Arnaud Soulet and Bruno Crémilleux. Adequate condensed representations of patterns. *Data Min. Knowl. Discov.*, 17(1) :94–110, 2008.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27 :379–423, 623–656, 1948.

[SVA97] Ramakrishnan Srikant, Quoc Vu, and Rakesh Agrawal. Mining association rules with item constraints. In *KDD*, pages 67–73, 1997.

[TKR+95] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hätönen, and H. Mannila. Pruning and grouping discovered association rules. In *ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47 – 52, 1995.

[TKS02] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41, 2002.

[WB97a] Pak Chung Wong and R. Daniel Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, pages 3–33, Washington, DC, USA, 1997. IEEE Computer Society.

[WB97b] Pak Chung Wong and R. Daniel Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, pages 3–33, Washington, DC, USA, 1997. IEEE Computer Society.

[WP06] Chao Wang and Srinivasan Parthasarathy. Summarizing itemset patterns using probabilistic models. In *KDD*, pages 730–735, 2006.

[WWT99] Pak Chung Wong, Paul Whitney, and Jim Thomas. Visualizing association rules for text mining. In *INFOVIS '99 : Proceedings of the 1999 IEEE Symposium on Information Visualization*, page 120, Washington, DC, USA, 1999. IEEE Computer Society.

[WXL99] Ke Wang, Chu Xu, and Bing Liu. Clustering transactions using large items. In *CIKM*, pages 483–490, 1999.

[Yan05] Li Yang. Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Trans. Knowl. Data Eng.*, 17(1) :60–70, 2005.

[YCHX05] Xifeng Yan, Hong Cheng, Jiawei Han, and Dong Xin. Summarizing itemset patterns : a profile-based approach. In *KDD '05 : Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 314–323, New York, NY, USA, 2005. ACM.

[YN04] Sadok Ben Yahia and Engelbert Mephu Nguifo. Emulating a cooperative behavior in a generic association rule visualization tool. In *ICTAI '04 : Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04)*, pages 148–155, Washington, DC, USA, 2004. IEEE Computer Society.

[Zak00] Mohammed Javeed Zaki. Generating non-redundant association rules. In *KDD*, pages 34–43, 2000.

[ZH02] Mohammed Javeed Zaki and Ching-Jiu Hsiao. Charm : An efficient algorithm for closed itemset mining. In *SDM*, pages 457–473, 2002.

[ZLTX05] Kaidi Zhao, Bing Liu, Thomas M. Tirpak, and Weimin Xiao. Opportunity map : a visualization framework for fast identification of actionable knowledge. In *CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 60–67, New York, NY, USA, 2005. ACM Press.

[ZP03] Mohammed J. Zaki and Benjarath Phoophakdee. Mirage : A framework for mining, exploring and visualizing minimal association rules. Technical report, Computer Science Dept., Rensselaer Polytechnic Inst, 2003.

Résumé :

L'Extraction de Connaissances à partir de Données (ECD) est une discipline dont l'objectif est de trouver de nouvelles connaissances, communément appelées motifs, à partir de bases de données. Elle repose sur des techniques issues de divers domaines tels que les bases de données, les statistiques ou encore l'intelligence artificielle. L'ECD est décrite comme un processus interactif qui consiste à préparer les données, extraire des connaissances à partir de ces données à l'aide d'algorithmes et interpréter les connaissances obtenues. L'interprétation des résultats d'extraction nécessite une exploration des connaissances. Dans ce mémoire, nous nous intéressons à cette étape d'exploration.

De nos jours, il existe beaucoup d'algorithmes d'extraction de connaissances. Ils produisent habituellement de grandes quantités de motifs. Pour faciliter l'exploration de ces motifs, deux approches sont souvent utilisées : la première approche consiste à résumer les ensembles de motifs extraits et la seconde approche repose sur la construction de représentations visuelles de ces motifs. Cependant, dans la plupart des travaux, les résumés ne sont pas structurés et ils sont proposés sans méthode d'exploration. D'autre part, les représentations visuelles n'offrent pas une vue globale des ensembles de motifs.

Notre première contribution est la définition d'un cadre générique qui permet de construire des résumés de grands ensembles de motifs à plusieurs niveaux de détail. Les résumés obtenus peuvent être structurés sous forme de cubes. Cette structuration permet d'explorer les ensembles de motifs via leurs résumés à l'aide d'opérateurs de navigation OLAP. Notre deuxième contribution est la proposition d'un algorithme qui fournit un premier résumé de taille inférieure à un seuil donné, pour initialiser l'exploration des motifs. Les résumés qu'il retourne sont obtenus en maximisant une mesure de qualité des résumés. Enfin, notre troisième contribution est l'instanciation de notre cadre avec les règles d'association. Dans ce contexte, nous proposons une mesure de qualité pour les résumés d'ensembles de règles d'association. Ensuite, nous testons notre algorithme sur des bases génériques de règles d'association en évaluant le temps d'exécution et la qualité des résumés qu'il produit.

Mots clés : Fouille de données, Cubes de données, Résumés d'ensembles de motifs, Règles d'association.

Abstract :

Knowledge Discovery in Databases (KDD) is a discipline whose goal is to find from databases new knowledge commonly named patterns. It is based on techniques from various fields such as databases, statistics or artificial intelligence. KDD is described as an interactive process of preparing the data, extracting knowledge from data using algorithms and interpreting the obtained knowledge. The interpretation of extraction results requires to explore the knowledge. In this thesis, we focus at this step of exploration.

Nowadays, there are many algorithms for extracting knowledge. They usually produce large amounts of patterns. To facilitate the exploration of these patterns, two approaches are often used : the first approach is to summarize the sets of extracted patterns and the second approach relies on the construction of visual representations of these sets of patterns. However, in most work, the summaries are not structured and they are proposed without a method of exploration. Moreover, visual representations do not provide an overview of the sets of patterns.

Our first contribution is the definition of a generic framework for constructing summaries of large sets of patterns at different levels of detail. The obtained summaries can be structured in the form of cubes. This structure allows to explore the sets of patterns through their summaries with OLAP navigation operators.

Our second contribution is the proposal of an algorithm which generates a relevant cube based summary not exceeding a user-specified size, to initialize the exploration of a given rule set. The summaries generated by the algorithm are obtained by maximizing a quality measure of these. Finally, our third contribution is the instantiation of our framework with association rules. In this context, we propose a new quality measure for summaries of sets association rule sets. We test our algorithm on generic bases of association rules by evaluating the execution time and the quality of the summaries it produces.

Keywords : Data mining, Data cubes, summarization of pattern sets, association rules.